

AperTO - Archivio Istituzionale Open Access dell'Università di Torino

**Investigating the use of gradient boosting machine, random forest and their ensemble to predict skin flavonoid content from berry physical-mechanical characteristics in wine grapes**

**This is the author's manuscript**

*Original Citation:*

*Availability:*

This version is available <http://hdl.handle.net/2318/1526764> since 2015-10-21T12:02:43Z

*Published version:*

DOI:10.1016/j.compag.2015.07.017

*Terms of use:*

Open Access

Anyone can freely access the full text of works made available as "Open Access". Works made available under a Creative Commons license can be used according to the terms and conditions of said license. Use of all other works requires consent of the right holder (author or publisher) if not exempted from copyright protection by the applicable law.

(Article begins on next page)



## UNIVERSITÀ DEGLI STUDI DI TORINO

This Accepted Author Manuscript (AAM) is copyrighted and published by Elsevier. It is posted here by agreement between Elsevier and the University of Turin. Changes resulting from the publishing process - such as editing, corrections, structural formatting, and other quality control mechanisms - may not be reflected in this version of the text. The definitive version of the text was subsequently published in: *Computers and Electronics in Agriculture* 117 (2015) 186–193; [dx.doi.org/10.1016/j.compag.2015.07.017](http://dx.doi.org/10.1016/j.compag.2015.07.017)

You may download, copy and otherwise use the AAM for non-commercial purposes provided that your license is limited by the following restrictions:

- (1) You may use this AAM for non-commercial purposes only under the terms of the CC-BY-NC-ND license.
- (2) The integrity of the work and identification of the author, copyright owner, and publisher must be preserved in any copy.
- (3) You must attribute this AAM in the following format: Creative Commons BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/deed.en>), <http://www.sciencedirect.com/science/article/pii/S0168169915002276>

1           **Investigating the use of gradient boosting**  
2           **machine, random forest and their ensemble to**  
3           **predict skin flavonoid content from berry**  
4           **physical-mechanical characteristics in wine**  
5           **grapes**

Luca Brillante<sup>1\*</sup>, Federica Gaiotti<sup>1</sup>, Lorenzo Lovat<sup>1</sup>, Simone Vincenzi<sup>2</sup>, Simone  
Giacosa<sup>3</sup>, Fabrizio Torchio<sup>3</sup>, Susana Río Segade<sup>3</sup>, Luca Rolle<sup>3</sup>, Diego Tomasi<sup>1</sup>

<sup>1</sup> CRA-VIT Council for Agricultural Research and Economics, Viticulture Research  
Center, Conegliano (TV), Italy.

<sup>2</sup> University of Padova, Centro Interdipartimentale per la Ricerca in Viticoltura ed  
Enologia, Legnaro (PD), Italy.

<sup>3</sup> University of Turin, Dipartimento di Scienze Agrarie, Forestali e Alimentari,  
Grugliasco (TO), Italy.

\*Corresponding author

E-mail:[brillanteluca@live.it](mailto:brillanteluca@live.it) (LB)

## 6 **Abstract**

7 Flavonoids are a class of bioactive compounds largely represented in grapevine and  
8 wine. They also affect the sensory quality of fruits and vegetables, and derived products.  
9 Methods available for flavonoid measurement are time-consuming, thus a rapid and  
10 cost-effective determination of these compounds is an important research objective.  
11 This work tests if applying machine learning techniques to texture analysis data allows  
12 to reach good performances for flavonoid estimation in grape berries.

13 Whole berry and skin texture analysis was applied to berries from 22 red wine grape  
14 cultivars and linked to the total flavonoid content. Three machine-learning techniques  
15 (regression tree, random forest and gradient boosting machine) were then applied.  
16 Models reached a high accuracy both in the external and internal validation. The  $R^2$   
17 ranged from 0.75 to 0.85 for the external validation and from 0.65 to 0.75 for the  
18 internal validation, while RMSE (Root Mean Square Error) went from  $0.95 \text{ mg g}^{-1}$  to  
19  $0.7 \text{ mg g}^{-1}$  in the external validation and from  $1.3 \text{ mg g}^{-1}$  to  $1.1 \text{ mg g}^{-1}$  in the internal  
20 validation.

21

22 **Key-words:** random forest; gradient boosting machine (GBM); wine-grape; flavonoids;  
23 texture analysis

24

## 25 **1. Introduction**

26 Flavonoids are a group of secondary metabolites widely distributed in plants, which  
27 greatly affect the sensory and nutritional quality of fruits and vegetables (Harnly et al.,  
28 2006). They represent a huge portion of soluble phenols present in grapevine (Braidot et  
29 al., 2008). Flavonoids are among the most important compounds for the quality of red

30 wine grapes because of their effect on wine sensory attributes (Ristic et al., 2010 is an  
31 example) and aging. The concentration of these compounds in wine depends, among  
32 other factors, on the quantity originally present in grapes (González-Neves et al., 2004).  
33 In the last ten years, flavonoids have received a very great attention from both  
34 researchers and the general audience because of their beneficial effect on human health  
35 (Yao et al., 2004). They have shown antioxidant (Lourenço et al., 2008),  
36 hypocholesterolemic (Gonzalez et al., 2014) and anti-inflammatory effects (Noll et al.,  
37 2009). Their nutraceutical properties are exploited in fresh table grapes, in  
38 pharmaceutical and cosmetic products derived from grape, and are a very appealing  
39 argument for wine marketing purposes.

40 Red grapes are richer in flavonoids than white ones, but their biosynthesis and  
41 concentration greatly depend on cultivar, vineyard practices, soil and climate  
42 (Kondouras et al., 2006). Grape maturity, and therefore the harvest date, is also another  
43 very important parameter because quantitative and qualitative modifications of tannins  
44 and anthocyanins (the two most represented flavonoid families in grape) happen during  
45 ripening (Kuhn et al., 2013).

46 Different methods based on spectrophotometry, chromatography, and mass spectrometry  
47 are usually used for the determination of flavonoids in fruits and vegetables (see Ignat et  
48 al., 2011 for a generic review and Lorrain et al., 2013 for the case of grapes and wine).  
49 Regarding grape analysis, these methods are all very accurate but they often require  
50 sample preparation and long analysis times. The problem is especially the time required  
51 for the extract preparation and purification, which has to be made by hand and can  
52 require berry peeling, solvent extractions, and other manipulations that strongly increase  
53 costs and limit the number of acquirable data. Industry and research will greatly benefit  
54 from a rapid and cost effective method to obtain a faster screening of flavonoids in

55 grapes. Such a method is at today lacking, although recently great advances have been  
56 made in this field by the use of Near InfraRed (NIR) spectroscopy coupled to  
57 chemometrics, in particular using partial least squares (PLS) regression models (Ferrer-  
58 Gallego et al., 2011, Rolle et al., 2012a, Cozzolino 2015).

59 During grape ripening, berries change not only their chemical composition, but also  
60 their mechanical properties: they soften, become less resilient, and the skin generally  
61 harden (Rolle et al., 2012b). In industry, these textural modifications are currently  
62 evaluated by sensory panels to help in the choice of the harvest date. Texture Analysis  
63 (TA) has shown to be an effective instrumental technique for an accurate evaluation of  
64 physical-mechanical characteristics of grapes (Letaief et al., 2008, Giordano et al.,  
65 2013, Battista et al., 2015). It is cost-effective as it [does not](#) require long times and  
66 reagents for sample preparation and analysis.

67 Although flavonoids and texture parameters belong to different grape properties, their  
68 values are both influenced by the berry ripening process. The phenolic ripeness of grape  
69 skin was found to be well assessed when the TA values were used (Río Segade et al.,  
70 2008), but the possibility of a predictive model has been never investigated, and neither  
71 an evaluation of possible chemometrics approaches to these parameters exists. A model  
72 linking the differences in berry mechanical properties and chemical composition  
73 induced by the grape ripeness could be an alternative to NIR methods for rapidly  
74 assessing the flavonoid contents at the berry level.

75 TA data are different from those obtained with NIR. In the first method, the number of  
76 measured parameters available as predictors is limited, and it is generally lower than the  
77 number of observations, i. e. the dataset is in a long format. Conversely, NIR datasets  
78 are wider, the number of wavelengths available as predictors is large and therefore PLS,  
79 a regression algorithm well suited to these situations, has been extensively applied

80 (Cozzolino 2015). With the reduced number of predictors present in TA, other learning  
81 algorithms could be effectively applied as an effort to better exploit the available  
82 information.

83 In this work, we will evaluate the use of regression trees and of two ways of combining  
84 them in order to achieve greater performances in predictions: Random Forest, RF  
85 (Breiman 2001), and gradient boosting machine, GBM (Friedman 2001). RF has shown  
86 to be a state-of-the art method, allowing the highest accuracy, but it is still not  
87 widespread to date. According to a [recent review by Scott et al., 2013](#) for chemometric  
88 classification problems (286 reviewed papers), RF is used in only 4.5% of the articles  
89 where machine-learning algorithms are applied. The same source evidences that  
90 boosting algorithm is even less used (1%).

91 The aim of the work was to evaluate different chemometric approaches in the evaluation  
92 of data obtained from parameters influenced by the grape ripening process, such as  
93 berry mechanical properties data and flavonoid content in berry skins. For this, the  
94 performances of RF and GBM algorithms were compared on a large dataset composed  
95 of approx. 800 berries belonging to 22 grapevine cultivars, their suitability for flavonoid  
96 content prediction in grape berries was evaluated on the basis of mechanical properties,  
97 and an informal explanation of the underlying algorithms was suggested. Furthermore, a  
98 predictive model was also developed. This approach could be used as an example for  
99 other compounds and fruits.

100

## 101 **2. Materials and Methods**

### 102 **2.1 Grape sampling**

103 Grapes from 22 red grapevine cultivars (*Vitis vinifera* L.) were sampled in the CRA-VIT

104 experimental collection (1.2 ha) located in Susegana (TV), Veneto Region (North-East  
105 Italy), in 2010 and 2011. Vines were 15 years old, grafted on SO4 rootstock  
106 (interspecific cross between *Vitis riparia* Michx. and *Vitis berlandieri* Planch.), and  
107 planted at 3.0 m between rows and 1.5 m between vines. They were Sylvoz pruned and  
108 trained with a vertical shoot position system. For each cultivar, samples were composed  
109 of approx. 3 kg of grape berries, which were picked up randomly from ten vines. In  
110 order to successfully compare berries at ripeness with adequate sugar content, the  
111 berries were calibrated using a densimetric method by berry flotation in different saline  
112 solutions (Rolle et al., 2011). This study was carried out only on the berries with sugar  
113 contents comprised between  $183\pm 8$  g L<sup>-1</sup> and  $217\pm 8$  g L<sup>-1</sup> corresponding to  $11.0\pm 0.5\%$   
114 (v/v) and  $13.0\pm 0.5\%$  (v/v) potential alcohol, respectively.

115 The sorted berries were visually inspected before analysis; those with damaged skins  
116 were discarded. For each variety studied, a sub-sample of 36 sorted berries (therefore a  
117 total of 792 berries for all cultivars together) was randomly selected for the  
118 determination of the physical-mechanical properties and then for the flavonoid content.  
119 As described in the successive section, single berries measurements were then averaged  
120 by three to compose a single sample for predictive modeling.

121

## 122 **2.2 Physical and mechanical properties**

123 Grape berries were singularly weighed, with an analytical laboratory balance Radwag  
124 AS 220/X (Radwag, Radom, Poland), and then a Texture Profile Analysis (TPA) non  
125 destructive mechanical test was performed for each of them as described by Letaief et  
126 al., 2008. It allowed the measurement of berry hardness (N, as H), cohesiveness  
127 (adimensional, as Co), gumminess (N, as G), springiness (mm, as S), chewiness (mJ, as  
128 Ch) and resilience (adimensional, as R). A puncture test (Letaief et al., 2008) was then

129 carried out on the same berries taken singularly to measure skin break force (N, as  $F_{sk}$ ),  
130 skin break energy (mJ, as  $W_{sk}$ ) and skin resistance to axial deformation ( $N\ mm^{-1}$ , as  
131  $E_{sk}$ ). All these measurements were performed on the equatorial position of whole berry,  
132 while skin thickness ( $\mu m$ , as  $Sp_{sk}$ ) was measured in the skin after manual removal from  
133 the pulp with a razor blade (Letaief et al. 2008, Río Segade et al. 2011a). Analyses were  
134 made with a Universal Testing Machine (UTM) TAxT2i texture analyzer (SMS-Stable  
135 Micro Systems, Godalming, Surrey, UK) equipped with a 5 kg load cell and a HDP/90  
136 platform. A SMS P/35 flat probe under 25% deformation, with a waiting period of 2s  
137 between the two compressions and a speed of  $1\ mm\ s^{-1}$ , was used for the TPA test. A  
138 SMS P/2N needle probe, with a test speed of  $1\ mm\ s^{-1}$  and a penetration depth of 3 mm,  
139 was used for the puncture test. A SMS P/2 flat probe, with a test speed of  $0.2\ mm\ s^{-1}$   
140 was used to measure  $Sp_{sk}$ . All data were acquired at 400 Hz and evaluated using the  
141 Texture Expert Exceed software, version 2.54.

142

### 143 **2.3 Skin flavonoid content**

144 After the skin thickness test, each berry skin was individually immersed for 4 hours in  
145 5 mL of a buffer solution containing 12% v/v ethanol,  $2\ g\ L^{-1}$  of  $Na_2S_2O_5$ ,  $5\ g\ L^{-1}$  of  
146 tartaric acid and adjusted to pH 3.20 with NaOH (Di Stefano et al., 1991). Each skin  
147 was then homogenized at 8000 rpm for 1 min with an Ultraturrax T18 (IKA  
148 Labortechnik, Staufen, Germany), and the extract was centrifuged for 10 min at  $3500 \times$   
149 g and  $20\ ^\circ C$ . The supernatant was then used for analysis after dilution with an ethanolic  
150 solution of HCl (70:30:1, ethanol:water:HCl, v/v) (Di Stefano et al., 1991). Total  
151 flavonoid index (TF) was determined by a spectrophotometric method, [reading the](#)  
152 [absorbance at 280 nm](#), using an Uvmini-1240 PC spectrophotometer (Shimadzu  
153 Scientific Instruments, Columbia, MD, USA) and expressed as  $mg\ g^{-1}$  berry of (+)-

154 catechin (Rolle et al., 2011, Di Stefano et al., 1991).

## 155 **2.4 Predictive modelling**

### 156 **2.4.1 Description of the used machine-learning techniques**

157 The relationship between predictors and the outcome was modeled using Regression  
158 Trees, RT (Breiman et al., 1984) and two derived techniques: RF (Breiman, 2001) and  
159 stochastic gradient boosting with trees as base learners; the latter will be here called  
160 Gradient Boosting Machine (GBM) in reference to the work where this technique first  
161 appeared (Friedman, 2001). A comprehensive description of these techniques cannot be  
162 given in few words, nevertheless the following paragraphs will try to briefly and lightly  
163 introduce the subject. Readers interested in more technical details can find worthwhile  
164 information in (Hastie, 2009) and in the help and vignettes of the cited R packages.

165 Regression trees are rule based models that split the whole dataset in groups where data  
166 tend to be homogeneous with respect to the response. In the technique used in this work,  
167 which is known as Classification And Regression Tree, CART (Breiman et al., 1984),  
168 data in the terminal nodes (the final groups that are no further partitioned) are simply  
169 averaged to predict the outcome. At the beginning, the entire dataset is split in two  
170 groups to minimize the overall sum of squares, by searching every value of every  
171 predictor. The technique is then recursive, these two groups are split again in two parts  
172 each to further reduce the prediction error, according to the available predictor values.  
173 This technique is also known as recursive partitioning because of its iterative nature.  
174 The number of groups duplicates at each split until the terminal nodes are so small that  
175 they cannot be further partitioned. However, these “full grown” trees generally overfit,  
176 in the sense that they tend to fit the noise other than the structure in the training data.  
177 Therefore, they achieve poor performances on the validation data despite having great  
178 performances on the training data. Their growth must therefore be controlled, and this

179 can be obtained by cross-validation procedures. Cross-validation is a form of internal  
180 validation, which is based on the use of a fraction only of the whole training data to  
181 develop the model, while using the remaining part for the validation. In k-fold cross  
182 validation, the training dataset is divided in k parts; k-1 parts are used to fit the model  
183 and the k<sup>th</sup> part is used to evaluate the structure of the model on simulated new data. The  
184 procedure is then iterative, and all k parts serve as validation once at a time. This allows  
185 determining the size of the trees enabling the best results on future unseen data.

186 A characteristic of regression trees is their instability, their structure can greatly vary  
187 with the data available for modeling. This property can appear at a first sight a  
188 deficiency of the method, instead it has become to be really useful and extremely well  
189 exploited by two state-of-the-art techniques in statistical learning such as bagging and  
190 boosting. These techniques are based on the “perturb and combine” strategies (Breiman,  
191 1996a) and on the idea that combined learners can outperform single ones. For this  
192 combination to be effective, single learners must be able to capture a part of the  
193 structure in the data that is not modeled by other learners. The plasticity of trees can be  
194 exploited for this purpose: by artificially varying the available data through re-sampling  
195 techniques, they can be induced to learn different aspects of the dataset. Partial  
196 predictions from ensemble of trees are then combined to obtain the final predictions.  
197 Bagging (Breiman, 1996b) and boosting (Freund & Schapire 1997) are two ways of  
198 combining learners. In bagging, trees are grown in parallel on a part of the available  
199 data, and predictions are then averaged across all trees. However, in boosting, trees are  
200 grown sequentially, and each successive tree models the residuals of the previous tree  
201 predictions. Bagging and boosting are the two techniques that, further optimized by  
202 increasing randomization, are respectively used in RF and GBM. In RF, trees are grown  
203 on re-sampled subsets of the training data by using only some of the available

204 predictors, randomly chosen at each split. Final predictions for each tree are then  
205 averaged. This parameter is called *mtry* and has to be set by the user, as well as the  
206 number and depth of trees in the forest. In GBM, trees are sequentially built to reduce  
207 the errors of the previous trees, but residuals are resampled and just a fraction is  
208 available for modeling at each iteration. Furthermore, learning is regularized through  
209 shrinkage, i.e. learning rate is slowed by allowing the use of just a fraction of the whole  
210 value for each residual. As occurred in RF, even in GBM, the number and depth of trees  
211 have to be selected by the user. Parameter selection, also called parameter tuning, is  
212 generally made using cross-validation or bootstrap techniques in order to minimize the  
213 performances of the algorithm on simulated new data, being not the error on the training  
214 set a robust choice because of overfitting.

215

#### 216 **2.4.2 Details about the used procedure**

217 A predictive model was built to predict the flavonoid content in berry skin using  
218 physical and mechanical properties of the whole berry and skin as inputs. The 792  
219 berries data (36 berries x 22 cultivars) were averaged by three, randomly selected inside  
220 the same cultivar, to obtain 264 averaged samples (12 samples composed by 3 berries  
221 for 22 cultivars). Prior to model fitting, data were partitioned and a random approx. 20%  
222 of data (53 samples) were left out from the training set for later use as test set. Data  
223 were then centered and scaled. In this work, models were tuned using 10 repetitions of 5  
224 folds cross-validation in order to optimize the Root Mean Square Error (RMSE) on the  
225 resampled data (more than 10 000 possible combinations were evaluated). Performances  
226 of the models were then compared on the same set of 100 bootstrap re-samples always  
227 using RMSE as a metric.

228 Data were analyzed with the R statistical software 3.1.2 (R Core Team, 2014) using the

229 packages rpart (Therneau et al., 2015), randomForest (Liaw & Wiener, 2002), gbm  
230 (Ridgeway, 2013), caret (Kuhn, et al., 2014).

## 231 **3. Results and Discussion**

### 232 **3.1 Descriptive analysis**

233 Fig. 1 shows the TF content for the 22 cultivars used in the experiment. Raboso,  
234 Ancellotta and Teroldego had the highest amount of TF in the skin of fresh berries (with  
235 a median of 8.72, 6.52 and 5.35 mg g<sup>-1</sup> berry, respectively) but also had the highest  
236 variance (1.52, 4.20 and 0.71 mg g<sup>-1</sup> berry, respectively). Gamay, Schiava gentile and  
237 Aleatico had the lowest concentration in these compounds (with a median of 0.97, 0.91  
238 and 0.98 mg g<sup>-1</sup> berry, respectively) and the lowest variance (0.01, 0.01 and 0.01,  
239 respectively). In the global dataset (Table 1), TF had a mean of 2.60 mg g<sup>-1</sup> berry, a  
240 median of 2.07 mg g<sup>-1</sup> berry and a variance of 4.16 mg g<sup>-1</sup> berry. The registered  
241 minimum was 0.69 mg g<sup>-1</sup> berry (Aleatico), while the maximum was 12.87 mg g<sup>-1</sup> berry  
242 (Ancellotta). Descriptive statistics for the physical-mechanical characteristics in the  
243 global dataset is shown in Table 1. Data were in agreement with those reported in  
244 scientific literature in several works (Zouid et al., 2013, Letaief et al., 2008, Río Segade  
245 et al., 2011a).

246 Table 2 shows Pearson correlations in the global dataset. Being TF the outcome of the  
247 developing model, good correlations with the available predictors would be welcomed,  
248 but *r* values for this variable were moderate. BW and S showed the highest correlations  
249 with TF (*r* values of -0.59 and -0.53, respectively, *p*-value < 0.001), followed by E<sub>sk</sub> and  
250 Sp<sub>sk</sub> (*r* values of 0.34 and 0.34, respectively, *p*-value < 0.001). The less related  
251 predictors were W<sub>sk</sub>, H and G, which did not show significant correlations. It should be  
252 noticed that the last two variables were well related to the anthocyanin extractability in  
253 the study published by (Zouid et al., 2013). In the present work, the number of cultivars

254 taken in account is greatly higher compared to the cited work, where only Cabernet-  
255 Franc was measured. Given the lack of significance, it could be hypothesized that the  
256 relation is not uniform but depends on the cultivar. The Spearman's method, used to  
257 highlight possible monotonic, but non-linear relations with TF, did not give association  
258 values higher than those already observed (data not shown).

259 The strongest correlations observed in the dataset ( $r$  values higher than 0.75,  $p$ -value <  
260 0.001) were those among the physical-mechanical predictors (H with G and Ch;  $F_{sk}$  with  
261  $W_{sk}$  and  $E_{sk}$ ; BW with S and SW), which is a consequence of the way they were  
262 measured or calculated (Letaief et al., 2008). H, G and Ch were strongly positive-related  
263 because G and Ch were calculated from H. Therefore, harder berries were also more  
264 gummy and chewy. Considering skin related mechanical properties,  $F_{sk}$  corresponds to  
265 the skin resistance to the needle probe penetration, while  $W_{sk}$  is represented by the area  
266 under the force/time curve.  $E_{sk}$  is defined as the slope of the stress–strain curve in the  
267 linear section.  $W_{sk}$  and  $E_{sk}$  were strongly positive-related to  $F_{sk}$ , so stiffer skins were  
268 also more resistant to the penetration and therefore harder. Furthermore, heavier berries,  
269 which are also bigger ones, had higher value of S and had, obviously, higher amount of  
270 skin. BW was retained instead of S because the relationship of BW with TF is well  
271 known in the literature.

272 It is important to highlight that the sugar content of berry showed some significant  
273 relations (alpha risk < 0.1) with other variables in the dataset, table 2, but the correlation  
274 was strong with none of them. The effect of stage of ripening on mechanical properties  
275 values was in general less determining in comparison to variety effect (Rio Segade et  
276 al., 2008). Cultivar variability of these properties across the 22 cultivars studied clearly  
277 dominated. Before continuing, it should be cleared that even if the correlations in Table  
278 2 let to make an idea of the main relationships in the dataset, possible multidimensional

279 relations were not taken into account. As an example, it seems logical that total skin  
280 weight could be the result of a linear combination between the berry size and the skin  
281 thickness for each berry.  $SW$  was highly related to a combination of  $BW$  and  $Sp_{sk}$  with a  
282  $r$  value of 0.87. This value was clearly higher than those of the single relations, being  
283 probably redundant the information of  $SW$  if used in a model also containing  $BW$  and  
284  $Sp_{sk}$ .

285

### 286 **3.2 Predictor filtering**

287 Sensibility to correlated predictors depends on the used statistical learning technique,  
288 but it is generally not welcomed because redundant and non informative inputs reduce  
289 model performances. When inference is the objective, the negative effect of correlated  
290 variables is even worse than for predictions alone. Furthermore, the measurement of a  
291 greater number of variables in order to apply a model would increase costs and time,  
292 and therefore a justification is required. Any of the three used learning techniques (RT,  
293 RF, GBM) completely fail when correlated predictors are present, the less sensitive  
294 technique probably being GBM because it shrinks effect estimates (Maloney et al.,  
295 2012), and the most sensitive one being RF (Strobl et al., 2007). In general, tree based  
296 techniques implicitly run feature selection because, if a predictor does not permit to  
297 reduce the residual sum of squares at any tree split, its contribution to the model is zero.  
298 However, if highly correlated predictors are present, the choice between them is  
299 somewhat random because they similarly reduce the sum of squares, and have a similar  
300 probability to be chosen for a given split. In RF, where an  $mtry$  number of predictors is  
301 sampled at each split (see section 2.4), the presence of correlated predictors increases  
302 the chance to sample similar information. It reduces randomization and therefore  
303 independence across trees; important assumption to optimize performances. In addition,

304 it dilutes the importance of key predictors and increases the importance of weak  
305 variables correlated to important ones (Strobl et al., 2007).

306 To account for these problems and to optimize model performances, predictors were  
307 first filtered to avoid correlation levels higher than 0.7 (according to Pearson correlation  
308 coefficients, Table 2) and therefore to reduce redundancy. The relationships of the  
309 predictors with the outcome were not considered for selecting predictors in this first  
310 phase. Feature selection was indeed performed successively using Recursive Feature  
311 Elimination (RFE). Four predictors were eliminated by this filtering step, which were  
312  $F_{sk}$ ,  $W_{sk}$ , H, and G.  $F_{sk}$  and  $W_{sk}$  were highly related ( $r = 0.92$ ,  $p < 0.001$ ).  $W_{sk}$  was not  
313 related to TF, contrarily to  $F_{sk}$ , but this last was also strongly related to  $E_{sk}$  ( $r = 0.81$ ,  $p <$   
314  $0.001$ ). A previous study has reported that  $E_{sk}$  is related to cellular maturity index  
315 (EA%) as predictors of anthocyanin extractability (Río Segade, et al., 2011b). Among  
316 H, G and Ch ( $r = 0.91-0.97$ ,  $p < 0.001$ ), this last was retained because it was also well  
317 related to anthocyanin extractability (Zouid et al., 2013). The information provided by  
318 SW in a model can be well approximated by a combined use of BW and  $Sp_{sk}$ , as  
319 previously explained. Furthermore,  $Sp_{sk}$  is considered as main texture parameter to  
320 predict anthocyanin extractability in winegrapes (Río Segade et al., 2011c). The final set  
321 of filtered predictors included BW,  $E_{sk}$ , Co,  $Sp_{sk}$ , Ch,  $W_{sk}$ , and R.

### 322 **3.3 Recursive feature elimination and model tuning**

323 Recursive feature elimination (RFE) (Guyon et al., 2002), a backward selection  
324 algorithm, was used in the way optimized by Ambroise & McLachlan 2002, and  
325 therefore including feature selection in the model building process. Predictors  
326 elimination was evaluated on the basis of the performances achieved on re-sampled sets  
327 obtained by k-fold cross validation ( $k = 5$ ). The process was run for RT, RF, and GBM,  
328 and models were also tuned to optimize performances during the process (see section

329 2.4). RT, RF, and GBM were all tuned using the same set of re-samples therefore  
330 ensuring consistency in the evaluation and allowing comparison across model  
331 performances.

332 In all three cases (RT, RF, and GBM), RFE suggested the use of all seven available  
333 predictors (BW,  $E_{sk}$ , Co,  $Sp_{sk}$ , Ch,  $W_{sk}$ , and R), and therefore all of them had some  
334 influence on the techniques evaluated. The relative predictor importance in all models is  
335 shown in Table 3. In this table, the influence was scaled between 0 and 100 to allow an  
336 easier comparison between models, but as already stated some of the selected predictors  
337 had zero influence. The 0 and 100 are relative values obtained by subtracting the  
338 minimum registered influence (across all predictors) from the individual influence for  
339 each predictor, and then by dividing for the difference between the maximum and the  
340 minimum registered influence. The influence of each predictor in the model varied  
341 according to the model. Considering a single tree (RT), the overall relative influence of  
342 each predictor was higher, because all predictors were used once or few times, and this  
343 avoided the predominance of very strong predictors such as BW. In RT,  $Sp_{sk}$  was the  
344 predictor that allowed the greatest error reduction. In ensembles, and with the  
345 perturbation of data imposed in RF and GBM methods, the influence of some strong  
346 predictors popped up and seems to take advantage over the others. This was more  
347 evident in GBM than in RF, which had an intermediate behavior. These comments are  
348 valid only for this study.

349 Model tuning suggested the use of 7 splits for RT, 1000 trees and  $mtry = 4$  for RF, 5000  
350 trees having 4 splits each and a shrinkage of 0.005 for GBM. Fig. 2 shows the tuned RT  
351 with the aim of illustrating the basic element also composing RF and GBM ensembles.  
352 Single trees are very easy to interpret and to allow making an idea of the relationships in  
353 the dataset. It is important to remember that they are fairly unstable, and small

354 perturbations in the dataset can completely change their structure. Therefore, trees just  
355 describe relationships relative to the data observed, and interpretations are difficultly  
356 generalizable. This is especially true for the lower splits. However, it is worthwhile to  
357 note that  $Sp_{sk}$ , which was the variable with the largest influence for RT (Table 3), acts in  
358 a controversial fashion. For the smallest berries, which were also the richest in  
359 flavonoids, a higher  $Sp_{sk}$  indicated a lower content of TF, while for the biggest berries,  
360 the inverse was true. It is also important to note the role of BW, which was negatively  
361 related to the amount in phenolic compounds as widely discussed in the literature  
362 ([Barbagallo et al., 2011](#)). Also  $E_{sk}$  seems to be an important parameter because elastic  
363 skins were associated to higher content in TF. BW,  $E_{sk}$ , and  $Sp_{sk}$  were the predictors  
364 with the highest influence in all models (Table 3).

365

### 366 **3.4 Model comparison**

367 Results of the tuned models are shown in Table 4 for both training and test data and for  
368 the cross-validated re-samples. It appears that all algorithms, starting from RT, tended to  
369 overfit the training set, which is probably a consequence of a training set too small  
370 when compared to the complexity of the relationships among predictors and between  
371 these and the outcome, as suggested by the weak correlations observed in Table 2.  
372 Despite this, the model accurately predicted the test set used as external validation.  
373 Predictions for the test set exceeded those obtained with 5-fold cross-validation, which  
374 can be considered an internal validation. Test sets are considered the ultimate proof of  
375 model performances, often neglecting cross-validation and bootstrap assessment  
376 methods. However, observations like the present one make us think about the method to  
377 prefer in model assessment. The performances observed over a test set could also be  
378 attributed to random select observations easier to predict than those contained in the

379 training set used in cross-validation. Resampling methods are more robust from this  
380 point of view, but they can be upward (i. e. pessimistic) biased, especially the bootstrap,  
381 even if an alternative to avoid such bias is available, but only for classification problems  
382 (Efron & Tibshirani 1997).

383 Fig. 3 shows the predictions on the [train and test sets](#) for all methods. Figs. 3a and 3b  
384 shows the blocky structure used in prediction by RT, where similar data were just  
385 predicted by the mean of the group they belong according to the rules in Fig. 2, and are  
386 therefore grouped also in predictions. In GBM and RF ensembles, the predictions are  
387 averaged from many trees and allow the methods to be more adaptable to the form of  
388 data and to also model non-linearity (and interactions). In Figs. 3c/d and 3e/f,  
389 predictions were no longer grouped for the same values of predictions. Comparing  
390 Figs 3c/d with 3e/f, it appears that both RF and GBM methods well predict the test set,  
391 but predictions, as shown by the location of points in the scatterplots, although similar  
392 were not exactly identical. Predictions obtained on the same re-sampled data by GBM  
393 and RF were highly correlated (0.82), however looking at Table 3 it appears that they  
394 did not make use of the same predictors in the same way. It will be possible that  
395 combining both methods in a single ensemble will boost the overall accuracy a little bit.  
396 These algorithms were combined by weighted average of their predictions, using a  
397 greedy optimization method as described in (Caruana et al., 2006). Combining methods  
398 in a single ensemble of models has the highest efficacy when algorithms are different,  
399 and therefore predictions uncorrelated. It is not the case here, where these assumptions  
400 are not really respected. However, the combination of RF and GBM brings to a nice  
401 improvement in model predictions. GBM and RF predictions were weighted 0.51 and  
402 0.49, respectively, for averaging, and the resulting RMSE of their ensemble was 1.05  
403  $\text{mg g}^{-1}$ , which was slightly lower than that obtained by single methods (Table 4). The

404 result on the test set for the RMSE was  $0.701 \text{ mg g}^{-1}$  and for  $R^2$  was 0.85. The  
405 corresponding predictions on the train set are in Fig. 3g and in the test set are shown in  
406 Fig. 3h.

407 To compare the results of this work with others found in the literature is somewhat  
408 difficult, because the use of texture analysis to predict flavonoid content in grape berries  
409 is novelty, and also a so varied dataset, containing 22 cultivars, is rare to be found.  
410 Texture analysis was already used to develop rapid method for the evaluation of total  
411 phenolic content and phenol extractability in grape seeds with a good accuracy (Rolle et  
412 al., 2013), and in skins but limited to the anthocyanin content (Rolle et al., 2012b, Rio  
413 Segade 2011c). In grape berries, however, a rapid evaluation of the phenolic content has  
414 generally been made using NIR spectroscopy, and several works have reached very  
415 good performances (Ferrer-Gallego et al., 2011). This work was performed on a single  
416 cultivar (Graciano), and data were expressed in  $\text{mg g}^{-1}$  of berry skins. In the present  
417 work, data were expressed in  $\text{mg g}^{-1}$  of whole berry, which from an industrial point of  
418 view could be more practical. The results of the last two studies are therefore not  
419 directly comparable.

420 The results obtained showed that RF and GBM, and even their average can reach a very  
421 high accuracy for TF prediction from physical-mechanical data obtained for many  
422 different cultivars. RF is simpler to perform and accurately tune than GBM.  
423 Furthermore, its use of features in the dataset less overfitted BW influence.

424 However, even if the performances of those algorithms were very high, it is also true  
425 that for real world application, model performances were still too low to be practically  
426 used in a generalized way. Results obtained with prediction could be useful for the  
427 comparison of the phenolic maturity of different vineyards, but at this time they were  
428 hardly suitable for the monitoring of TF during ripening for cultivars with low amounts

429 of flavonoids. Conversely, they could be used for those cultivars with very high  
430 amounts of flavonoids (such as Raboso, Ancellotta and Teroldego), because a reduced  
431 relative error in prediction.

432 It is possible that, being the physical-mechanical characteristics linked to total  
433 flavonoids in a way that depends on the cultivar and is not universal, cultivar-specific  
434 calibration will be necessary to improve model performances. This will probably allow  
435 the use of TA to monitor grape ripening even for cultivars with low amount in  
436 flavonoids. Cultivar- specific calibration or the inclusion of the cultivar as a categorical  
437 term in the developed models was not possible in this work because the number of  
438 observations by cultivar was too low.

439 To further increase model accuracy, it will also be interesting to test the average of more  
440 than three berries for a single sample in order to improve the accuracy in the TA  
441 predictors. It could also be interesting to normalize the results using other properties or  
442 evaluated parameters. Finally, it will be important to acquire more data and to develop  
443 cultivar-specific calibrations. Indeed, except BW, which had a homogeneous behavior  
444 for all 22 cultivars in the dataset, other physical-mechanical parameters greatly varied  
445 by the cultivar, and general patterns were weak.

446

## 447 **4. Conclusions**

448 This work collected and assessed a large and varied dataset of texture analysis data and  
449 flavonoid content from the analysis of every single berry. It tried to evaluate different  
450 machine-learning algorithms to assess their suitability to model the relationships  
451 between physical-mechanical characteristics of grape and the concentration of skin  
452 flavonoids. The reason for modeling such a relation is that grape berries show changes  
453 in their physical-mechanical properties during ripening, which are variety dependent.

454 The approaches evaluated here (RF and GBM) are state-of-the art techniques, but have  
455 still rarely been used in chemometrics. This work brings an interesting case-study while  
456 also trying to simply and informally explain the way these methods work, starting from  
457 their basic element, RT. It will serve as an introduction and will offer some valuable  
458 insights for food scientists interested in learning more about these techniques or  
459 searching for domain-specific examples of application.

460 Presented models are able to capture a huge portion of the variability in the dataset, as  
461 shown by the reached  $R^2$  and accuracy (given by the RMSE), and they can be useful for  
462 a fast screening of many cultivars, because it does not ask for sample preparation and  
463 extraction, but not yet for fine measurements. It should also be considered that, even if  
464 the number of cultivars in this study was high, universal considerations cannot be  
465 inferred because, as already reported in the discussion, the evolution of physical-  
466 mechanical parameters with ripening could be different across cultivars. Therefore it is  
467 probable that conclusions obtained from this study could be different when developing  
468 models for a single cultivar, especially in the role and importance of the used physical  
469 predictors. It is also highly probable that machine-learning techniques, once applied on  
470 single cultivars or on groups of cultivars presenting a similar evolution of physical-  
471 mechanical properties with the ripening, will reach outstanding performances and could  
472 allow a rapid and accurate estimation of ripening-influenced parameters like TF in grape  
473 berries.

## 474 **Abbreviations**

475 **BW** Berry Weight

476 **CART** Classification And Regression Tree

477 **Co** Cohesiveness

478 **Ch** Chewiness

479 **E<sub>sk</sub>** Skin Young's modulus

480 **F<sub>sk</sub>** Skin break force  
481 **G** Gumminess  
482 **GBM** Gradient Boosting Machine  
483 **H** Hardness  
484 **R** Resilience  
485 **RF** Random Forest  
486 **RFE** Recursive Feature Elimination  
487 **RT** Regression Tree  
488 **S** Springiness  
489 **Sp<sub>sk</sub>** Skin thickness  
490 **SW** Skin Weight  
491 **W<sub>sk</sub>** Skin break energy  
492 **TA** Texture Analysis  
493 **TF** Total Flavonoids Index

494

## 495 **References**

496 Ambroise, C., McLachlan, G.J., 2002. Selection bias in gene extraction on the basis of  
497 microarray gene-expression data. *PNAS* 99, 6562–6566.

498 Barbagallo, M.G., Guidoni, S., Hunter, J.J., 2011. Berry size and qualitative  
499 characteristics of *Vitis vinifera* L. cv. Syrah. *South African J. Enol. Vitic.* 32, 129–  
500 136.

501 Battista, F., Tomasi, D., Porro, D., Caicci, F., Giacosa, S., Rolle, L., 2015. Winegrape  
502 berry skin thickness determination: comparison between histological observations  
503 and texture analysis determination. *Ital. J. Food Sci.* 27, 136-141.

504 Braidot, E., Zancani, M., Petrusa, E., Peresson, C., Bertolini, A., Patui, S., Macri, F.,  
505 Vianello, A., 2008. Transport and accumulation of flavonoids in grapevine (*Vitis*  
506 *vinifera* L.). *Plant Signal. Behav.* 3, 626–632.

507 Breiman, L., 1996. Bias, variance, and arcing classifiers. In Technical report 460  
508 Statistics Department University of California.

509 Breiman, L., 1996. Bagging predictors. *Mach. Learn.* 24, 123–140.

510 Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and  
511 regression trees. *Statistics and probability series*, Wadsworth, Belmont, CA.

512 Breiman, L., 2001. Random forests. *Mach. Learn.* 45, 5–32.

513 Cozzolino, D., 2015. The role of visible and infrared spectroscopy combined with

514 chemometrics to measure phenolic compounds in grape and wine samples.  
515 *Molecules* 20, 726–737.

516 Caruana, R., Munson, A., Alexandru, N.-M. Getting the most out of ensemble selection.  
517 *Int. Conf. Data Min.* 2006; 1–12.

518 Caruana, R., Niculescu-Mizil, A., Crew, G., Ksikes, A., 2011. Ensemble selection from  
519 libraries of models. *Proc. ICML '04* 2011; 34, 1–21.

520 Di Stefano, R., Cravero, M.C., 1991. Metodi per lo studio dei polifenoli dell'uva. *Riv. di*  
521 *Vitic. ed Enol.* 44, 37–45.

522 Efron, B., Tibshirani, R., 1997. Improvements on cross-validation: the 632+ bootstrap  
523 method. *J. Am. Statistic. Assoc.* 92, 438, 548–560.

524 Ferrer-Gallego, R., Hernández-Hierro, Rivas-Gonzalo, J., Escribano-Bailón, M.T., 2011.  
525 Determination of phenolic compounds of grape skins during ripening by NIR  
526 spectroscopy. *LWT-Food Sci. Technol.* 44, 847–853.

527 Freund, Y., Schapire, R., 1997. A decision-theoretic generalization of on-line learning  
528 and an application to boosting. *J. Comput. Syst. Sci.* 55, 119–139.

529 Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine.  
530 *Ann. Stat.* 29, 1189–1232.

531 Giordano M., Zecca O., Belviso S., Reinotti M., Gerbi V., Rolle L., 2013. Volatile  
532 fingerprint and physico-mechanical properties of 'Muscat blanc' grapes grown in  
533 mountain area: a first evidence of the influence of water regimes. *Ital. J. Food Sci.*  
534 25, 329-338.

535 Gonzalez, J., Donoso, W., Sandoval, N., Reyes, M., Gonzalez, P., Gajardo, M., Morales,  
536 E., Neira, A., Razmilic, I., Yuri, J.A., Moore-Carrasco, R., 2015. Apple peel  
537 supplemented diet reduces parameters of metabolic syndrome and atherogenic  
538 progression in ApoE  $-/-$  Mice. *Evidence-Based Complement. Altern. Med.* (2015)

539 González-Neves, G., Charamelo, D., Balado, J., Barreiro, L., Bochicchio, R., Gatto, G.,  
540 Gil, G., Tessore, A., Carbonneau, A., Moutonet, M., 2004. Phenolic potential of  
541 Tannat, Cabernet-Sauvignon and Merlot grapes and their correspondence with  
542 wine composition. *Anal. Chim. Acta* 513, 191–196.

543 Guyon, I., Weston, J., Barnhill, S., Vladimir, V., 2002. Gene selection for cancer  
544 classification using Support Vector Machines. *Mach. Learn.*, 46, 389–422.

545 Harnly, J.M., Doherty, R.F., Beecher, G.R., Holden, J.M., Haytowitz, D.B., Bhagwat,  
546 Gebhardt, S., 2006. Flavonoid content of U.S. fruits, vegetables, and nuts. *J. Agric.*  
547 *Food Chem.* 54, 9966–9977.

548 Hastie, T., Tibshirani, R., Friedman, J., 2009. *The elements of statistical learning: data*  
549 *mining, inference, and prediction*, 2nd ed. Springer Netherlands.

550 Ignat, I., Volf, I., Popa, V.I., 2011. A critical review of methods for characterisation of  
551 polyphenolic compounds in fruits and vegetables. *Food Chem.* 126, 1821–1835.

- 552 Kuhn, N., Guan, L., Dai, Z. W., Wu, B. H., Lauvergeat, V., Gomès, E., Li, S.H., Godoy,  
553 F., Arce-Johnson, P., Delrot, S., 2013. Berry ripening: Recently heard through the  
554 grapevine. *J. Exp. Bot.* 65, 4543-4559.
- 555 Kuhn, M., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, T.,  
556 2014. Caret: classification and regression training. R package version 6.0-37.  
557 URL: <http://CRAN.R-project.org/package=caret>
- 558 Koundouras, S., Marinos, V., Gkoulioti, A., Kotseridis, Y., Van Leeuwen, C., 2006.  
559 Influence of vineyard location and vine water status on fruit maturation of non-  
560 irrigated cv. Agiorgitiko (*Vitis vinifera* L.). Effects on wine phenolic and aroma  
561 components. *J. Agric. Food Chem.* 54, 5077–5086.
- 562 Liaw, A. and Wiener, M., 2002. Classification and regression by randomForest. *R*  
563 *News* 2(3), 18—22.
- 564 Letaief, H., Rolle, L., Gerbi, V., 2008. Mechanical behavior of wine grapes under  
565 compression tests. *Am. J. Enol. Vitic.*, 59, 323–329.
- 566 Lorrain, B., Ky, I., Pechamat, L., Teissedre, P.L., 2013. Evolution of analysis of  
567 polyphenols from grapes, wines, and extracts. *Molecules* 18(1), 1076–100.
- 568 Lourenço, F., Gago, B., Barbosa, R.M., De Freitas, V., Laranjinha, J. LDL isolated from  
569 plasma-loaded red wine procyanidins resist lipid oxidation and tocopherol  
570 depletion. *J. Agric. Food Chem.* 2008; 56, 3798–3804.
- 571 Maloney, K.O., Schmid, M., Weller, D.E., 2012. Applying additive modelling and  
572 gradient boosting to assess the effects of watershed and reach characteristics on  
573 riverine assemblages. *Methods Ecol. Evol.* 3, 116–128.
- 574 Noll, C., Hamelet, J., Matulewicz, E., Paul, J.L., Delabar, J.M., Janel, N., 2009. Effects  
575 of red wine polyphenolic compounds on paraoxonase-1 and lectin-like oxidized  
576 low-density lipoprotein receptor-1 in hyperhomocysteinemic mice. *J. Nutr.*  
577 *Biochem.* 20, 586–596.
- 578 R Core Team, 2014. *R: A language and environment for statistical computing.* R  
579 Foundation for Statistical Computing, Vienna, Austria. URL: [http://www.R-](http://www.R-project.org/)  
580 [project.org/](http://www.R-project.org/).
- 581 Ridgeway, G., 2013. Gbm: generalized boosted regression models. R package  
582 version 2.1. <http://CRAN.R-project.org/package=gbm>
- 583 Ristic, R., Bindon, K., Francis, L.I., Herderich, M.J., Iland, P.G., 2010. Flavonoids and  
584 C13-norisoprenoids in *Vitis vinifera* L. cv. Shiraz: Relationships between grape and  
585 wine composition, wine colour and wine sensory properties. *Aust. J. Grape Wine*  
586 *Res.* 16, 369–388.
- 587 Río Segade, S., Rolle, L., Gerbi, V., Orriols, I., 2008. Phenolic ripeness assessment of  
588 grape skin by texture analysis. *J. Food Compos. Anal.* 21, 644–649.
- 589 Río Segade, S., Orriols, I., Giacosa, S., Rolle, L., 2011a. Instrumental texture analysis  
590 parameters as winegrapes varietal markers and ripeness predictors. *Int. J. Food*

591 Prop. 14, 1318–1329.

592 Río Segade, S., Soto Vázquez, E., Orriols, I., Giacosa, S., Rolle, L., 2011b. Possible  
593 use of texture characteristics of winegrapes as markers for zoning and their  
594 relationship with anthocyanin extractability index. *Int J Food Sci Tech.* 46,  
595 386–394.

596 Río Segade, S., Giacosa S., Gerbi V., Rolle L., 2011c. Berry skin thickness as main  
597 texture parameter to predict anthocyanin extractability in winegrapes. *LWT-Food*  
598 *Sci. Technol.* 44, 392–398.

599 Rolle, L., Río Segade, S., Torchio, F., Giacosa, S., Cagnasso, E., Marengo, F., 2011.  
600 Influence of grape density at harvest date on changes in phenolic composition,  
601 phenol extractability indices, and instrumental texture properties during ripening.  
602 *J. Agric. Food Chem.* 59, 8796–8805.

603 Rolle, L., Torchio, F., Lorrain, B., Giacosa, S., Río Segade, S., Cagnasso, E., Gerbi, V.,  
604 Teissedre, P.L., 2012. Rapid methods for the evaluation of total phenol content and  
605 extractability in intact grape seeds of Cabernet-Sauvignon: Instrumental  
606 mechanical properties and FT-NIR spectrum. *J. Int. Sci. Vigne Vin* 46, 29-40.

607 Rolle, L., Torchio, F., Ferrandino, A., Guidoni, S., 2012b. Influence of wine-grape skin  
608 hardness on the kinetics of anthocyanin extraction. *Int. J. Food Prop.* 15, 249-261.

609 Rolle, L., Giacosa, S., Torchio, F., Perenzoni, D., Río Segade, S., Gerbi, V., Mattivi, F.,  
610 2013. Use of instrumental acoustic parameters of winegrape seeds as possible  
611 predictors of extractable phenolic compounds. *J. Agric. Food Chem.* 61, 8752–  
612 8764.

613 Scott, I.M., Lin, W., Liakata, M., Wood, J.E., Vermeer, C.P., Allaway, D., Ward, J.L.,  
614 Draper, J., Beale, M.H., Corol, D.I., Baker, J.M., King, R.D., 2013. Merits of  
615 random forests emerge in evaluation of chemometric classifiers by external  
616 validation. *Anal. Chim. Acta* 801, 22–33.

617 Therneau, T., Atkinson, B., Ripley, B., 2015. rpart: Recursive Partitioning and  
618 Regression Trees. R package version 4.1-9. URL: [http://CRAN.R-](http://CRAN.R-project.org/package=rpart)  
619 [project.org/package=rpart](http://CRAN.R-project.org/package=rpart)

620 Strobl, C., Boulesteix, A.-L., Zeileis, A., Hothorn, T., 2007. Bias in random forest  
621 variable importance measures: illustrations, sources and a solution. *BMC*  
622 *Bioinformatics*, 8, 25.

623 Yao, L.H., Jiang, Y.M., Shi, J., Tomás-Barberán, F. A., Datta, N., Singanusong, R.,  
624 2004. Flavonoids in food and their health benefits. *Plant foods Hum. Nutr.* 59,  
625 113–122.

626 Zouid, I., Siret, R., Jourjon, F., Mehinagic, E., Rolle, L. 2013. Impact of grapes  
627 heterogeneity according to sugar level on both physical and mechanical berries  
628 properties and their anthocyanins extractability at harvest. *J. Texture Stud.* 44, 95–  
629 103.

**Table 1.** Descriptive statistics for the global dataset of physical-mechanical properties and total flavonoid content composed of 792 berries from 22 red wine grape cultivars.

	<b>Min.</b>	<b>Max.</b>	<b>Mean</b>	<b>Median</b>	<b>Var.</b>
<b>BW (g)</b>	1.31	4.13	2.57	2.54	0.331
<b>F<sub>sk</sub> (N)</b>	0.18	1.01	0.65	0.66	0.024
<b>W<sub>sk</sub> (mJ)</b>	0.11	1.47	0.64	0.64	0.040
<b>E<sub>sk</sub> (N mm<sup>-1</sup>)</b>	0.15	0.47	0.29	0.29	0.004
<b>Sp<sub>sk</sub> (μm)</b>	122.33	315.00	209.69	201.00	1599.699
<b>SW (g)</b>	0.12	0.45	0.26	0.25	0.005
<b>H (N)</b>	1.36	5.68	3.19	3.05	0.655
<b>Co (adimens.)</b>	0.60	0.89	0.79	0.80	0.003
<b>G (N)</b>	1.13	4.45	2.51	2.43	0.377
<b>S (mm)</b>	1.62	3.00	2.43	2.48	0.069
<b>Ch (mJ)</b>	1.86	11.89	6.21	6.01	3.193
<b>R (adimens.)</b>	0.31	0.51	0.45	0.41	0.001
<b>TF (mg g<sup>-1</sup>)</b>	0.69	12.87	2.60	2.07	4.164

**Table 2.** Pearson correlations for the global dataset of physical-mechanical properties, total flavonoid and total soluble solids content obtained from 792 berries from 22 red wine grape cultivars

	BW	F <sub>sk</sub>	W <sub>sk</sub>	E <sub>sk</sub>	Sp <sub>sk</sub>	SW	H	Co	G	S	Ch	R	TSS
<b>BW</b>													
<b>F<sub>sk</sub></b>	-0.13 *												
<b>W<sub>sk</sub></b>	0.04 ***	0.92 ***											
<b>E<sub>sk</sub></b>	-0.29 ***	0.81 ***	0.54 ***										
<b>Sp<sub>sk</sub></b>	0.06	-0.07	-0.12	0.03									
<b>SW</b>	0.76 ***	-0.21 ***	-0.11	-0.25 ***	0.46 ***								
<b>H</b>	0.17 **	0.54 ***	0.30 ***	0.75 ***	0.06	0.14 *							
<b>Co</b>	0.03	-0.11	-0.06	-0.17 **	-0.01	0.08	-0.22 ***						
<b>G</b>	0.19 **	0.53 ***	0.31 ***	0.73 ***	0.05	0.16 **	0.97 ***	0.01					
<b>S</b>	0.88 ***	-0.06	0.02	-0.10	-0.06	0.62 ***	0.31 ***	0.15 *	0.36 ***				
<b>Ch</b>	0.44 ***	0.42 ***	0.26 ***	0.57 ***	0.04	0.35 ***	0.91 ***	0.05	0.95 ***	0.61 ***			
<b>R</b>	0.06	-0.23 ***	-0.15 *	-0.29 ***	-0.20 **	0.01	-0.38 ***	0.58 ***	-0.24 ***	0.25 ***	-0.14 *		
<b>TSS</b>	-0.11	-0.01	-0.02	0.01	-0.02	0.00	-0.05	0.04	-0.05	-0.11	-0.07	0.04	
<b>TF</b>	-0.59 ***	0.13 *	-0.03	0.34 ***	0.34 ***	-0.18 **	0.00	-0.13 *	-0.03	-0.53 ***	-0.20 **	-0.19 **	0.03

\*\*\* =  $p$ -value < 0.001, \*\* =  $p$ -value < 0.01, \* =  $p$ -value < 0.05, . =  $p$ -value < 0.1

**Table 3.** Relative influence of each predictor in all tested algorithms. Influence was scaled between 0 and 100 to allow an easier comparison.

	<b>RT</b>	<b>GBM</b>	<b>RF</b>
<b>C</b>	0.00	6.20	4.23
<b>R</b>	17.00	0.00	0.00
<b>W<sub>sk</sub></b>	51.25	1.31	0.98
<b>Ch</b>	35.41	2.79	15.80
<b>Sp<sub>sk</sub></b>	56.41	3.49	3.03
<b>E<sub>sk</sub></b>	100.00	4.77	26.77
<b>BW</b>	42.07	100.00	100.00

---

**Table 4.** Results ( $R^2$  and Root Mean Squared Error, RMSE) of the tested algorithm on the training set, the external validation (test set) and the internal validation (10 repetitions of 5-fold Cross-Validation, CV). For cross-validation estimations of the standard deviation for both metrics are also shown. RMSE results are expressed in  $\text{mg g}^{-1}$  berry, lower the error better the performances of the model.

	<b>Train RMSE</b>	<b>Train R2</b>	<b>Test RMSE</b>	<b>Test R2</b>	<b>CV RMSE</b>	<b>CV R2</b>	<b>CV RMSE SD</b>	<b>CV R2 SD</b>
<b>RT</b>	0.817	0.849	0.951	0.752	1.286	0.650	0.198	0.101
<b>RF</b>	0.419	0.965	0.729	0.836	1.071	0.754	0.148	0.063
<b>GBM</b>	0.364	0.971	0.745	0.836	1.074	0.753	0.147	0.058

**Table 3.** Relative influence of each predictor in all tested algorithms. Influence was scaled between 0 and 100 to allow an easier comparison.

	<b>RT</b>	<b>GBM</b>	<b>RF</b>
<b>C</b>	0.00	6.20	4.23
<b>R</b>	17.00	0.00	0.00
<b>W<sub>sk</sub></b>	51.25	1.31	0.98
<b>Ch</b>	35.41	2.79	15.80
<b>Sp<sub>sk</sub></b>	56.41	3.49	3.03
<b>E<sub>sk</sub></b>	100.00	4.77	26.77
<b>BW</b>	42.07	100.00	100.00

---

**Table 4.** Results ( $R^2$  and Root Mean Squared Error, RMSE) of the tested algorithm on the training set, the external validation (test set) and the internal validation (10 repetitions of 5-fold Cross-Validation, CV). For cross-validation estimations of the standard deviation for both metrics are also shown. RMSE results are expressed in  $\text{mg g}^{-1}_{\text{berry}}$ , lower the error better the performances of the model.

	<b>Train RMSE</b>	<b>Train R2</b>	<b>Test RMSE</b>	<b>Test R2</b>	<b>CV RMSE</b>	<b>CV R2</b>	<b>CV RMSE SD</b>	<b>CV R2 SD</b>
<b>RT</b>	0.817	0.849	0.951	0.752	1.286	0.650	0.198	0.101
<b>RF</b>	0.419	0.965	0.729	0.836	1.071	0.754	0.148	0.063
<b>GBM</b>	0.364	0.971	0.745	0.836	1.074	0.753	0.147	0.058

### Figure Captions:

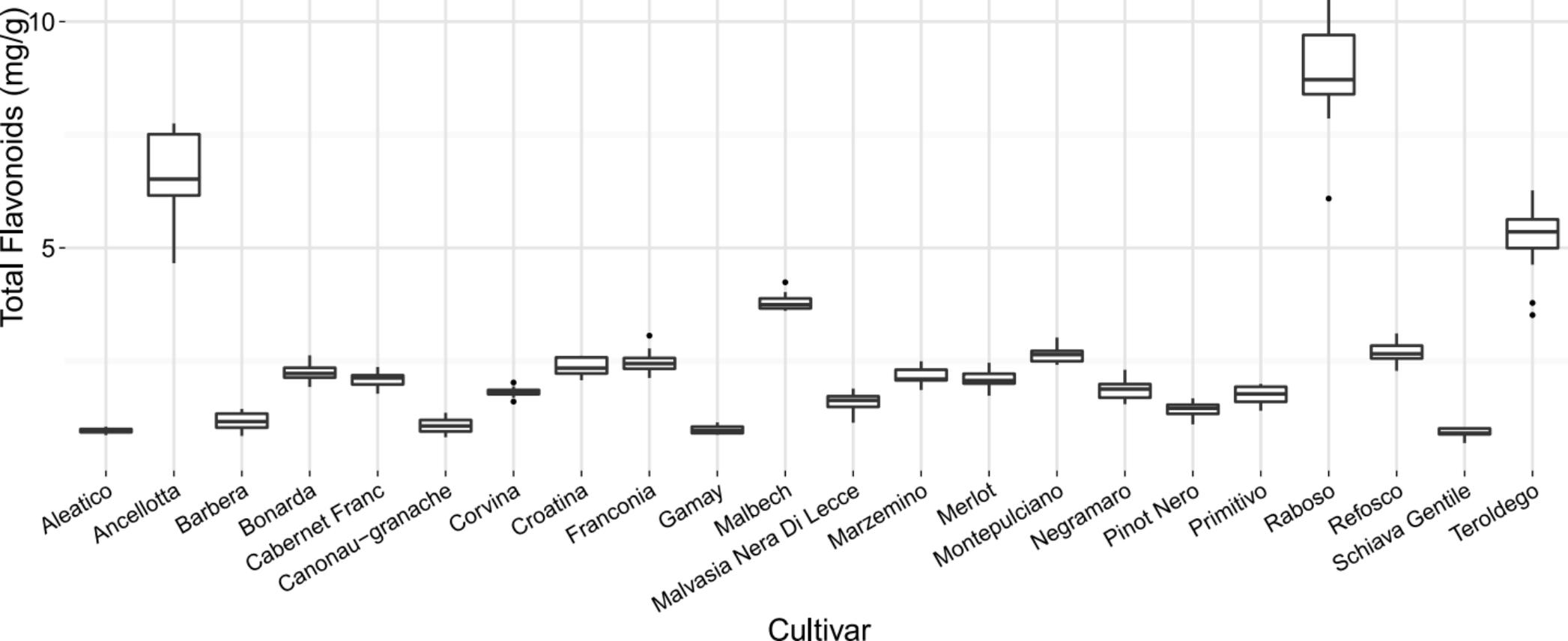
**Fig. 1. Flavonoid content for all cultivars in the experiment.** Flavonoid content ( $\text{mg g}^{-1}_{\text{berry}}$ ) for the 22 red wine grape cultivars used in the experiment.

**Fig. 2. Best regression tree model to predict total flavonoids from texture analysis data** Figure representing the tuned RT on the training dataset. The whole training dataset is recursively splitted in two parts, according to the predictor that allows the greatest reduction in the residual sum of squares. The selected predictor at each split is shown inside the ellipse, while just under there is the rule used for splitting which is a corresponding predictor value. Here numbers are expressed in the original measure unit of each predictor, readers are kindly referred to table 1 for the complete list. The optimal number of splits according to the results of the cross-validation procedure was equal to 7.

**Fig. 3. Results of the machine-learning techniques on train and test data-sets.** Relationships between observed and predicted TF data ( $\text{mg g}^{-1}$ ) over the train test and the test set used as external validation for all algorithms. Solid black line is the identity line, while the dashed gray line is a linear regression (OLS) applied to the data and the filled gray region is its 95% confidence interval.

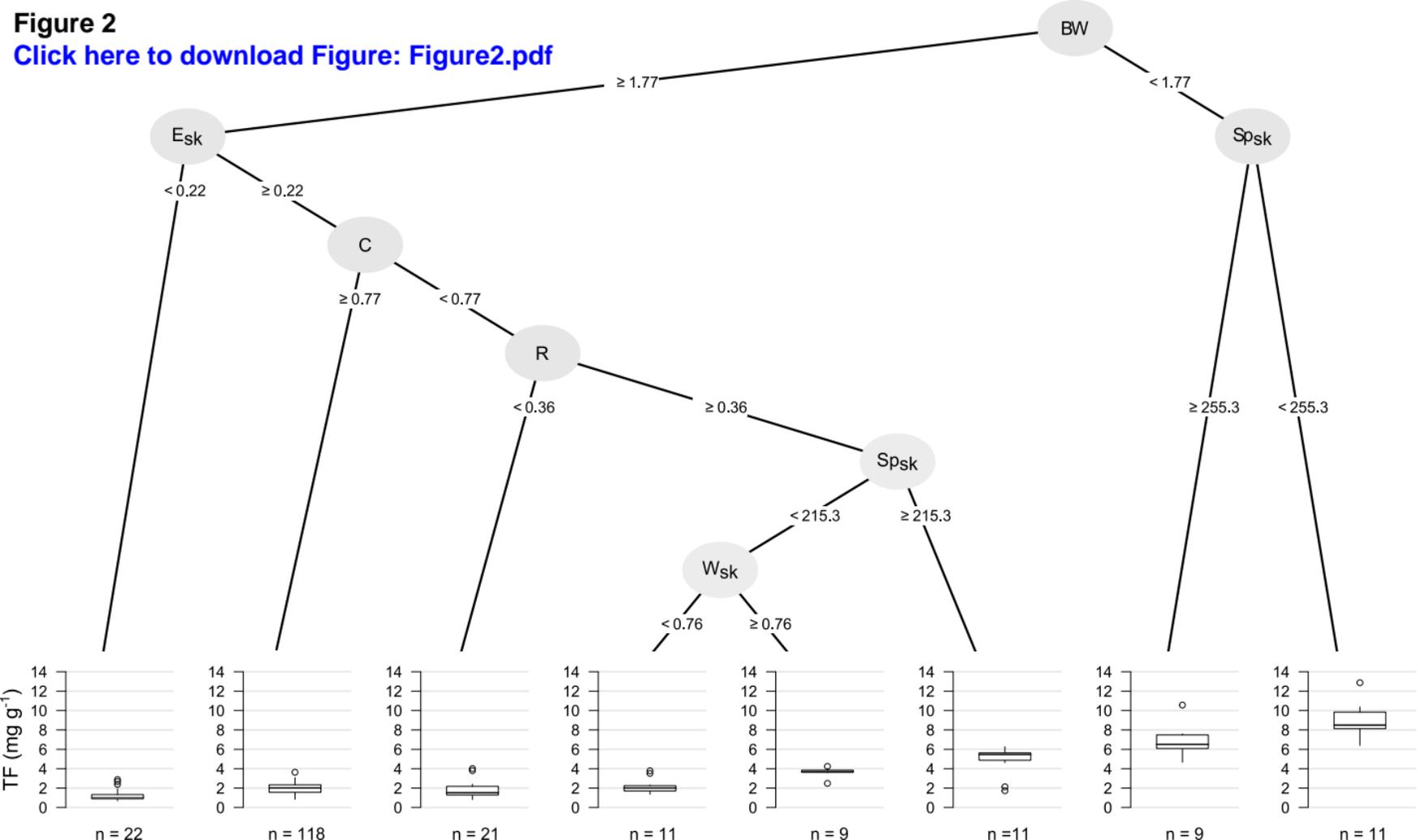
**Figure 1**

[Click here to download Figure: Figure1.pdf](#)



**Figure 2**

[Click here to download Figure: Figure2.pdf](#)



**Figure 3**  
[Click here to download Figure: Figure3.pdf](#)

