CENTENARY OF THE OIV
45th WORLD CONGRESS OF VINE AND WINE
14-18 OCTOBER 2024 FRANCE-DIJON
OIV
Available on IVES Conference Series vine & wine
DOI : https://doi.org/10.58233/V9H1atSf

Multi-mineral profile and AI: wine authentication and identification

Leticia Sarlo[1,2], Coraline Duroux[2], Théodore Tillement[2], François Lux[1,3], Olivier Tillement[1]

[1] Institut Lumière-Matière, UMR 5306, Université Claude Bernard Lyon 1-CNRS, Université de Lyon, Villeurbanne Cedex 69100, France
[2] M&Wine, 305 rue des Fours, 69270 Fontaines Saint Martin, France
[3] Institut Universitaire de France (IUF), Paris

Abstract. The use of a wine's mineral profile (MWP) as a stable and distinctive fingerprint can revolutionize wine authentication. By employing inductively coupled mass spectrometry, we assessed the concentration of ^{11}B, ^{23}Na, ^{24}Mg, ^{27}Al, ^{28}Si, ^{31}P, ^{34}S, ^{35}Cl, ^{39}K, ^{43}Ca, ^{45}Sc, ^{47}Ti, ^{51}V, ^{52}Cr, ^{55}Mn, ^{56}Fe, ^{59}Co, ^{60}Ni, ^{63}Cu, ^{66}Zn, ^{75}As, ^{79}Br, ^{85}Rb, ^{88}Sr, ^{89}Y, ^{90}Zr, ^{93}Nb, ^{111}Cd, ^{118}Sn, ^{127}I, ^{133}Cs, ^{137}Ba, ^{139}La, ^{140}Ce, ^{141}Pr, ^{146}Nd, ^{182}W, ^{205}Tl, ^{208}Pb, ^{238}U with minimal sample preparation. MWP is shaped at least by soil composition and winemaking techniques. This last factor may overshadow the terroir when defining this profile, hindering origin-related information extraction. However, the integration of artificial intelligence (AI) presents itself as a solution. More than 19,000 MWPs were analysed, laying the groundwork for a machine-learning algorithm to assess wine's country, region and main grape variety. Extreme Gradient Boosting was employed, exceeding scores of areas under the receiving operating characteristic curves of 0.9 for country, French wine region and grape variety classification. This performance enables a specificity of wine authentication up to 99%, demonstrating the potential of combining AI and MWP analysis. This study highlights the importance of comprehensive MWP datasets for advancing AI applications in origin verification, offering a promising tool for the wine industry to enhance security and consumer trust.

1. Introduction

Wine is a popular beverage with an important economic impact worldwide. It can be subject to fraudulent practices [1], which has led to an increasing interest in techniques to assure its traceability and authentication [2]. The three prevalent approaches in the literature are DNA analysis [3], determination of organic compounds, i.e., polyphenols and volatiles compounds, or mineral elements [1], both of which may be considered fingerprints of a wine.

For the identification of grape varieties, DNA analysis has been studied as an avenue of research. It selects characteristic sequences of the variety on the genetic material recovered from the beverage [4]. However, studies suggest that only wines under a year post-bottling may be analysed, as the DNA degrades over time, which hinders sample identification [4,5]

Isotope measurement gives information about wine's both organic and inorganic profile. Isotope ratio-mass spectrometry (IRMS) [6] and liquid chromatography-IRMS [7] are two of the techniques used for this quantification. Nonetheless, the preparation of samples is expensive and time-consuming, impeding the creation of a comprehensive database.

Many analytical techniques can be used to determine the organic compound profile, such as gas chromatography-mass spectrometry [8], high-performance liquid chromatography-diode array detection [9] and nuclear magnetic resonance [10]. Even though this profiling is widely used, these molecules are sensitive to oxidation, ageing of the wine and its storage conditions [11], hampering the comparison of the same sample over the years.

To address the limitations imposed by the organic profile, the measurement of the elemental inorganic content has been explored as an alternative for verifying the origin of wine. [12,13]. These elements can be divided into three main categories depending of their concentrations. The first category is referred to as macro elements, comprising for example K, Na, Mg and Ca whose concentrations are in the range of 10 to 1000 mg/kg. The second category is microelements, with concentrations in the range of 0.1 to 10 mg/kg, including for example Fe, Cu or Mn [12]. Both categories assemble

elements essential to plant growth and development as well as non-essential [14] while the third category comprises non-essential elements such as rare earth, Ag, Pb or U, with a range of concentrations between 0.1 and 1000 µg/kg [12]. Some factors to influence their concentrations are grape maturity, winemaking practices, soil type and properties and the composition of the mother rock [14,15].

Because of these wide ranges of concentrations as well as the variety of elements to be measured, inductively coupled plasma-mass spectrometry (ICP-MS) is widely used [1]. In the literature, it is used in association with classical chemometrics analysis [16,17], as well as machine learning (ML) classification algorithms [18,19].

Although many efforts have been made to develop authentication methods, current literature often focuses on specific parameters like individual countries [13], regions [20], wine appellations [18], or grape varieties [12,19]. This narrowing frequently results in limited sample sizes, typically staying under 100 samples, which constrains the general applicability of the findings as well as their statistical significance.

As the concern for wine origin verification increases, the need to conciliate the cost with the reliability of the analysis becomes of utmost importance. In this study, we show a quick and cost-efficient ICP-MS semi-quantitative (SQ) method of wine analysis, quantifying 40 elements, who integrate the Mineral Wine Profile (MWP), on over 200 samples per day. The resulting MWP constituted an oenotheque comprising of more than nineteen thousand samples, which were then used to train ML algorithms to verify a wine's origin. We aim to establish a timeless tool which will be able to fully authenticate a wine after the analysis of only a 30 mL sample.

2. Materials and methods

2.1. Reagents and materials

A total of 19431 wines originating from commerce as well as international contest were analysed in this study. More details about their origin are described in

Figure 1. For each wine, approximately 30 mL were collected in certified metal-free tubes (VWR®). The sample was diluted in the proportion 1:3 with nitric acid 1% (v/v), prepared with ultrapure water (MilliQ®, 18.2 mΩ.cm) and nitric acid Suprapur® grade (69% (v/v), Roth), and 10 µg/L of indium standard solution, prepared with 1000 mg/L indium standard in HNO_3 4% (v/v) purchased from SCP Science. This dilution allows the storage of wine in acid conditions, preserving the MWP over time by hindering precipitation and adsorption by the tube walls.

Before analysis, a second dilution of 1:5 is done with HNO_3 1% (v/v), dilution of nitric acid Suprapur® grade (69% (v/v), Roth) with ultrapure water (MilliQ®, 18.2 mΩ.cm). This dilution factor (1:15) has been shown to minimize matrix effects [21].

A tuning solution containing 1 µg/L of Ce, Co, Li, Tl, and Y in 2% HNO_3 (v/v) (Agilent Technologies) is used in the beginning of the analysis for mass calibration and performance validation. A multi-element standard (VWR®, reference 85006.186) with 100 mg/L of Ag, As, B, Ba, Be, Bi, Ca, Cd, Co, Cr, Cu, Fe, K, Li, Mg, Mn, Mo, Na, Ni, Pb, Sb, Se, Sr, Ti, Tl, V, and Zn, in 5% HNO_3 (v/v) was diluted to obtain the semi-quantitative calibration standard of concentration 20 µg/L. To control and validate this calibration, a wine of commercial origin is used as reference. It is prepared following the same procedure used for the wine sample described previously and will be referend as control wine in this study.

2.2. Mineral Wine Profile Determination

The MWPs were determined by ICP-MS analyses, performed between June 2022 and March 2024 at M&Wine, Lyon-France, and the Institut de Sciences Analytiques, Université Claude Bernard Lyon 1, using different quadrupole ICP-MS equipment. Most of the measurements were performed using Agilent Technologies simple quadrupole-ICP-MS 7850 with an integrated autosampler, SPS 4. A micromist nebulizer was employed for all measurements. To minimize polyatomic interferences, the collision cell was set to Helium mode for all elements and the flow rate was 5 mL/min. For the operating conditions, the following parameters were set: 1550 W forward power, 1 L/min carrier gas flow, 15 L/min plasma gas flow and 1 L/min auxiliary gas flow. A tuning solution was employed before each analysis to adjust the remaining parameters to optimize the signal. Analysis of a control wine is done at the beginning, middle and end of each sample sequence. Blanks and the 28-element standard are reanalysed every 40 samples.

The SQ analysis enabled the determination of the elemental concentrations using the 28-element standard solution. The following elements were analysed, with 100 sweeps and one replicate: ^{11}B, ^{23}Na, ^{24}Mg, ^{27}Al, ^{28}Si, ^{31}P, ^{34}S, ^{35}Cl, ^{39}K, ^{43}Ca, ^{45}Sc, ^{47}Ti, ^{51}V, ^{52}Cr, ^{55}Mn, ^{56}Fe, ^{59}Co, ^{60}Ni, ^{63}Cu, ^{66}Zn, ^{75}As, ^{79}Br, ^{85}Rb, ^{88}Sr, ^{89}Y, ^{90}Zr, ^{93}Nb, ^{111}Cd, ^{115}In, ^{118}Sn, ^{127}I, ^{133}Cs, ^{137}Ba, ^{139}La, ^{140}Ce, ^{141}Pr, ^{146}Nd, ^{182}W, ^{205}Tl, ^{208}Pb, ^{238}U. All but ^{115}In, the internal standard, constitute the MWP.

2.3. Statistical analysis and sample classification

Before any analysis, values lower than the limit of quantification (LOQ) were replaced for 10^{-4}. For the columns with less than 100 samples with values lower than LOQ, the sample was excluded from the dataset. Praseodymium, neodymium, scandium and samarium had more than 60% of values lower than LOQ so they were considered outliers and were not included in the analysis. Silicon was not included in the analysis because of detection limitations.

One-way ANOVA was performed for each label in the country, French region and principal grape variety, when the number of samples was greater than 50.

In a previous study [22], our group has seen a natural separation in data and its classification is possible using ML algorithms. Sample classification was performed as described in the aforementioned study with changes. Labels with less than 50 samples present in the database were not used for model training nor testing. Samples with unknown labels were also not employed.

As done previously, the following six models were compared by means of the area under the receiver operating characteristic curve (AUC). This metric indicates the likelihood that a classifier will correctly rank a randomly selected positive instance above a randomly selected negative instance. Random guessing would give an AUC of 0.5, whereas a perfect classifier achieves 1.0 [23]. All of the models were developed in Python, version 3.9.19.

Extreme Gradient Boosting (XGB) is a boosting ensemble learning algorithm that combines the predictions of multiple decision trees to achieve the final classification [24]. It was implemented using the *XGBoost* library [24].

Artificial Neural Networks (ANN) is a technique comprising interconnected nodes organized in different layers, the first being the input layer, the last the output layer and in between both there are one or more intermediate layers [18]. It was created using the *TensorFlow* library [25].

The following ML models were implemented using the *scikit-learn* library [26].

Random Forest (RF) consists of a group of tree predictors. Each tree is based on random vector values that are sampled independently and follow the same distribution for all trees in the forest [27]

k-Nearest Neighbours (k-NN) is a method where given a set of known classes and an unknown sample, its label will be determined by comparing to the most frequently occurring label among its k-Nearest Neighbours [28]

Support Vector Machines (SVM) is based on the mapping of input vectors to a high-dimensional feature space where a hyperplane is constructed to separate data points into distinct classes [29].

Logistic Regression (LR) is a binary classifier which uses a sigmoid function to map predictions to probabilities between 0 and 1. The classification is made based on whether the output exceeds a predefined threshold [30].

No hyperparameter tuning was done during the comparison between algorithms. The dataset was divided in train and test sets in the proportion 80:20. The model chosen was XGBoost. Hyperparameter tuning was performed using grid search and they are presented in Table 1.

Table 1. Hyperparameters used in XGBoost model for the classification of country, French wine producing region and grape variety.

Hyperparameter	Country	Region	Variety
n_estimators	200	200	90
max_depth	8	8	10
reg_lambda	2	2	1.5
learning_rate	0.1	0.1	0.1
gamma	0	0	0
colsample_bytree	0.6	0.6	0.6

All the others were the standard parameters following the documentation [24]. The database was divided in train, validation and test sets in the proportion 70:15:15, respectively, in a random and stratified fashion. To improve the training of the model, a space filling technique, Synthetic Minority Over-sampling Technique (SMOTE), was employed after split only on the training set.

This approach uses the creation of synthetic samples in order to over-sample the minority class [31]. This allows for better overall performance of the model, by balancing the class distributions.

The metrics chosen to measure model performance were accuracy, ratio of correctly classified samples, and specificity, probability of a negative sample being predicted as negative by the algorithm.

3. Results and discussion

3.1. Elemental composition analysis in wine samples

This research was conducted using a database comprising of 19431 MWP from wines deriving from international competitions as well as from commercial origin. They were distributed in 52 different countries, 290 wine producing regions and 264 different main grape varieties. As multivarietal wines were present in the database, when referencing the main grape variety in this study, it is the variety present in the highest percentage. This repartition is illustrated in

Figure 1.

Because of the diverse origin and varietal composition of the samples, the elemental concentration had great variability. This behaviour may be also influenced by the different vinification techniques employed and environmental conditions. When comparing to the existing literature, variations within the same vineyard [12] and country [32] were recorded. The values found in our set are in agreement to the ranges offered in literature for macro, micro and trace elements. Some of these ranges are show in Figure 2, where the difference of concentration between categories is also explored.

In Figure 2, the difference of concentration between classes for a same element is visible. For readability, not all elements were shown but this difference was confirmed

for all elements by one-way ANOVA. These elements were chosen because they exemplify the three groups of mineral elements present in wine: macro, micro and trace elements.

In regard to macronutrients, there are essential and non-essential elements for plant growth. Potassium, calcium and magnesium are part of the first group, with their origin associated to the soil, as the rootstock distribute throughout the plant until the grapes [12], but also with cultivation practices [33]. Sodium is a non-essential macronutrient with the origin associated with soil and proximity to the sea [34]. They were found to contribute in the origin classification of French [33], Spanish [35] and Argentinean [34] wines as well as improving distinction of groups when associated to rare earth elements [36].

Iron, manganese, nickel, copper and zinc are part of the essential micronutrients and chromium, cobalt and aluminium, the non-essentials. They have a natural origin, soil, but also are influenced by anthropogenic sources [15,37]. One example is the material which the wine may have contact with, stainless steel, bronze, or brass may influence elemental content differently [37]. Pesticides and phytosanitary products are another source as they may contain Cu, Zn, Ni, Mn [38]. These elements may also interfere in wine quality as they may be responsible for haze formation, such as Cu, Fe, Mn, Al, Ni and Zn, with the first tree also participating in reactions during maturation that impact acetaldehyde content [38]. It is also hypothesized, for concentration larger than 1mg/L, that Cu may contribute to a metallic taste in the final product [37].

Lastly, two trace elements were represented in this plot, lanthanum and yttrium. They are part of the rare earth elements which is a group considered as markers of origin and variety, therefore used in their classification [39–42]. Their content is extremally impacted by the use of bentonite [43,44], used for wine clarification and stabilization during its production, and they alone do not outperform the combination of macro, micro and other trace elements [36].

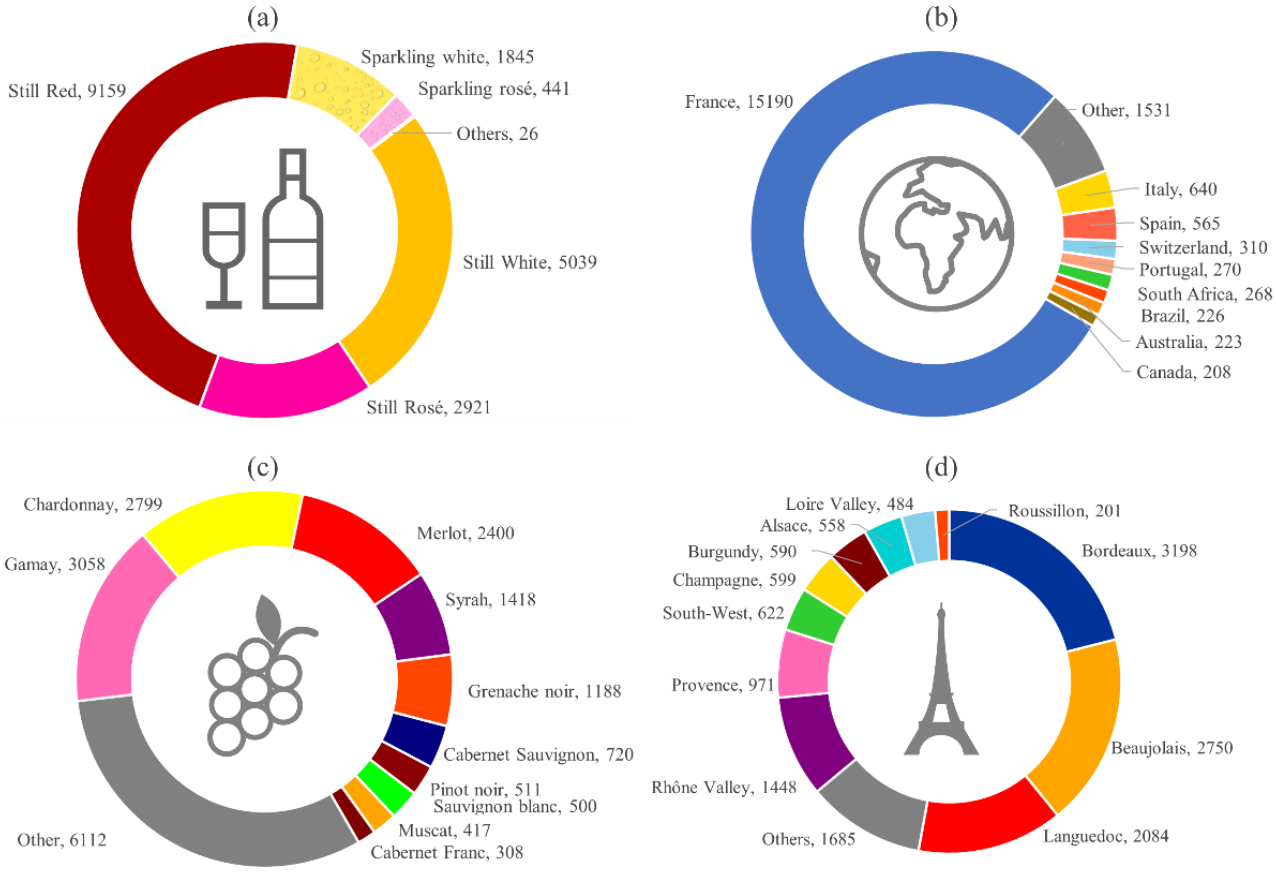

Figure 1. Repartition of the 19431 analysed wines in the databased used in this study. It was done based on wine type (a), country (b), grape variety (c), and French wine region (d). Unlabelled wines as well as labels with less than 200 occurrences are grouped in the "Others" label. For the coloured version, the reader is invited to check the online version of this document.

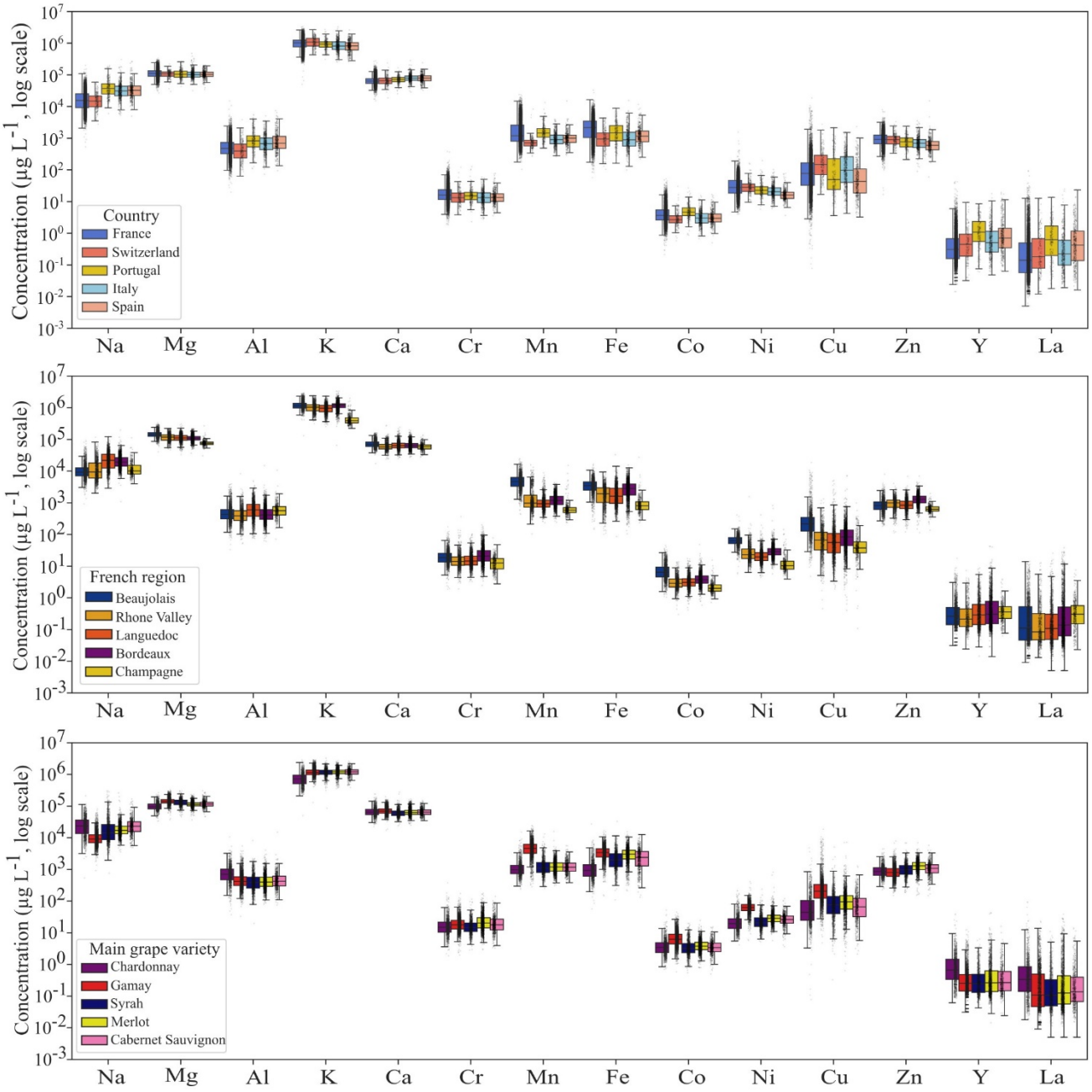

Figure 2. Distribution of the log-transformed concentration for Na, Mg, Al, K, Ca, Cr, Mn, Fe, Co, Ni, Cu, Zn, Y and La. The categories compared are, from top to bottom, country of origin, with labels France, Spain, Italy, Switzerland and Portugal; French region, labels being Beaujolais, Rhone Valley, Champagne, Languedoc and Bordeaux; and main grape variety, labels are Chardonnay, Gamay, Syrah, Merlot and Cabernet Sauvignon. The whiskers in the boxplot correspond to 1.5 the interquartile range, the horizontal line in the box is the median. For the coloured version, the reader is invited to check the online version of this document.

All of these factors contribute to the differences presented in Figure 2. It is evident, when comparing categories per element in each graph, that different elements will distinguish a category from the others. Some clear examples are the region of Champagne, represented in yellow on the second graph and the Gamay variety, coloured red, on third graph. These are two distinct categories as the Champagne is mainly composed of sparkling wines with defining and regulated viticultural practices, and the Gamay is a variety specific to the Beaujolais region. These particularities only contribute to the separation of these categories and shed light to the previous origins of minerals listed.

Additionally, when comparing variety and region, it is also important to bring the distinction of winemaking methods between red, white and rosé wines. This is indicative of different metal contents which has already been explored in other studies [45,46]. In Figure 2, Gamay, a typical red variety shows itself less concentrated in Na, but more in Mn, Fe, Co, Ni and Cu when compared to the other groups. Chardonnay, usually employed in white wine, is more concentrated in Na, Al, and Y while less in Mg, K and Fe.

Because of these differences in concentration, it was envisioned to use the MWP to distinguish the categories of wines in our database. This has been made for smaller datasets [18,19,34–36,41,47] with various chemometrics and ML methods, i.e. neural networks [18], random forest [19] and orthogonal partial least squares discriminant analysis [13]. The variety of techniques that may be employed motivates an initial selection of a ML algorithm before the development of the model.

3.2. Machine learning algorithm selection

As to assure the most adequate ML strategy, the AUC of the binary classification for each label was calculated. This was repeated 10 times. The results were averaged and are shown in

Table 2.

Table 2. Comparison between six ML models tested for classifying wine origin and grape variety. The average AUC score was computed for 10 iterations for each label. The best performing model for each category, based on the highest AUC, is in bold.

Model	Mean AUC		
	Country	French Region	Grape Variety
RF	0.954	0.952	0.872
k-NN	0.846	0.860	0.740
SVM	0.963	0.949	0.891
LR	0.935	0.911	0.863
XGB	**0.980**	**0.970**	**0.923**
ANN	0.935	0.910	0.859

These values show promising results for all of the models, as all were close to 1. XGB presents itself as the most adapted model for the classification of our dataset and therefore was optimized, by means of hyperparameter tuning, for each category classification.

3.3. Origin and variety classification

The model's performance in the validation and tests sets for each category are shown in

Table 3,

Table 4 and

Table 5. The metrics chosen were accuracy (Acc) and specificity (Spc). The model presents great results for the binary classification of countries and regions, reaching up to 99% of accuracy for multiple labels during validation as well as testing. This attest for the robustness of the model for predicting a wine's origin.

In regard to country classification, it is remarkable that both metrics are always superior to 90%. Other studies have attempted country classification notably Forina *et al.*, which also had the largest sample size to our knowledge (1188 wines) [48]. The ML algorithm used in our study had more samples as well as outperformed theirs when comparing the specificity, mean of 90% versus 80%. This attest to the reliability of our model to identify a negative sample.

Table 3. Accuracy (Acc) and specificity (spc) of the prediction of the validation (val) and test sets for country classification using XGB. The values presented are the mean of 10 iterations.

Country	**Acc_{val}**	**Acc_{test}**	**Spc_{val}**	**Spc_{test}**
France	94.4%	93.9%	91.2%	93.0%
Italy	96.8%	92.2%	97.5%	92.2%
Spain	97.9%	92.3%	98.3%	92.4%
Switzerland	99.1%	97.0%	99.3%	97.1%
Portugal	98.2%	94.9%	98.5%	94.9%
South Africa	99.2%	95.8%	99.4%	95.9%
Australia	99.7%	98.4%	99.8%	98.5%
Brazil	99.8%	98.7%	99.9%	98.8%
Canada	99.6%	99.0%	99.7%	99.0%
Romania	99.1%	94.5%	99.3%	94.5%
Moldova	99.4%	96.8%	99.5%	96.8%
Hungary	99.1%	95.7%	99.2%	95.8%

Greece	98.9%	93.1%	99.2%	93.2%
Bulgaria	99.1%	95.5%	99.3%	95.6%
Austria	99.0%	94.8%	99.3%	94.8%
Germany	99.1%	90.9%	99.3%	91.0%
Belgium	99.6%	91.5%	99.7%	91.5%
Slovakia	99.1%	91.7%	99.3%	91.7%

Within a country, there may be different regions known for their wine production. This leads to wines in a same country having different mineral fingerprints, as they also correlate to the region of production. This is a subject explored in plenty in the literature in the distinction of different countries regions, such as Italy [49], Romania [47] and China [50].

To delve deeper in this subject, we have also classified French wines coming from 13 different regions. Remarkable accuracy and specificity are achieved, surpassing 87% for all datasets. These values are in agreement with those found by Wu *et al.* [33].

Table 4. Accuracy (Acc) and specificity (spc) of the prediction of the validation (val) and test sets for French region classification using XGB. The values presented are the mean of 10 iterations.

Region	**Acc_{val}**	**Acc_{test}**	**Spc_{val}**	**Spc_{test}**
Bordeaux	95.8%	95.6%	96.0%	95.5%
Beaujolais	98.1%	97.9%	98.5%	97.9%
Languedoc	92.4%	90.2%	93.7%	90.1%
Rhone Valley	94.3%	92.1%	95.6%	92.4%
Provence	94.9%	91.1%	95.8%	91.0%
Southwest France	94.5%	87.1%	95.8%	87.1%
Champagne	99.3%	98.8%	99.5%	98.9%
Burgundy	96.4%	92.4%	96.9%	92.5%
Alsace	98.0%	95.7%	98.4%	95.9%
Loire Valley	95.7%	88.1%	96.8%	88.1%
Roussillon	98.2%	92.6%	98.7%	92.7%
Savoie	99.0%	96.4%	99.2%	96.4%
Corsica	99.6%	94.8%	99.7%	94.9%

As shown by many different studies in the literature, a wine's elemental profile is a source of information to distinguish the grape varieties [13,51–53]. Nonetheless, these studies are limited geographically, so evaluating international samples contributes to the knowledge already created in the literature.

In Table 5, the accuracy and specificity of the validation and test sets are detailed. Notable results are obtained with values surpassing 70% for the test set. Comparing to the previous tables, an increase in the difference between the metrics of validation and test sets is seen. This may be due to many factors such as the disclosure of varieties not being mandatory in the labels nor their percentages if the wine is multivarietal [54]. Additionally, the rootstock is another factor that may impact the uptake of minerals and, consequently, the MWP [55]. Despite these factors, the algorithm still has impressive results for varietal distinction, which shows an importance of the grape variety to a wine's composition.

Table 5. Accuracy (Acc) and specificity (spc) of the prediction of the validation (val) and test sets for main grape variety classification using XGB. The values presented are the mean of 10 iterations.

Variety	**Acc_{val}**	**Acc_{test}**	**Spc_{val}**	**Spc_{test}**

Gamay	97.6%	97.2%	98.5%	97.6%
Chardonnay	90.4%	88.1%	91.1%	87.6%
Merlot	91.9%	91.6%	92.5%	91.9%
Syrah	90.2%	83.1%	91.8%	82.8%
Grenache noir	90.1%	83.4%	91.5%	82.9%
Cabernet Sauvignon	88.4%	77.8%	89.9%	77.8%
Pinot noir	92.4%	83.2%	93.5%	83.3%
Sauvignon blanc	95.0%	90.8%	95.8%	90.9%
Muscat	95.6%	87.4%	96.1%	87.4%
Cabernet Franc	92.2%	75.6%	93.3%	75.6%
Cinsault	96.1%	87.9%	96.7%	87.8%
Grenache blanc	95.4%	84.6%	95.9%	84.6%
Viognier	95.9%	76.3%	96.5%	76.3%
Malbec	97.0%	86.5%	97.4%	86.6%
Grenache	96.1%	82.9%	96.5%	82.8%
Cinsault noir	96.3%	82.4%	96.7%	82.5%
Riesling	98.3%	89.4%	98.5%	89.4%
Carignan noir	96.7%	83.7%	97.1%	83.7%
Pinot meunier	98.5%	93.7%	98.7%	93.8%
Pinot gris	97.4%	87.3%	97.7%	87.3%
Sémillon	96.9%	84.6%	97.3%	84.6%
Rolle	96.1%	84.3%	96.5%	84.3%
Tempranillo	98.7%	94.9%	98.9%	94.9%
Mourvèdre	92.2%	74.8%	92.5%	74.9%
Vermentino	96.6%	80.5%	97.0%	80.5%
Gewurztraminer	98.2%	86.6%	98.4%	86.6%
Pinot blanc	97.3%	84.8%	97.6%	84.8%
Roussanne	97.4%	79.9%	97.6%	80.0%
Chenin Blanc	97.9%	84.7%	98.2%	84.7%

When compared to the literature, the studies are often restricted in size, with the number of samples varying from less than hundreds of samples up to a thousand. The creation of a large MWP dataset permits the development of a polyvalent model for origin verification of wine.

This study's findings highlight the impressive predictive capabilities in identifying the country, region of wine production and variety of a wine. Future prospects should focus in the correlation of wine production practices, microclimate and wine's sensory attributes in a sub-regional scale in order to further deepen the knowledge about the MWP's formation and impact. Its association with AI emerges as a vital tool for such investigations.

Using a large and diverse dataset, this study developed Extreme Gradient Boosting models that achieved mean accuracies of 95% for country classification, 93% for French wine region classification, and 85% for grape variety classification. All of the test sets accuracies are illustrated in

Figure 3. Additionally, the model specificity reached up to 99% when assessing a wine's origin, based solely on its MWP.

Integrating MWP and AI is essential for advancing the wine industry, addressing the evolving demands of contemporary consumers for detailed origin authentication beyond mere geographical regions.

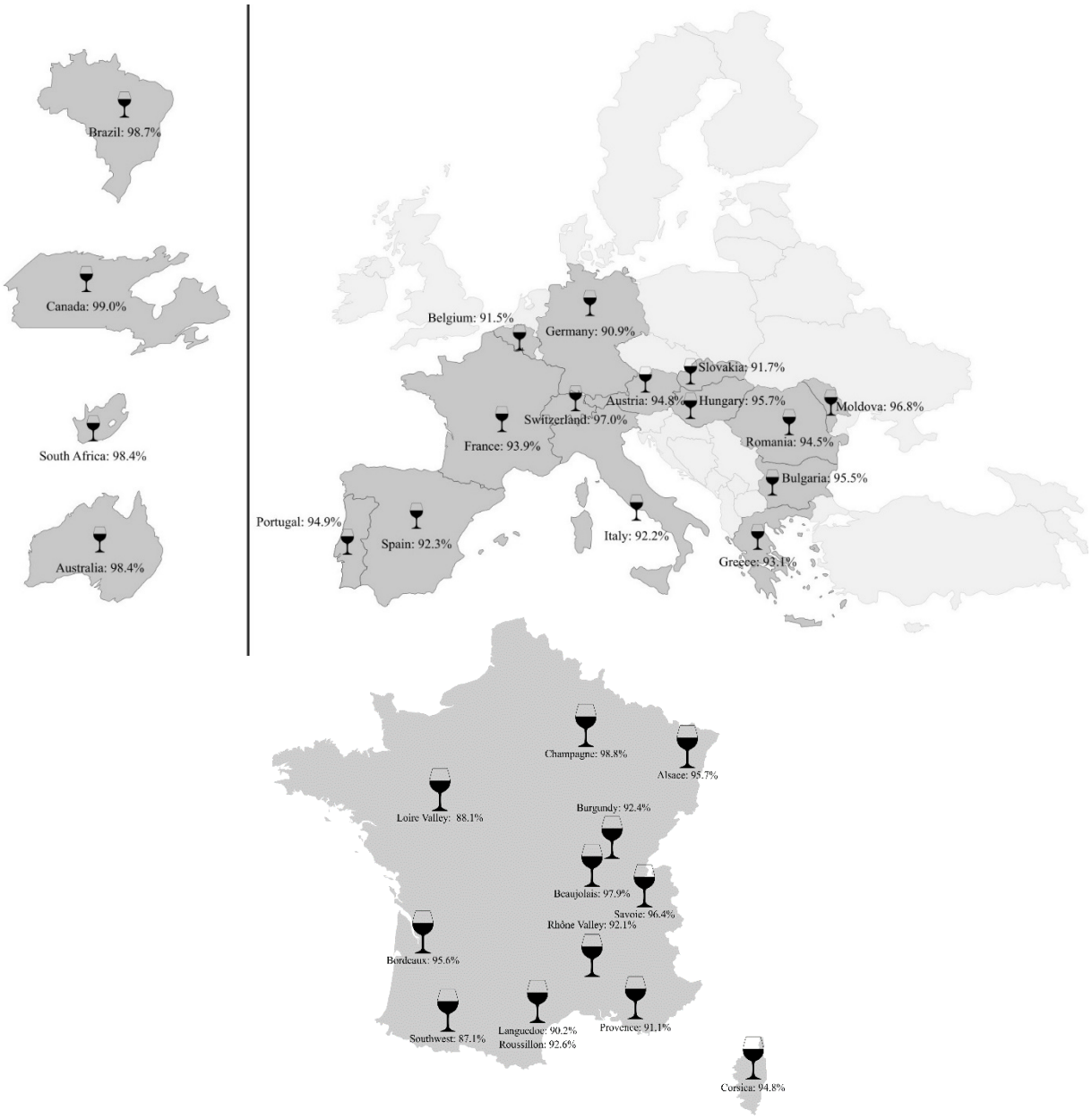

Figure 3. Graphical representation of the classification accuracy for countries (top) and French wine regions (bottom). The accuracy was calculated based on the test set. The scale is not representative of the size of the countries. Mean total accuracy for country is 95% and for region, 93%.

4. Références

1. A. Popîrdă, C. E. Luchian, V. V. Cotea, L. C. Colibaba, E. C. Scutarașu, and A. M. Toader, Agriculture **11**, 225 (2021).
2. A. G. Potortí, V. Lo Turco, M. Saitta, G. D. Bua, A. Tropea, G. Dugo, and G. Di Bella, Natural Product Research **31**, 1000 (2017).
3. M. M. Baleiras-Couto and J. E. Eiras-Dias, Analytica Chimica Acta **563**, 283 (2006).
4. C. Villano, M. T. Lisanti, A. Gambuti, R. Vecchio, L. Moio, L. Frusciante, R. Aversano, and D. Carputo, Food Control **80**, 1 (2017).
5. S. Zambianchi, G. Soffritti, L. Stagnati, V. Patrone, L. Morelli, and M. Busconi, Food Control **142**, 109249 (2022).
6. C. Li, X. Kang, J. Nie, A. Li, M. A. Farag, C. Liu, K. M. Rogers, J. Xiao, and Y. Yuan, Food Chemistry **398**, 133896 (2023).
7. M. Perini and L. Bontempo, TrAC Trends in Analytical Chemistry **147**, 116515 (2022).
8. M. Schartner, J. M. Beck, J. Laboyrie, L. Riquier, S. Marchand, and A. Pouget, Commun Chem **6**, 247 (2023).
9. V. Merkytė, E. Longo, G. Windisch, and E. Boselli, Foods **9**, 1785 (2020).
10. I. Le Mao, G. Da Costa, and T. Richard, OENO One **57**, 15 (2023).
11. L. Zhang, Q. Liu, Y. Li, S. Liu, Q. Tu, and C. Yuan, Current Research in Food Science **6**, 100418 (2023).
12. C. K. Tanabe, J. Nelson, R. B. Boulton, S. E. Ebeler, and H. Hopfer, Molecules **25**, 2552 (2020).
13. K. Pasvanka, M. Kostakis, M. Tarapoulouzi, P. Nisianakis, N. S. Thomaidis, and C. Proestos, Separations **8**, 119 (2021).
14. D. Bertoldi, R. Larcher, M. Bertamini, S. Otto, G. Concheri, and G. Nicolini, J. Agric. Food Chem. **59**, 7224 (2011).
15. P. Pohl, TrAC Trends in Analytical Chemistry **26**, 941 (2007).
16. F. R. S. Bentlin, F. H. Pulgati, V. L. Dressler, and D. Pozebon, J. Braz. Chem. Soc. **22**, 327 (2011).
17. M. M. M. Lima, D. Hernandez, and R. C. Runnebaum, ACS Food Sci. Technol. **3**, 1646 (2023).
18. G. Astray, C. Martinez-Castillo, J.-C. Mejuto, and J. Simal-Gandara, Journal of Food Composition and Analysis **102**, 104043 (2021).
19. N. L. Da Costa, J. P. B. Ximenez, J. L. Rodrigues, F. Barbosa, and R. Barbosa, Eur Food Res Technol **246**, 1193 (2020).
20. P. Alonso Gonzalez, E. Parga-Dans, P. Arribas Blázquez, O. Pérez Luzardo, M. L. Zumbado Peña, M. M. Hernández González, Á. Rodríguez-Hernández, and C. Andújar, PLoS ONE **16**, e0258739 (2021).
21. S. Catarino, A. S. Curvelo-Garcia, and R. B. D. Sousa, Talanta **70**, 1073 (2006).
22. L. Sarlo, C. Duroux, Y. Clément, P. Lanteri, F. Rosseti, O. David, A. Tillement, P. Gillet, A. Hagège, L. David, M. Dumoulin, R. Marchal, T. Tillement, F. Lux, and O. Tillement, OENO One.
23. T. Fawcett, Pattern Recognition Letters **27**, 861 (2006).
24. T. Chen and C. Guestrin, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, San Francisco California USA, 2016), pp. 785–794.
25. M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mane, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viegas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, (2015).
26. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, A. Müller, J. Nothman, G. Louppe, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, (2012).
27. L. Breiman, Machine Learning **45**, 5 (2001).
28. T. Cover and P. Hart, IEEE Transactions on Information Theory **13**, 21 (1967).
29. C. Cortes and V. Vapnik, Mach Learn **20**, 273 (1995).
30. D. R. Cox, Journal of the Royal Statistical Society: Series B (Methodological) **20**, 215 (1958).
31. N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, Journal of Artificial Intelligence Research **16**, 321 (2002).
32. P. Kment, M. Mihaljevič, V. Ettler, O. Šebek, L. Strnad, and L. Rohlová, Food Chemistry **91**, 157 (2005).
33. H. Wu, G. Lin, L. Tian, Z. Yan, B. Yi, X. Bian, B. Jin, L. Xie, H. Zhou, and K. M. Rogers, Food Chemistry **339**, 127760 (2021).

34. M. P. Fabani, R. C. Arrúa, F. Vázquez, M. P. Diaz, M. V. Baroni, and D. A. Wunderlin, Food Chemistry **119**, 372 (2010).

35. P. Paneque, M. T. Álvarez-Sotomayor, A. Clavijo, and I. A. Gómez, Microchemical Journal **94**, 175 (2010).

36. Z. Temerdashev, M. Bolshov, A. Abakumov, A. Khalafyan, A. Kaunova, A. Vasilyev, O. Sheludko, and A. Ramazanov, Molecules **28**, 4319 (2023).

37. M. Gajek, A. Pawlaczyk, and M. I. Szynkowska-Jozwik, Molecules **26**, 214 (2021).

38. B. Tariba, Biol Trace Elem Res **144**, 143 (2011).

39. V. G. Mihucz, C. J. Done, E. Tatár, I. Virág, G. Záray, and E. G. Baiulescu, Talanta **70**, 984 (2006).

40. M. Aceto, E. Robotti, M. Oddone, M. Baldizzone, G. Bonifacino, G. Bezzo, R. Di Stefano, F. Gosetti, E. Mazzucco, M. Manfredi, and E. Marengo, Food Chemistry **138**, 1914 (2013).

41. N. Jakubowski, R. Brandt, D. Stuewer, H. R. Eschnauer, and S. Görtges, Fresenius J Anal Chem **364**, 424 (1999).

42. A. E. Martin, R. J. Watling, and G. S. Lee, Food Chemistry **133**, 1081 (2012).

43. M. del M. Castiñeira, R. Brandt, N. Jakubowski, and J. T. Andersson, J. Agric. Food Chem. **52**, 2953 (2004).

44. S. Catarino, M. Madeira, F. Monteiro, F. Rocha, A. S. Curvelo-Garcia, and R. B. De Sousa, J. Agric. Food Chem. **56**, 158 (2008).

45. G.-D. Dumitriu (Gabur), C. Teodosiu, I. Morosanu, O. Plavan, I. Gabur, and V. V. Cotea, Journal of Food Composition and Analysis **100**, 103935 (2021).

46. J. Griboff, M. Horacek, D. A. Wunderlin, and M. V. Monferrán, Front. Sustain. Food Syst. **5**, 657412 (2021).

47. O. R. Dinca, R. E. Ionete, D. Costinel, I. E. Geana, R. Popescu, I. Stefanescu, and G. L. Radu, Food Anal. Methods **9**, 2406 (2016).

48. M. Forina, P. Oliveri, H. Jäger, U. Römisch, and J. Smeyers-Verbeke, Chemometrics and Intelligent Laboratory Systems **99**, 127 (2009).

49. M. E. Conti, M. Rapa, C. Simone, M. Calabrese, G. Bosco, S. Canepari, and M. L. Astolfi, Food Control **158**, 110226 (2024).

50. Food Research International **163**, 112165 (2023).

51. J. Pořízka and P. Diviš, Journal of Elementology **23**, (2018).

52. Z. Temerdashev, A. Khalafyan, A. Abakumov, M. Bolshov, V. Akin'shina, and A. Kaunova, Heliyon e29607 (2024).

53. I. Feher, D. A. Magdas, A. Dehelean, and C. Sârbu, J Food Sci Technol **56**, 5225 (2019).

54. International Organisation of Vine and Wine, (2024).

55. S. M. Olarte Mantilla, C. Collins, P. G. Iland, C. M. Kidman, R. Ristic, P. K. Boss, C. Jordans, and S. E. P. Bastian, Am J Enol Vitic. **69**, 32 (2018).