Machine learning feature extraction for predicting the ageing of olive oil

Arnaud Gucciardi[a,b], Safouane El Ghazouali[a], Umberto Michelucci[a,c], and Francesca Venturini[a,d]

[a]TOELT llc, Machine Learning Research and Development, Birchlenstr. 25, 8600 Dübendorf, Switzerland
[d]Institute of Applied Mathematics and Physics, Zurich University of Applied Sciences, Technikumstrasse 9, 8401 Winterthur, Switzerland
[b]Artificial Intelligence Laboratory, University of Ljubljana, Ljubljana, Slovenia
[c]Lucerne University of Applied Sciences and Arts, Computer Science Department, Lucerne, Switzerland

ABSTRACT

Monitoring the quality of extra virgin olive oil (EVOO) during its life cycle is of particular importance due to its influence on health-related characteristics and its significance for the oil industry. For this reason it is critical to find an easy-to-perform, non-destructive and affordable method to monitor the quality of EVOO and detect its degradation due to ageing. The following study explores a machine learning approach based on fluorescence measurements for predicting oil changes arising from the ageing process. The proposed method specifically predicts the quality parameters that are required for an olive oil to qualify as extra virgin. In particular, the two properties considered in this analysis are the UV absorbance at 232 and 268 nm (K_{232} and K_{268}), both critical markers of the quality of extra virgin oil. To achieve this goal, a large dataset of fluorescence measurements was analysed, comprising 720 excitation-emission matrices of twenty-four different oils initially labeled as extra virgin. The samples were aged under accelerated conditions at 60 °C in the dark for nine weeks and their properties were measured at ten different time steps during the process.

Two different machine learning pipelines were implemented for the prediction of K_{232} and K_{268}. In a first approach, the model was trained on all the ten ageing steps of each oil and learned to predict all the ten steps of an unseen oil. In a second approach, the model was trained on one single ageing on multiple oils and step for all the oils and learned to predicta single ageing step. The results demonstrate the potential of the proposed approach.

Keywords: Fluorescence spectroscopy; olive oil; machine learning; artificial neural networks; quality control; explainability; convolutional neural networks

1. INTRODUCTION

Ensuring the quality of olive oil, due to its health-related attributes and its significance to the oil industry, has garnered increasing attention, especially in the context of ageing. Dynamic changes in chemical properties during the ageing process require accurate monitoring to assess quality degradation. In this pursuit, predicting how specific oil properties evolve over time becomes crucial. This study introduces and validates a novel machine learning approach based on fluorescence spectroscopy designed to forecast changes in critical chemical parameters associated with the qualification of olive oil as extra virgin quality during the ageing process.

The quality assessment of extra virgin olive oil (EVOO) is performed through chemical analyses and sensory evaluation by certified laboratories and panelists.[1,2] Fluorescence spectroscopy, which has the advantages of being a rapid, cost-efficient, and non-destructive technique, has been proposed as an alternative to these time-consuming and expensive analyses. This technique has been successfully used to monitor the freshness of EVOO

Send correspondence to F. Venturini: francesca.venturini@zhaw.ch

Data Science for Photonics and Biophotonics, edited by Thomas Bocklitz,
Proc. of SPIE Vol. 13011, 130110A · © 2024 SPIE
0277-786X · doi: 10.1117/12.3017680

due to the natural presence of fluorophores in olive oil, the strongest of which is chlorophyll.[3–5] The advantage of fluorescence measurements is that they can be performed on undiluted samples, therefore eliminating the need for any sample handling, and significantly simplifying thir use.

To achieve this objective, this study used a large dataset comprising 720 fluorescence excitation-emission matrices (EEMs) of 24 olive oils, initially labeled as extra virgin. The oils underwent accelerated ageing for up to nine weeks, with properties measured at ten different time steps. The focus of this investigation lies in predicting two key quality parameters essential for extra virgin olive oil classification: the UV absorbance at 232 and 268 nm (K_{232} and K_{268}), both recognized as crucial markers of extra virgin oil oxidation. The trained model uses the excitation-emission matrices at different ageing steps as input for the model to predict the quality parameters at the corresponding time intervals.

This paper evaluates the proposed machine learning model using two different training strategies. In the first strategy, the model is trained on the evolution of the EEM and absorbance parameters over ten steps. The model's task is to predict the entire evolution of the absorbance parameters of an unseen oil from the EEMs. In the second strategy, a model is trained on each ageing step and, therefore, focusses on measuring the model's ability to predict absorbance parameters from a single EEM. The findings presented in this work show how the information related to the oxidation and ageing process is contained in the fluorescence data and demonstrate that fluorescence spectroscopy can be a quantitative tool for the quality assessment of olive oil.

2. MATERIAL AND METHODS

2.1 Olive oil samples and experimental setup

For this study, twenty-four different extra virgin olive oils were used. The oils, commercially available at the Migros and Coop supermarkets in Switzerland, were chosen to be as heterogeneous as possible in terms of price classes and production regions to identify characteristics that are common to all EVOO. The list of the EVOOs is reported in Table 1.

The ageing was performed under accelerated conditions at 60 °C in the dark following the modified Schaal oven test Celsius.[6] This approach enables the analysis of the oxidation in a shorter time because it reproduces oxidative changes similar to those observed under actual shelf life conditions.[7] The quality of EVOOs at the various phases of ageing can be monitored using UV absorption spectroscopy, according to the European Regulation and its amendments[1,2] through the absorbance in the at 232 nm K_{232} and the absorbance at 268 nm, K_{268}.

The fluorescence EEMs of each of the samples were acquired with an Agilent Cary Eclipse Fluorescence Spectrometer. All measurements were made at a constant temperature of 22 °C on undiluted samples.

Detailed information on the acquisition parameters and procedures of both spectroscopic techniques are reported in a separate paper.[8] The complete dataset including both UV-absorption and fluorescence measurements is publicly available.[9]

2.2 Machine learning methods

The goal of this work is to predict two UV absorption parameters, K_{232} and K_{268}, at the various ageing steps using excitation emission matrices as input. Given that K_{232} and K_{268} are continuous variables, this task falls within the realm of regression and supervised learning. For regression operations, many machine learning algorithms have been proposed. Among these, AdaBoost is widely recognized for its efficiency and flexibility, especially when dealing with complex data.[10] Therefore, in this work, AdaBoost with decision trees as learners (more on that below) has been used to test the feasibility of the described approach: namely, extracting the UV parameters from the excitation emission matrices. Various tests, not reported in this paper, show that other methods, such as decision trees or random forests, perform poorly compared to AdaBoost.

AdaBoost, short for Adaptive Boosting,[10–12] is a type of ensemble learning method that combines the strengths of multiple weak learners (i.e., simple regression models) to create a more robust and accurate predictive model. The algorithm assigns higher weights to misclassified data points in each iteration, effectively down-weighting the influence of outliers during training. This allows the algorithm to focus on the most challenging instances while improving its predictive capabilities.

Table 1. Olive oils samples analyzed in this study. Geographical origin: IT=Italy, ES=Spain, GR=Greece, PT=Portugal, EU=European not specified.

Label	Sample description	Origin
A	Coop Naturaplan Italienisches Olivenöl (BIO)	IT
B	Hacuinda Don Paolo	ES
C	Monocultivar Nocellara Bio	IT
D	Monini, Toscano IGP	IT, Tuscany
E	Monini, Classico	IT
F	Oliva, Favola	IT
G	Migros, M Classics	ES
H	Alexis, Manaki	GR
I	Migros, Bio Italienische Olivenöl	IT
J	Alnatura, Natives Olivenöl extra	ES
K	Migros, Bio Griechisches Olivenöl	GR
L	Demeter, Spanisches Olivenöl	ES
M	Filippo Berio, Il Classico	EU
N	Demeter, Bio Coop Naturaplan Portugisisches Olivenöl	PT
O	Castillo, Don Felpe	ES
P	Coop, Naturaplan Bio Griechisches Olivenöl	GR
Q	Demeter, Son Naava	ES, Mallorca
R	Iliada, Kalamata PDO	GR
S	Sapori d'Italia, Sicilia	IT, Sicily
T	San Giuliano, Sardegna DOP	IT, Sardinia
U	Coop, Naturaplan Bio Spanisches Olivenöl	ES
V	San Giuliano, Fruttato	IT
W	San Giuliano, L'Originale	IT
X	Coop, Qualité-Prix	IT, ES, GR

Fluorescence spectroscopy data, and specifically EEMs, have been used as input in various deep learning and machine learning methods for information extraction.[13–17] EEMs can be used as input, either as two-dimensional matrices or flattened one-dimensional arrays obtained by transforming the matrix's rows into a single, linear sequence of the pixel values. Since AdaBoost does not understand two-dimensional structures as input (it is designed to process one-dimensional arrays), the EEMs have been flattened out to train the algorithm. An additional advantage of the AdaBoost algorithm is that it can perform well without the need for preprocessing such as feature scaling. This can be advantageous when dealing with datasets with features of different scales or with an unknown range of variations. In the case of EEMs this advantage is not exploited, as each pixel has the same value range from 0 to 1000 due to instrumental constraints.

In this work, the AdaBoost algorithm was trained to predict the values of the parameters K_{232} and K_{268}, the UV absorbance at 232 and 268 nm, and thus quantify the ageing process. The AdaBoost parameters were the same for each case: 50 estimators, with the base estimator being a decision tree of depth 3, a linear loss function, and a stable learning rate of 1. For each ageing step three measurements were performed for each oil. The EEMs used for algorithm training were obtained from the average of the three measurements.[9] The pixel values of each EEMs were normalized, transforming their value to be between 0 and 1.

The models were trained with two different split strategies (A) and (B) visually represented in Fig. 1. In strategy A, the model learns to predict all ten values of the UV-parameters for each oil at the same time. The inputs consist of 23 arrays, each consisting of ten flattened EEMs (one for each ageing step) for each oil. The model is then tested on an unseen oil (see Section 2.3) and, starting from the EEMs at the ten different ageing steps, it predicts the K_{232} and K_{268} values at the ten ageing steps simultaneously. Although it is a good indicator of the general predictability of patterns related to the parameters K_{232} and K_{268}, it requires the ten measurements of the ageing process, which may not be available in practice. To test a more practical approach, the alternative strategy B was tested, which focusses on predicting individual ageing steps. In strategy B, the

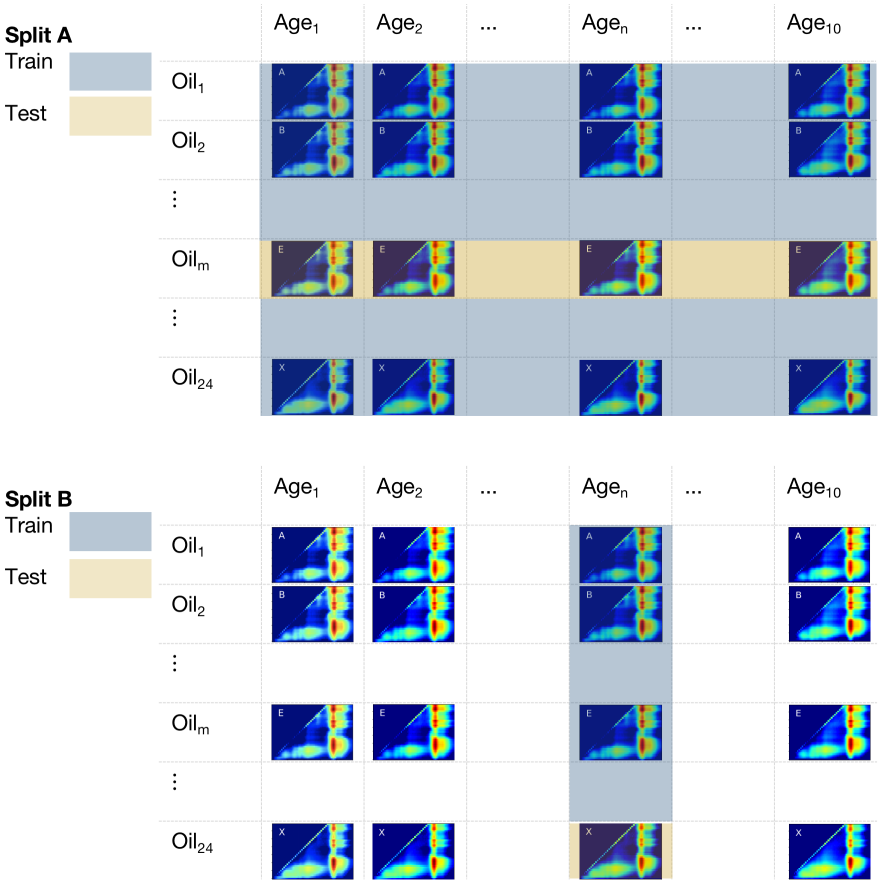

Figure 1. Training split strategies employed for the regression on the the K_{232} and K_{268} parameters. To predict the chemical behaviour of a single through the ten ageing steps of an oil, two different strategies are used: the first one predicts the behaviour of a single oil for the ten ageing steps, and the second one predicts each ageing step of an oil based on the behaviour of the remaining oil at that given age.

model learns to predict the parameters K_{232} and K_{268} of a single ageing step. For this purpose, a different model is trained for each ageing step. Using a leave-one-out approach (see Section 2.3), each model uses as input the EMMs of 23 oils and learns to predict K_{232} and K_{268} at this specific ageing step. The test is performed on an unseen oil at this ageing step. This is then repeated for all ageing steps.

(A) The input to the model is a batch of the ten EEMs at different ageing steps of an oil. The model is trained on a leave-one-out setting, where the training set consists of twenty-three batches of ten EEMs. The labels of the K_{232} and K_{268} for the remaining test oil are simultaneously predicted for the ten different ageing steps.

(B) The inputs to the model are all the oil samples EEMs at a given ageing step, except for the left out oil, which is used for the test. The values of K_{232} and K_{268} of the left-out oil are predicted at a specific ageing step. This process is repeated for each of the ten ageing steps.

The metric for the evaluation of the performance of the models is the Mean Absolute Error (MAE)

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{1}$$

where n represents the number of observations (either ten for a single oil or 240 for all oils together), y_i is the actual value for observation i ($K_{232,i}$ or $K_{268,i}$) and $\hat{y}_i$ is the predicted value. For the training strategy A, the

absolute error (AE) is calculated for each oil at all ageing steps and then averaged over the 24 models. For the training strateg B, the AE is calculated at each ageing step for all the 24 oils and then averaged over the ten ageing steps.

2.3 Validation

Due to the small size of the dataset, a leave-one-out (LOO) validation approach was used to validate the models.[18] LOO validation is a preferred method for evaluating the performance of a machine learning model when the dataset is small. In the LOO setting, a single oil is removed from the dataset, and the model training is completed on the remaining oils. The oil removed is then used as a test point to evaluate the model's performance. This process is repeated for each oil in the dataset, for both training approaches described in Section 2.2. The resulting predictions are used to assess the generalization properties of AdaBoost class of models. In strategy A, the left-out input is a stack of ten EEMs from ten ageing steps for a single olive oil, and in strategy B, the left-out input is an EEM for a single olive oil. To compare the prediction with both strategies, the MAEs are calculated and compared as described above.

3. RESULTS

The selected AdaBoost regressor used in this work learns to predict the absorbance parameters K_{232} and K_{268} from the fluorescence EEMs using the split strategies described in Section 2.2. The detailed results for each of the 24 oils and ten ageing steps obtained using the split strategy B are shown in Figures 2 and 3. The split strategy B is chosen because of its higher relevance for practical applications.

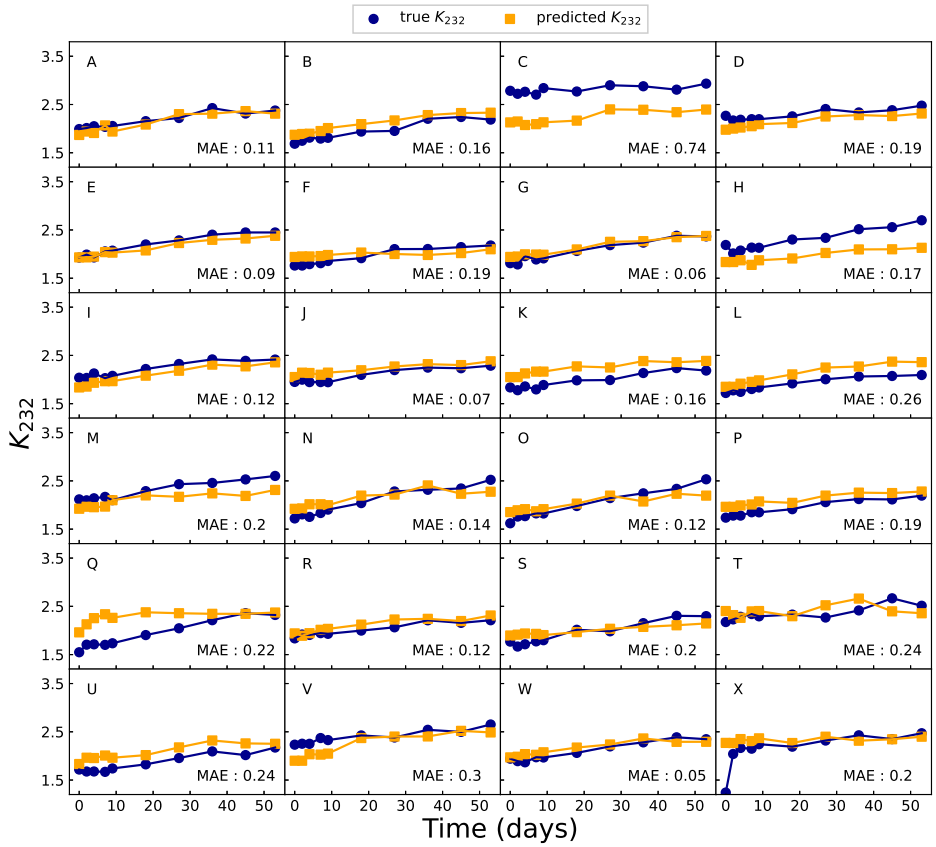

Figure 2. Prediction of the K_{232} parameter using the split strategy B and LOO validation. Each subgraph shows both the ground (circles, blue) and the predicted value (squared, yellow). For each oil the Mean Absolute Error obtained as average over the 10 steps is reported at the bottom right corner of each graph.

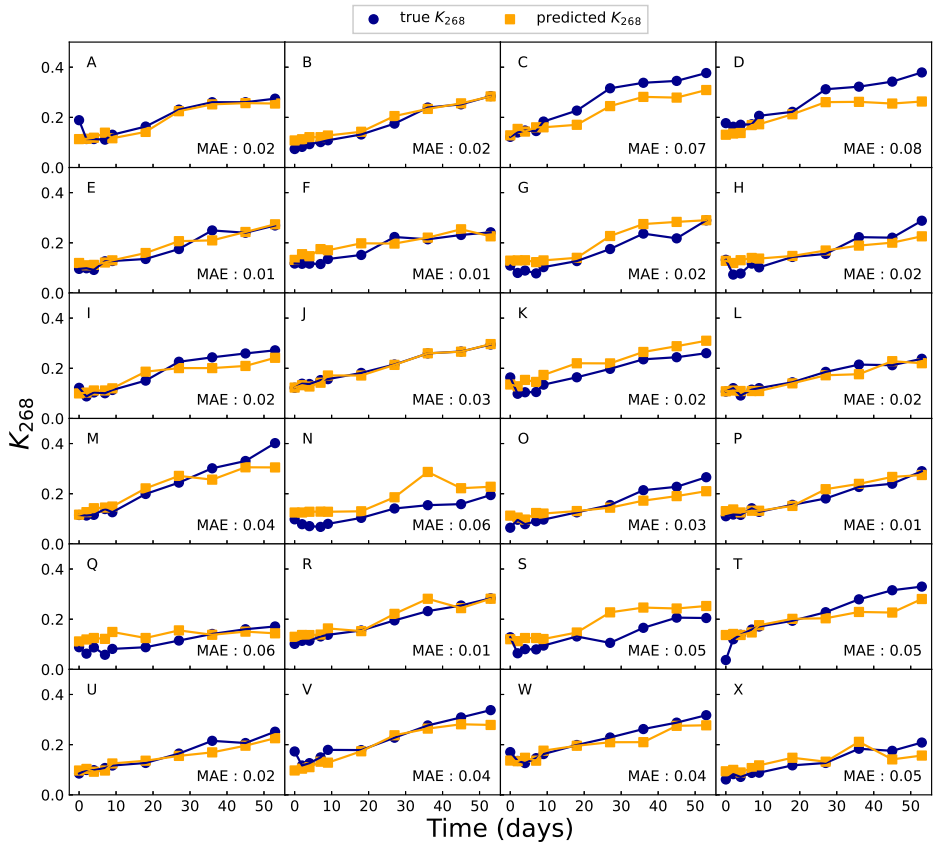

Figure 3. Prediction of the K_{268} parameter using the split strategy B and LOO validation. Each subgraph shows both the ground (circles, blue) and the predicted value (squared, yellow). For each oil the Mean Absolute Error (MAE) obtained as average over the ten steps is reported at the bottom right corner of each graph.

The predictions of the absorbance at 232 nm K_{232} is very good for all oils except for the oil C (Figure 2). This is not surprising because the oil C has from the beginning a K_{232} which is above the limit of the European Regulations, meaning the oil is already not edible immediately after opening the fresh bottle. Therefore, the model trained on EVOO is not good at predicting the evolution of K_{232} of an oil which is not EVOO from the beginning. Similarly the oil X has an unusual behaviour at the ageing step 0 which is not learnable from the model. Other predictions that present errors, e.g. on oils H,L and Q, still approximate the ageing behavior with relative stability. The most accurate evolution predictions are observed on oil samples G and W with MAEs of 0.06 and 0.05 respectively.

The predictions of the absorbance at 268 nm K_{268} are stable for all the oils and ageing steps (Figure 3). Similarly to what observed for K_{232}, the prediction for the C oil is an outlier, with the second highest MAE at 0.07.

An overview of all the results is given by the distribution of AEs for all oils and ageing steps for both parameters and split strategies that is shown in Fig. 4. An occurrence count is a prediction on a single ageing step and oil. The predictions arising from the two split strategies are very similar when comparing the MAEs. The MAE for the K_{232} obtained with split strategy A is 0.18 and 0.19 with split strategy B. Similarly, the MAEs for the prediction of the K_{268} parameter are 0.028 and 0.033, respectively.

These results are summarized in Table 2. For comparison, the error on the labels due to the experimental procedure and instrumentation is also reported in the table (as label error). This error was determined by making multiple measurements under identical conditions of the same sample and calculating three times the standard deviation. Table 2 shows that while for K_{232} the MAE of the prediction is approximately three times

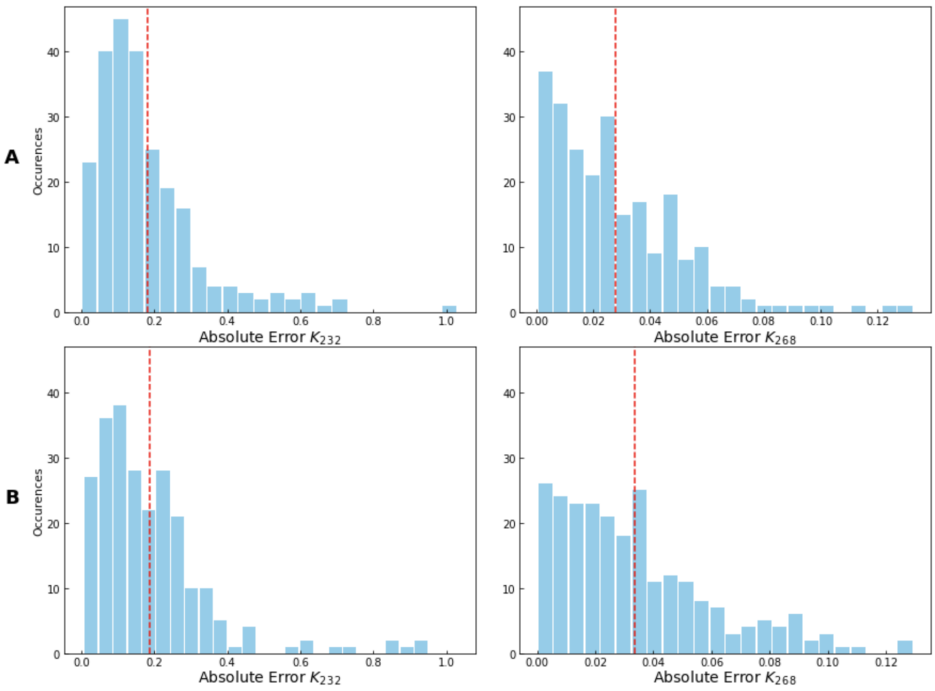

Figure 4. Distribution of the absolute error (AE) on the prediction for both K_{268} and K_{232}. An occurrence count is a prediction on a single ageing step and oil. Top panels: Split strategy A; bottom panels: split strategy B. The red dashed lines mark the average of the MAE.

the experimental error on the label, the MAE for the prediction of K_{268} is only slightly higher. This indicates that the K_{268} is the parameter whose changes are most strongly correlated with changes in fluorescence intensity. Also, the changes of the K_{268} during to the ageing is more significant, with an average rate of change of 0.004/day, and a total change of the average of 143%. For comparison, the average rate of change of K_{232} is 0.008/day and the change of the average over 53 days is 26%, making the latter a less sensitive indicator of deterioration.[19]

Table 2. Mean Absolute Error (MAE) for the prediction for both split strategies and experimental error on the label.

Predicted parameter	MAE with split A	MAE with split B	Label error
K_{232}	0.18	0.19	0.06
K_{268}	0.028	0.033	0.02

In the majority of cases and for both K_{232} and K_{268} parameters, the ageing trend and final ageing step value are accurately predicted thanks to the low relative MAE errors.

4. CONCLUSION

This study demonstrates the effectiveness of the AdaBoost Regressor machine learning model applied to fluorescence measurements in accurately predicting changes in chemical quality indicators during the ageing process of olive oil. Specifically, the parameters predicted from the excitation-emission matrices are the UV absorbance at 232 and 268 nm (K_{232} and K_{268}), both critical markers of the quality of extra virgin oil. The two distinct training strategies demonstrated the model's robustness and ability to capture complex non-linear relationships at all ageing steps. This work shows the potential of fluorescence spectroscopy combined with machine learning to develop new optical methods for monitoring olive oil quality throughout its entire life cycle. The advantages of fluorescence spectroscopy are that it is a user-friendly, non-destructive, rapid technique, that does not require any sample preparation. As such, the integration of machine learning methods in fluorescence spectroscopy measurement pipelines can become a valuable alternative to traditional chemical analysis techniques.

ACKNOWLEDGMENTS

This work was supported by the project "SUSTAINABLE" funded by the European Union's Horizon 2020 Project H2020-MSCA-RISE-2020 Grant No. 101007702.

REFERENCES

[1] "Commission regulation (eec) no. 2568/91 of 11 july 1991 on the characteristics of olive oil and olive-residue oil and on the relevant methods of analysis official journal l 248, 5 september 1991," *Offic. JL* **248**, 1–83 (1991).

[2] "Commission implementing regulation no 1348/2013 of december 17 2013," *Official Journal of the European Union* **338**, 31–67 (2013).

[3] Sikorska, E., Khmelinskii, I. V., Sikorski, M., Caponio, F., Bilancia, M. T., Pasqualone, A., and Gomes, T., "Fluorescence spectroscopy in monitoring of extra virgin olive oil during storage," *International journal of food science & technology* **43**(1), 52–61 (2008).

[4] Karoui, R. and Blecker, C., "Fluorescence spectroscopy measurement for quality assessment of food systems—a review," *Food and Bioprocess technology* **4**(3), 364–386 (2011).

[5] Lobo-Prieto, A., Tena, N., Aparicio-Ruiz, R., García-González, D. L., and Sikorska, E., "Monitoring virgin olive oil shelf-life by fluorescence spectroscopy and sensory characteristics: A multidimensional study carried out under simulated market conditions," *Foods* **9**(12), 1846 (2020).

[6] Evans, C., List, G., Moser, H. A., and Cowan, J., "Long term storage of soybean and cottonseed salad oils," *Journal of the American Oil Chemists' Society* **50**(6), 218–222 (1973).

[7] Mancebo-Campos, V., Fregapane, G., and Desamparados Salvador, M., "Kinetic study for the development of an accelerated oxidative stability test to estimate virgin olive oil potential shelf life," *European Journal of Lipid Science and Technology* **110**(10), 969–976 (2008).

[8] Venturini, F., Fluri, S., and Baumgartner, M., "Dataset of fluorescence eem and uv spectroscopy data of olive oils during ageing," *Data* **8**(81), 1–6 (2023).

[9] Venturini, F., Fluri, S., and Baumgartner, M., "Dataset of fluorescence eem and uv spectroscopy data of olive oils during ageing," *Mendeley data* .

[10] Chengsheng, T., Huacheng, L., and Bing, X., "Adaboost typical algorithm and its application research," in [*MATEC Web Conf*], **139**(1), 00222, Springer (2017).

[11] Ding, Y., Zhu, H., Chen, R., and Li, R., "An efficient adaboost algorithm with the multiple thresholds classification," *Applied Sciences* **12**(12) (2022).

[12] Solomatine, D. P. and Shrestha, D. L., "Adaboost. rt: a boosting algorithm for regression problems," in [*2004 IEEE International Joint Conference on Neural Networks (IEEE Cat. No. 04CH37541)*], **2**, 1163–1168, IEEE (2004).

[13] Zhang, X., Yang, J., Lin, T., and Ying, Y., "Food and agro-product quality evaluation based on spectroscopy and deep learning: A review," *Trends in Food Science & Technology* **112**, 431–441 (2021).

[14] Batista-Andrade, J. A., Vega, D. I., McClain, A., and Blaney, L., "Using multilinear regressions developed from excitation-emission matrices to estimate the wastewater content in urban streams impacted by sanitary sewer leaks and overflows," *Science of The Total Environment* **906**, 167736 (2024).

[15] Xu, R.-Z., Cao, J.-S., Feng, G., Luo, J.-Y., Feng, Q., Ni, B.-J., and Fang, F., "Fast identification of fluorescent components in three-dimensional excitation-emission matrix fluorescence spectra via deep learning," *Chemical Engineering Journal* **430**, 132893 (2022).

[16] Rutherford, J. W., Larson, T. V., Gould, T., Seto, E., Novosselov, I. V., and Posner, J. D., "Source apportionment of environmental combustion sources using excitation emission matrix fluorescence spectroscopy and machine learning," *Atmospheric Environment* **259**, 118501 (Aug. 2021).

[17] Nguyen, X. C., Seo, Y., Park, H.-Y., Begum, M. S., Lee, B. J., and Hur, J., "Tracking the sources of dissolved organic matter under bio- and photo-transformation conditions using fluorescence spectrum-based machine learning techniques," *Environmental Technology & Innovation* **31**, 103179 (Aug. 2023).

[18] Michelucci, U. and Venturini, F., "Estimating neural network's performance with bootstrap: A tutorial," *Machine Learning and Knowledge Extraction* **3**(2), 357–373 (2021).

[19] Venturini, F., Fluri, S., Mejari, M., Baumgartner, M., Piga, D., and Michelucci, U., "Shedding light on the ageing of extra virgin olive oil: Probing the impact of temperature with fluorescence spectroscopy and machine learning techniques," *LWT* **191**, 115679 (2024).