Detection of soft-refined oils in extra virgin olive oil using data fusion approaches for LC-MS, GC-IMS and FGC-Enose techniques: The winning synergy of GC-IMS and FGC-Enose

Food Control

Detection of soft-refined oils in extra virgin olive oil using data fusion approaches for LC-MS, GC-IMS and FGC-Enose techniques: the winning synergy of GC-IMS and FGC-Enose

--Manuscript Draft--

Manuscript Number:	FOODCONT-D-21-00923R2
Article Type:	Research Paper
Keywords:	Soft-deodorization; deacidification; FGC-Enose; adulteration; GC-IMS; LC-MS; soft-refined olive oil (SROO); data fusion
Corresponding Author:	Michele Suman Parma, ITALY
First Author:	Michele Suman
Order of Authors:	Michele Suman
	Alessandra Tata
	Andrea Massaro
	Tito Damiani
	Roberto Piro
	Chiara Dall'Asta
Abstract:	Extra virgin olive oil (EVOO) is frequently adulterated by mixing it with soft refined oils (SROO). The differentiation of EVOO from its blends with SROO is not possible with the most common approaches, and, for this reason, the discriminating power of liquid chromatography-high resolution mass spectrometry (LC-MS), gas-chromatography ion mobility spectrometry (GC-IMS) and flash gas-chromatography electronic nose (FGC-Enose) was examined previously. Here, the combination of the above-mentioned techniques for an improvement in classification power of the methods is explored. A total of 43 commercial EVOOs and 18 illegal mixtures of SROO with EVOO were previously analysed by LC-(+/-)MS , GC-IMS and FGC-Enose. Low-level and mid-level data fusion of the four datasets were performed. The merged unique fingerprints were submitted to partial least squared discriminant analysis (PLS-DA), and the extrapolated most informative variables were used to build support vector machine (SVM) classifiers. Statistical indicators were calculated and compared to find out the best classifier. The results of PLS-DA-SVM strategies on the combination of datasets demonstrated that, after low-level data fusion, the discriminatory capability of the two merged GC-based techniques was remarkably improved as compared to the individual techniques. This indicates that merging the datasets before PLS-DA better retrieves the most informative variables and, thus, enhances group separation and classification of unknowns. The combination of LC(+/-)MS datasets, both by mid- and low-level data fusion, did not show significant enhancement in terms of discrimination of EVOO from SROO as compared to the individual LC(+)MS matrix. The low-level combination of the four datasets (LC(+/-)MS, GC-IMS, FGC-Enose) was successful, although this laborious option is not a viable path in industry quality assurance. This study primarily provides new paths for the authentication of EVOO, taking advantage of merging multimodal LC-(+/-)MS, GC-IMS and FGC-Enose data, with consequent improvement in the performances of the classification models. The most promising results were achieved by the low-level data fusion of GC-IMS and FGC-Enose data.
Suggested Reviewers:	Philipp Weller p.weller@hs-mannheim.de Expertise in Analytical Data Fusion Approaches
	Marta Ferreiro marta.ferreiro@uca.es Expertise in GC-IMS Analysis

	Tullia Gallina Toschi Tultullia.gallinatoschi@unibo.it Expertise in Olive Oil topics and fraud issues

To the kind attention of:
Editor

Dr. Q. Rao, PhD
(Food Science; Food Chemistry; Food Quality; Food Safety)
Florida State University College of Human Sciences Nutrition,
Food & Exercise Sciences, Tallahassee, Florida, United States of America
Food Control - Elsevier

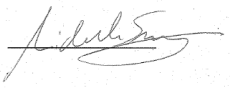
Authors:
*Alessandra Tata[1], Andrea Massaro[1], Tito Damiani[2], Roberto Piro[1], Chiara Dall'Asta[2], Michele Suman[3,4] *****

[1] Istituto Zooprofilattico Sperimentale delle Venezie, Laboratorio di Chimica Sperimentale, Vicenza, Italy
[2] Department of Food and Drug, University of Parma, Parma, Italy
[3] Department of Analytical Food Science, Barilla G. e R. Fratelli S.p.A., Parma, Italy
[4] Department for Sustainable Food Process, Catholic University Sacred Heart, Piacenza, Italy

Title:
"Advantages and disadvantages of data fusion of LC-MS, GC-IMS and FGC-Enose techniques in the authentication of extra virgin olive oil"

Dear Editor,
the present paper describes an original data-fusion exercise devoted to face recent fraud issues within Extra Virgin Olive Oil food chain. In particular taking into account of LC-(+/-)MS, GC-IMS and FGC-Enose analytical data, low-level data fusion of GC-IMS and FGC-ENose datasets demonstrated to be effective in order to generate an optimal model within a new framework for the authentication of EVOO.
It was a positive synergic effort among a control authority (Istituto Zooprofilattico), an academic (University of Parma) and an industrial (Barilla Advanced Research Labs) research labs.
The present manuscript has not been previously submitted/published and is not currently in press, under review or being considered for publication by another journal. Therefore, we would like you to evaluate it for publication and we would be honored in case it will be taken into consideration.

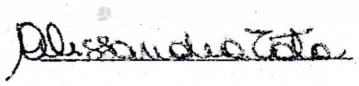
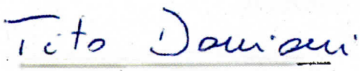
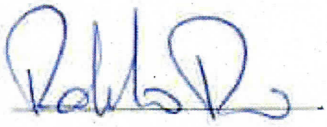
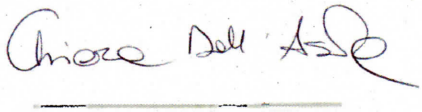
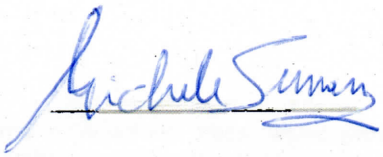
On behalf of all the authors
Yours sincerely.
Michele Suman



Parma, 6th April 2021

Dr. Michele Suman, PhD
Barilla G.R. F.lli SpA
Research, Development & Quality
Food Safety & Authenticity Research Manager
Food Safety Fellow Technical Ladder
Adjunct Professor of AgriFood Authenticity at Catholic University Sacred Heart – Milan/Piacenza
Chair ILSI Process Related Compounds & Natural Toxins Task Force
Chair Italian National Normative Organization (UNI) - Food Authenticity Commission
Scientific Board Member Italian Chemistry Society-Food Chemistry Inter-divisional Group
Via Mantova 166 - 43100 Parma (Italy)
☏ phone **+39 0521 262332**
☏ mobile **+39 3386938349**
✉ mail **michele.suman@barilla.com**
web **www.barillagroup.com**

This statement is signed by all the authors to indicate agreement that the above information is true and correct (*a photocopy of this form may be used if there are more than 10 authors*):

Author's name (typed)	Author's signature	Date
ALESSANDRA TATA		04/03/2021
ANDREA MASSARO		04/03/2021
TITO DAMIANI		04/03/2021
ROBERTO PIRO		04/03/2021
CHIARA DALL'ASTA		04/03/2021
MICHELE SUMAN		10/03/2021

Conflicts of Interest Statement

Manuscript title: ADVANTAGES AND DISADVANTAGES OF DATAFUSION OF LC-MS, GC-IMS AND FGC-ENOSE ~~MULTIPLE MASS SPECTROMETRIC~~ TECHNIQUES IN THE AUTHENTICATION OF EXTRA VIRGIN OLIVE OIL

The authors whose names are listed immediately below certify that they have NO affiliations with or involvement in any organization or entity with any financial interest (such as honoraria; educational grants; participation in speakers' bureaus; membership, employment, consultancies, stock ownership, or other equity interest; and expert testimony or patent-licensing arrangements), or non-financial interest (such as personal or professional relationships, affiliations, knowledge or beliefs) in the subject matter or materials discussed in this manuscript.

Author names:

ALESSANDRA TATA
ANDREA NASSARO
TITO DAMIANI
ROBERTO PIRO
CHIARA DALL'ASTA
MICHELE SUMAN

The authors whose names are listed immediately below report the following details of affiliation or involvement in an organization or entity with a financial or non-financial interest in the subject matter or materials discussed in this manuscript. Please specify the nature of the conflict on a separate sheet of paper if the space below is inadequate.

Author names:

To the kind attention of:
Editor

Dr. Q. Rao, PhD
(Food Science; Food Chemistry; Food Quality; Food Safety)
Florida State University College of Human Sciences Nutrition,
Food & Exercise Sciences, Tallahassee, Florida, United States of America
Food Control - Elsevier

Parma, 14th October 2021

Dear Editors,
with reference to the manuscript entitled "Detection of soft-refined oils in extra virgin olive oil using data fusion approaches for LC-MS, GC-IMS and FGC-Enose techniques: the winning synergy of GC-IMS and FGC-Enose", which we have submitted to your attention, we would like, as requested, to indicate the correspondent suggested highlights:

Highlights

- Extra virgin olive oil (EVOO) can be adulterated by mixing it with soft refined oils (SROO)
- LC-MS, GC-IMS and FGC-ENose were evaluated for their fraud detection potentialities
- Low-level and mid-level data fusion of those analytical dataset were performed
- The discriminatory capability of the two merged GC-based techniques was significantly improved
- Combining GC-based techniques, data fusion and a PLS-DA-SVM strategy provides a new framework for effective authentication of EVOO

Please do not hesitate to contact me for any other needs.
Best Regards.
Yours sincerely.
Michele Suman

Dr. Michele Suman, PhD

Barilla G.R. F.lli SpA
Research, Development & Quality
Food Safety & Authenticity Research Manager
Food Safety Fellow Technical Ladder
Adjunct Professor of AgriFood Authenticity at Catholic University Sacred Heart – Milan/Piacenza
Chair ILSI Process Related Compounds & Natural Toxins Task Force
Chair Italian National Normative Organization (UNI) - Food Authenticity Commission
Scientific Board Member Italian Chemistry Society-Food Chemistry Inter-divisional Group
Via Mantova 166 - 43100 Parma (Italy)
☏ phone **+39 0521 262332**
☏ mobile **+39 3386938349**

✉ mail **michele.suman@barilla.com**
web **www.barillagroup.com**

To the kind attention of:
Prof. Andrea Armani, DVM, PhD,
Dipartimento di Scienze Veterinarie
Viale delle Piagge, 2, 56124 Pisa (PI)
e-mail: andrea.armani@unipi.it
Editor Food Control / Elsevier

Ms Ichiko Charis Howells
On Behalf of the Editorial Board - Food Control

Authors:
*Alessandra Tata [a], Andrea Massaro [a], Tito Damiani [b], Roberto Piro [a], Chiara Dall'Asta [b], Michele Suman [c,d] **

[a] Istituto Zooprofilattico Sperimentale delle Venezie, Laboratorio di Chimica Sperimentale, Vicenza, Italy
[b] Department of Food and Drug, University of Parma, Parma, Italy
[c] Department of Analytical Food Science, Barilla G. e R. Fratelli S.p.A., Parma, Italy
[d] Department for Sustainable Food Process, Catholic University Sacred Heart, Piacenza, Italy

Dear Editor,
with reference to the manuscript entitled "Detection of soft-refined oils in extra virgin olive oil using data fusion approaches for LC-MS, GC-IMS and FGC-Enose techniques: the winning synergy of GC-IMS and FGC-Enose", which we have submitted to your attention, we would like, as requested, to make the following CRediT Statements:

CRediT author statement

Terms, Definition, Conceptualization: *Tata*, *Massaro*,

Ideas, formulation or evolution of overarching research goals and aims: *Tata, Piro, Dall'Asta, Suman*

Methodology, Development or design of methodology; creation of models: *Tata*, *Massaro*

Validation, Verification, whether as a part of the activity or separate, of the overall replication/ reproducibility of results/experiments and other research outputs: *Tata*, *Massaro*, *Damiani*

Application of statistical, mathematical, computational, or other formal techniques to analyze or synthesize study data: *Tata*, *Massaro*

Investigation, Conducting a research and investigation process, specifically performing the experiments, or data/evidence collection: *Tata*, *Massaro*, *Damiani*

Writing - Original Draft, Preparation, creation and/or presentation of the published work, specifically writing the initial draft: *Tata*, *Massaro*

Writing - Review & Editing: *Damiani, Piro, Dall'Asta, Suman*

Preparation, creation and/or presentation of the published work, specifically visualization/ data presentation: *Tata, Massaro*

Supervision, Oversight and leadership responsibility for the research activity planning and execution, including mentorship external to the core team: *Piro, Dall'Asta, Suman*

Project administration, Management and coordination responsibility for the research activity planning and execution: *Piro, Dall'Asta, Suman*

Please do not hesitate to contact me for any other needs.
Yours sincerely.
On behalf of all the authors
Michele Suman

Parma, 14th October 2021

Dr. Michele Suman, PhD
Barilla G.R. F.lli SpA
Research, Development & Quality
Food Safety & Authenticity Research Manager
Adjunct Professor of AgriFood Authenticity at Catholic University Sacred Heart – Milan/Piacenza
Chair ILSI Process Related Compounds & Natural Toxins Task Force
Chair Italian National Normative Organization (UNI) - Food Authenticity Commission
Via Mantova 166 - 43100 Parma (Italy)
☏ mobile **+39 3386938349**
✉ mail **michele.suman@barilla.com**
web **www.barillagroup.com**

To the kind attention of:
Prof. Andrea Armani, DVM, PhD,
Dipartimento di Scienze Veterinarie
Viale delle Piagge, 2, 56124 Pisa (PI)
e-mail: andrea.armani@unipi.it
Editor Food Control

Ms Ichiko Charis Howells
On Behalf of the Editorial Board - Food Control
Food Control - Elsevier

Authors:

*Alessandra Tata [a], Andrea Massaro [a], Tito Damiani [b], Roberto Piro [a], Chiara Dall'Asta [b], Michele Suman [c,d] **

[a] Istituto Zooprofilattico Sperimentale delle Venezie, Laboratorio di Chimica Sperimentale, Vicenza, Italy
[b] Department of Food and Drug, University of Parma, Parma, Italy
[c] Department of Analytical Food Science, Barilla G. e R. Fratelli S.p.A., Parma, Italy
[d] Department for Sustainable Food Process, Catholic University Sacred Heart, Piacenza, Italy

Title:

"Detection of soft-refined oils in extra virgin olive oil using data fusion approaches for LC-MS, GC-IMS and FGC-Enose techniques: the winning synergy of GC-IMS and FGC-Enose"

Answers to the comments and suggestions from the reviewers – FOODCONT-D-21-00923R1

Dear Editor,
with reference to your opinion on publication of the present work, we would like to thank again for the valuable final review we received. We are honored that this submitted paper can be accepted for publication based on last fine tunings accordingly to the reviewer(s)' comments.
Therefore, the manuscript has been modified according to these reviewers' requests. The detailed responses to the comments and suggestions are reported here below.

Reviewer 1 Comments:

The manuscript can be accepted after this revision, but the title should be changed since, it seems that confirm the utility of data fusion of data obtained from LC-MS, GC-IMS and FGC-Enose techniques for the detection of soft refined oils in extra virgin olive oil. The authors should revise the manuscript because the way of presenting the results can be confusing since till the end of the manuscript it cannot be found that the combination of GC-IMS and FGC-Enose fingerprints using a low-level data fusion approach is the most powerful classification tool.

Response: The reviewer has raised an interesting point, therefore, we modified the title accordingly and we made clear in the abstract that the low-level combination of GC-IMS and FGC-Enose is the most powerful.

Resultant changes to the title: "Detection of soft-refined oils in extra virgin olive oil using data fusion approaches for LC-MS, GC-IMS and FGC-Enose techniques: the winning synergy of GC-IMS and FGC-Enose"

Resultant changes to the abstract: "The most promising results were achieved by the low-level data fusion of GC-IMS and FGC-Enose data."

Reviewer 3 Comments:

I really appreciate that all my comments addressed in the first review have been properly explained by the authors.

Yet, I regret to say that I do not agree with the authors comment: "Having a small dataset (60 samples) due to reasons explained above, therefore, the proportion of our data split was based on the concept that the more training data we have, the better our model will be. In other words, big training data maximizes the performance of the model and provides higher confidence in the resulting accuracy".

The key point for both the training and test set is to be representative of the case under study.

Regarding the training set, it should contain as many samples as required to proper cover the data variability. Let's us say (just as an example) that with 20 samples all the variability is considered, therefore 20 samples are enough. There are several papers/algorithms that deal with training sample selection, such as Kennard-Stone, PCA score distribution, etc. The final number of training samples is strong depending on the sample/data distribution, whether it is homogeneous or heterogeneous. I am aware that it is not a simple decision, but models build with lower number of samples (in order to increase the test set) might be checked.

Test set is used to check the performance of the model, if not enough test samples are used, the performance values based on the test set are not reliable. If that the case, (as it happens in that paper with only 6 test samples) then the best option is to used cross-validation instead of an independent test set.

Response: We thank the reviewer for raising this interesting point. Actually, we tested the models with the same independent samples used in the previous studies from Damiani et al 2020 and Cavanna et al. 2020. Indeed, the three authentic EVOO samples of the test set were previously selected with a Kennard-Stone algorithm, while the other three "NOT EVOO" samples (DEO3, DEO_DEA2, and Mix D) chosen with the aim to predict both pure adulterated samples and mixtures. In order to clarify this point, we added this info to the manuscript. We also removed from the manuscript the comment related to the "concept that big training data maximizes the performance of the model and provides higher confidence in the resulting accuracy on test set"

Resultant changes to material and methods: "The test set was comprised of three authentic EVOO (CP-30, CP-31, CP-32) and three SROO (DEO3, DEO_DEA2, MIX_D) as previously done by Cavanna et al 2020 and Damiani et al 2020. The three authentic EVOO samples of the test set were selected with a Kennard-Stone algorithm, while the other three "NOT EVOO" samples (DEO3, DEO_DEA2, and Mix D) chosen with the aim to predict both pure adulterated samples and mixtures (Cavanna et al 2020)."

Parma, 14th October 2021

On behalf of all the authors. Best regards.

Dr. Michele Suman, PhD
Barilla G.R. F.lli SpA
Research, Development & Quality
Food Safety & Authenticity Research Manager
Adjunct Professor of AgriFood Authenticity at Catholic University Sacred Heart – Milan/Piacenza
Chair ILSI Process Related Compounds & Natural Toxins Task Force
Chair Italian National Normative Organization (UNI) - Food Authenticity Commission
Via Mantova 166 - 43100 Parma (Italy)
☏ mobile **+39 3386938349**
✉ mail **michele.suman@barilla.com**
💻 web **www.barillagroup.com**

Detection of soft-refined oils in extra virgin olive oil using data fusion approaches for LC-MS, GC-IMS and FGC-Enose techniques: the winning synergy of GC-IMS and FGC-Enose

Alessandra Tata [a], Andrea Massaro[a], Tito Damiani[b], Roberto Piro[a], Chiara Dall'Asta[b], Michele Suman[c,d] *

[a]Istituto Zooprofilattico Sperimentale delle Venezie, Laboratorio di Chimica Sperimentale, Vicenza, Italy

[b] Department of Food and Drug, University of Parma, Parma, Italy

[c] Department of Analytical Food Science, Barilla G. e R. Fratelli S.p.A., Parma, Italy

[d] Department for Sustainable Food Process, Catholic University Sacred Heart, Piacenza, Italy

*Corresponding author: Dr. Michele Suman Advanced Research Laboratory, Barilla G. e R. Fratelli S.p.A., Parma, Italy

email: michele.suman@barilla.com; michele.suman@unicatt.it

Abbreviations: LC-MS: liquid chromatography-mass spectrometry; GC-IMS: gas-chromatography ion mobility spectrometry; FGC-Enose: flash gas-chromatography electronic nose; EVOO Extra Virgin Olive Oil; SROO: soft-refined olive oil; PLS-DA, Partial Least Squared Discriminant Analysis; SVM, support vector machine; ROC curve, Receiver Operating Characteristic curve; AUC area under the curve.

Keywords: Soft-deodorization, deacidification, FGC-Enose, adulteration, GC-IMS, LC-MS, soft-refined olive oil (SROO), data fusion

Abstract

Extra virgin olive oil (EVOO) is frequently adulterated by mixing it with soft refined oils (SROO). The differentiation of EVOO from its blends with SROO is not possible with the most common approaches, and, for this reason, the discriminating power of liquid chromatography-high resolution mass spectrometry (LC-MS), gas-chromatography ion mobility spectrometry (GC-IMS) and flash gas-chromatography electronic nose (FGC-Enose) was examined previously. Here, the combination of the above-mentioned techniques for an improvement in classification power of the methods is explored.

A total of 43 commercial EVOOs and 18 illegal mixtures of SROO with EVOO were previously analysed by LC-(+/-)MS , GC-IMS and FGC-Enose. Low-level and mid-level data fusion of the four datasets were performed. The merged unique fingerprints were submitted to partial least squared discriminant analysis (PLS-DA), and the extrapolated most informative variables were used to build support vector machine (SVM) classifiers. Statistical indicators were calculated and compared to find out the best classifier. The results of PLS-DA-SVM strategies on the combination of datasets demonstrated that, after low-level data fusion, the discriminatory capability of the two merged GC-based techniques was remarkably improved as compared to the individual techniques. This indicates that merging the datasets before PLS-DA better retrieves the most informative variables and, thus, enhances group separation and classification of unknowns. The combination of LC(+/-)MS datasets, both by mid- and low-level data fusion, did not show significant enhancement in terms of discrimination of EVOO from SROO as compared to the individual LC(+)MS matrix. The low-level combination of the four datasets (LC(+/-)MS, GC-IMS, FGC-Enose) was successful, although this laborious option is not a viable path in industry quality assurance.

This study primarily provides new paths for the authentication of EVOO, taking advantage of merging multimodal LC-(+/-)MS, GC-IMS and FGC-Enose data, with consequent

improvement in the performances of the classification models. The most promising results were achieved by the low-level data fusion of GC-IMS and FGC-Enose data.

1. Introduction

Due to its high economic value and unique sensorial and nutritional characteristics, extra virgin olive oil (EVOO) is considered at high risk of fraud(Casadei *et al.*, 2021). Recently, more sophisticated adulterations have been developed. The mixtures of EVOO with soft deacidified and soft deodorized olive oils are considered the most critical frauds because they are not easily detectable by regular methods(Conte *et al.*, 2020).

The detection of soft refined products in EVOO has been recently attempted by near infrared (NIR) spectroscopy (Gertz, Matthäus and Willenberg, 2020) and diacylglycerol determination(Gómez-Coca *et al.*, 2020). Recently, the adulteration of EVOO with soft-refined olive oil (SROO) has raised the interest of our research group, as four non-targeted methods capable of detecting this fraud were developed and validated separately; these were liquid chromatography-mass spectrometry (LC-MS) in positive and negative ion mode, gas-chromatography ion mobility spectrometry (GC-IMS) and flash gas-chromatography electronic nose (FGC-Enose) (Damiani *et al.*, 2020; Cavanna *et al.*, 2020).

Data fusion is a chemometric technique that merges the outcomes of multiple analytical sources. It has recently emerged as an attractive means to enhance the prediction power of a model for food authentication (Callao and Ruisánchez, 2018; Hu *et al.*, 2019; Márquez *et al.*, 2016). Low-level data fusion is a valuable chemometric strategy capable of concatenating multiple datasets and improving the classification performances by retrieving the discriminative variables from different techniques (Andrade *et al.*, 2021). Mid-level data fusion aims at merging datasets by reducing their high dimensionality and teasing out solely the most informative variables capable of codifying each group in the study (Jandric *et al.*, 2021; Tata *et al.*, 2021; Riuzzi *et al.*, 2021).

Data fusion models were applied to EVOO for the detection of its adulteration with vegetable oils (Schwolow *et al.*, 2019; Li, Xiong and Min, 2019), the assessment of its geographical origin (Casale *et al.*, 2010a; Casale *et al.*, 2012; Pizarro *et al.*, 2013; Nescatelli *et al.*, 2014; Bajoub *et al.*, 2017) and the reveal of sensory defects (Borràs *et al.*, 2016).

Most of the common data fusion models applied to EVOO have merged data from analytical techniques that provide similar information, such as Raman, near infrared and medium infrared spectroscopies (Casale *et al.*, 2010b; Li, Xiong and Min, 2019; Pizarro *et al.*, 2013; Bevilacqua *et al.*, 2013; Jiménez-Carvelo, Lozano and Olivieri, 2019; Casale *et al.*, 2012; Bragolusi *et al.*, 2021) or chromatographic profiles recorded at three different wavelengths (Nescatelli *et al.*, 2014). On the other hand, data fusion could be very useful when complementary information is fused and included in one unique model (Schwolow *et al.*, 2019; Assis *et al.*, 2019; Borràs *et al.*, 2016; Casale *et al.*, 2010a; Casale *et al.*, 2007).

In the present study, data from the three complementary techniques, each of them characterized by distinct information (volatile and non-volatile chemical profiles) were merged by low and mid-level data fusion for the discrimination of authentic EVOO and fraudulent SROO blends. Although promising results have been achieved in food authentication assessment (Damiani *et al.*, 2020; Cavanna *et al.*, 2020), reports on the combination of data from different mass spectrometric techniques for the improvement of detection of the SROO blends are still limited. The present study aimed to evaluate the enhanced prediction power obtained by low-level and mid-level data fusion and outline any possible disadvantages.

The comparison was carried out through the estimation of statistical indicators, i.e., accuracy, sensitivity, specificity, for a training set and probability of predictions for a set of validation samples. To the best of our knowledge, this is the first study exploring data fusion strategies for the detection of SROO blends in EVOO.

2. Materials and methods

2.1 Dataset collection and analysis

The datasets used for this study were acquired in our previous studies (Damiani *et al.*, 2020; Cavanna *et al.*, 2020). Therefore, all the details about sample collection and analyses are reported in detail in our previous publications.

Briefly, a total of 43 commercial Italian EVOOs, obtained over three harvest seasons (i.e., 2015/2016, n = 18; 2016/2017, n = 8; 2017/2018, n = 17), were considered as authentic samples. In addition, soft-deodorization and deacidification were carried out on commercial virgin and lampante olive oils to create counterfeit soft-refined samples (SROO).

In order to create counterfeited samples potentially compliant with the legislation, the official EVOO physic-chemical quality parameters(Regulation, 2016) were analysed in these refined oils.

Based on the obtained results, 18 illegal blends were prepared at different percentages by mixing the so-obtained SROO with authentic EVOOs randomly chosen from the sample set.

Authentic and counterfeit olive oil samples were analysed using three different techniques, namely GC-IMS, FGC-Enose, and LC-(+/-)MS.

Partially satisfactory classification models were obtained from the separate volatile profiles (Damiani et al. 2020) and from the LC-MS profiles (Cavanna et al. 2020).

2.2 Data fusion strategies and multivariate statistical analysis

In order to improve the prediction of authentic and adulterated EVOO, LC-(+/-)MS, GC-IMS and FGC-Enose data were merged via both low level and mid-level data fusion strategies using RStudio 3.6.2 and Metabonalyst 5.0 web platform.

2.2.1 Low-level data fusion

Each dataset was pre-processed by removing the C^{13} isotopes and the *m/z* ions with more than 75% of non-acquired intensities (missing values) across all the samples. Each dataset was normalized by sum and scaled by Pareto. Each pre-processed dataset was split into training set (55 samples) and test set (6 samples). The test set was comprised of three authentic EVOO (CP-30, CP-31, CP-32) and three SROO (DEO3, DEO_DEA2, MIX_D) as previously done by Cavanna *et al* 2020 and Damiani *et al* 2020. The three authentic EVOO samples of the test set were selected with a Kennard-Stone algorithm, while the other three "NOT EVOO" samples (DEO3, DEO_DEA2, and Mix D) chosen with the aim to predict both pure adulterated samples and mixtures (Cavanna et al 2020).

Low-level data fusions of: i) two LC-MS instrumental ion modes; ii) GC-IMS and FGC-Enose, and; iii) multimodal LC-MS and FGC-Enose and GC-IMS were performed.

The pre-processed signals of each training set were simply concatenated, mean-centered and processed as a unique fingerprint of the samples.

The merged training sets were submitted to the supervised partial least squared discriminant analysis (PLS-DA) with the aim of extrapolating the most informative variables.

The PLS-DA variables with coefficients >55 were retained and used to construct the linear SVM classification models which was validated on the merged test set.(Massaro *et al.*, 2021)

The criterion used to extrapolate the most significant features was based on the inspection of PLS-DA coefficient plot (not shown) reporting the informative variables in a descending order (from the one with highest coefficient to that with the lowest).

The "elbow" of the graph, where the coefficient of the informative variables leveled off, was considered as limit point.

The variables placed to the right of this point, corresponding to coefficient equal to 55, were retained as significant.

2.2.2 Mid-level data fusion

Briefly, each pre-processed dataset (split into training and test sets) was submitted to supervised PLS-DA. We selected the first five components of the PLS-DA of each dataset and we retrieved from them the most significant variables. As recommended by Hair et al (Hair *et al.*, 2006) only the ions with absolute values for PLS-DA loadings >0.3 were retained and used to build the SVM classification models. Further details of the mid-level data fusion strategy adopted can be found elsewhere (Massaro *et al.*, 2021)

2.2.3 Validation of the classification model

Support vector machine (SVM) classification models were built with the extrapolated molecular features using the Biomarker Analysis section of Metaboanalyst 5.0 after low-level and mid-level data fusions. Each SVM model was cross-validated by Monte Carlo cross validation (MCCV) using a repeated, balanced sub-sampling procedure. In details, the MCCV split training data in 2/3 for training the model and 1/3 for testing it.

For each iteration, the training/test split was different. In the first iteration, the model was tested on training data and test errors were calculated. After 100 iterations, the average of the test errors was determined and sensitivity (true positive rate), specificity (true negative rate) and accuracy were calculated.

The overall prediction power of the SVM models was estimated based on the area under the curve (AUC) of the receiver operating characteristic (ROC) curve. Finally, the SVM models were tested for their ability to classify six samples from the merged test set that was withheld previously.

3. Results

The combination of analytical methods (LC(+/-)-MS, GC-IMS and FGC-Enose) was assessed to evaluate possible improvements in discriminating of EVOO from their blends with SROO. First, a low-level data fusion of LC(+/-)-MS datasets was conducted. The resultant global data matrix was split into training and test sets. The training set was submitted to multivariate statistical analysis by means of PLS-DA (**Figure 1A**). A good trend of separation was observed, with Component1 and Component2 capable of explaining 35.7 % and 10.2% of the data variance, respectively (**Figure 1B**). The *m/z* values and associated retention times with a higher discriminatory capacity (coefficient >55) were retained and used to build a SVM classifier. The SVM model was cross-validated by MCCV on the training set (**Figure 1A**, right side) with accuracy, sensitivity and specificity reaching 0.94, 0.93 and 0.95 respectively (**Table 1**). The ROC curve, a graph plotting true positive and false positive rates of the SVM classification model at all classification thresholds, showed an AUC equal to 0.97 (**Figure 1C**). These excellent accuracy, sensitivity and specificity parameters increased in the blinded verification which was able to correctly classify 6/6 samples. The results of the predictions on the test set, and the correlated probabilities, can be visualised in **Table S1** of the supplementary material. The averaged probability of all samples is above 96%.

Subsequently the combination of GC-IMS and FGC-Enose approaches by low-level data fusion was evaluated. To this aim, GC-IMS and FGC-Enose datasets were both split into training and test sets, concatenated, and PLS-DA was performed on the fused data (**Figure 2A**). The PLS-DA score plot is reported in **Figure 2B** with the EVOO samples grouped decently by Component1 and Component2. The SVM model, built with the selected variables, showed an accuracy, sensitivity and specificity on the training set of 0.96, 0.93 and 0.97, respectively, and an AUC of the ROC curve equal to 0.99 (**Table 1** and **Figure 2C**). The SVM correctly classified

6/6 samples in the test set with an averaged probability above 93%, although with a low probability of predicting one sample (MIX_D) (**Table S2**).

Finally, the LC(+/-)MS, GC-IMS and FGC-Enose datasets were merged by a low-level data fusion approach. Compared to the previous two techniques, the score plot showed improved clustering of the two groups (authentic and non-authentic EVOO) in the study, with the first and second components, C1 and C2, explaining 26.9% and 11.0 % of the total variance of the model, respectively (**Figure 3**).

The results of the cross-validation of SVM, built with the variables with coefficient >55 retrieved from fused-PLS-DA, are shown in **Table 1**. In this case, the SVM model built with the combination of the most informative variables of the three techniques reached an accuracy, sensitivity and specificity on the training set of 0.96, 0.93 and 0.97 respectively and an AUC of the ROC curve equal to 0.98 (**Table 1** and **Figure 3C**). The SVM correctly classified 6/6 samples in the test set with an averaged probability above 93%, although with a low probability of predicting the sample MIX_D (**Table S3**).

Mid-level data fusion was also attempted for the alternated combination of all four datasets (**Figures S1, S2** and **S3** of the supplementary material), with less satisfactory results, especially in terms of the ROC curve in cross-validation and the probability of predictions for the test set, as compared to the low-level data fusion.

For this reason, a summary of mid-level data fusion results of the cross-validation of the SVM models and their validation on the merged test set are only shown in the supplementary material (**Tables S4-S7**).

Note that the best classification performances in this case were achieved by the mid-level data fusion of the two LC matrices. (**Table S4**). With the mid-level data fusion of the four datasets less trustable classifier was obtained (**Table S4**).

4. Discussion

In previous studies, GC-IMS, FGC-Enose and LC(+/-)MS datasets were statistically analysed separately. Headspace-based techniques (i.e., GC-IMS, FGC-Enose) showed great potential as rapid screening platforms and exhibited remarkable reproducibility over the time; yet, the EVOO's volatile fingerprint seemed to be heavily affected by chemical changes occurring in ordinary shelf-life conditions. On the other hand, LC-MS enabled the identification of fraud-specific markers; however, it suffers of limited sensitivity (i.e., fraud detected at >40% SROO addition). In this study, we want to explore the possibility of merging the data and evaluate possible improvements in the discrimination of genuine EVOO from SROO. In particular, the main aim was to provide a robust data fusion approach to be coupled with quick fingerprint analysis that could be applied in an industrial environment for rapid EVOO authentication. Low-level fusion was first used to pick up correlations between variables of different blocks of data. Low-level fusion is based on the simple concatenation of data to which a single model is applied to pick up correlations between variables belonging to different datasets (Biancolillo *et al.*, 2014; Borràs *et al.*, 2015). It has the limitations of high volume of features, which is difficult to handle, and the possible predominance of one data source over the others. In order to exclude this possible issue, we checked the number of variables of each dataset. We had a thousand variables in each LC-MS dataset and a total of one-hundred thirty variables in the GC matrices. Besides the predominance of the LC-MS source, the difference

in block sizes did not affect the PLS-DA weighting of the GC variables that appear as the most significant features in low-level data fusion of the four datasets.

On the other hand, mid-level data fusion is characterized by an initial high dimensional data reduction, by means of supervised or unsupervised tools capable of extracting the most informative variables from each separate dataset (Pirro *et al.*, 2014; Borràs *et al.*, 2015).

After both low-level and mid-level data fusion, PLS-DA–SVM strategies were applied to concatenated datasets to obtain classification rates for cross-validation and validation on the test set. The SVM models that followed the mid-level data fusion provided less powerful classification, and for this reason, results were included in SI only, and are not discussed further.

In the individual techniques, the LC(+)MS profiles showed high accuracy, R^2 and Q^2 (Cavanna et al., 2020). The accuracy is the capability of the model to correctly classify the samples, the R^2 parameter indicates the goodness of fit of the PLS-DA model (how well it explains the dataset) and Q^2 provides a measure of exactness between the predicted and actual data (Triba et al., 2015; Worley and Powers, 2013). Further details are reported elsewhere (Anderssen et al., 2006; Westerhuis et al., 2008). It is worth noting that LC-MS is a highly informative technique that can be used for the identification of chemical markers to be further used in target analysis. While being extremely powerful, this approach is costly and requires high-level laboratory skills. Its application in an industrial environment is, therefore, suggested only for explorative analysis or for confirmatory purposes, whereas it cannot be applied for routine controls. Although in the present study we used linear SVM as the classifier instead of PLS-DA (PLS-DA was employed just to extrapolate the most informative variables used to build the classification model), it does not seem that either the mid or the low-level combination of the two datasets resulted in improvements to the classification figures of merit. However, the performance obtained from the fusion of the LC-(+/-)MS can be regarded as a benchmark for evaluating the discrimination potential shown by the data fusion applied to volatile fingerprints.

In the individual techniques, the soft independent modelling by class analogy (SIMCA) models developed on the GC-IMS and FGC-Enose fingerprint datasets were able to correctly recognize the SROO blends as non-authentic products, even at the lowest adulteration percentage (i.e. 10%) (Damiani *et al.*, 2020). Only one EVOO sample was wrongly recognized as not EVOO, confirming the high potential of the two separately employed techniques (Damiani *et al.*, 2020). After the application of low-level data fusion, the SVM model developed herein achieved extremely high sensitivity, specificity and accuracy with fully correct predictions for the test set. In contrast to our previous study, we were able to include EVOO 15/16 (CP_1-CP_12), oils that negatively altered the performance of our previous SIMCA model (Damiani *et al.*, 2020). Therefore, the chemometric approach followed in the present work, and based on the fusion of both volatile fingerprint datasets, showed an improvement in the discrimination potential of the model compared to each technique alone. This fused dataset approach is able to overcome the difficulties related to partial overlap of EVOO's chemical features in the volatile fraction characteristics, thereby differentiating oil resulting from fraudulent practice from naturally aged oil subjected to long storage conditions.

When compared to the SVM model obtained by fusing LC-(+)MS and LC-(-)MS datasets, the quality parameters on the training set were slightly higher for the GC-fused model, while the probability of correct prediction in the validation test set was lower (0.93 versus 0.96 for GC-fused and LC-fused model, respectively), even though the same outcomes for sample classification were seen.

Overall, it can be concluded these the two models are almost comparable in terms of classification performances, although the GC-fused model showed undeniable advantages in terms of cost-effectiveness and ease of handling in an industrial quality control routine approach.

On the other hand, it must be underlined that MS offers the opportunity to identify the chemical markers responsible for classification, and to monitor them over time. Therefore, its superior use for explorative and confirmatory purposes is without question.

To gain a comprehensive overview of the potential of data fusion in EVOO classification, all four datasets were fused, and the resultant model was compared to the previous one in terms of performance.

In this case, the statistical indicators obtained in both mid- and low-level data fusion were still satisfactory, but lower than those obtained from the combination of the two GC-based approaches. We observed a low AUC when running the mid-level data fusion of the four datasets (**Table S4**).

On the other hand, considering the analytical and chemometric efforts required to collect and fuse datasets from four different techniques, with little to no improvement obtained in the overall model, this approach is far from offering a useful solution currently applicable within industrial production monitoring.

In conclusion, the combination of GC-IMS and FGC-Enose fingerprints using a low-level data fusion approach is the most powerful classification tool we know of to date that could be used for identifying soft refinement of EVOO in an industrial quality assurance setting. Notably, this approach is based on datasets obtained using cost-effective and easy-to-handle techniques.

5. Conclusion

Data fusion strategies to authenticate EVOO were tested taking into account of LC-(+/-)MS, GC-IMS and FGC-Enose analytical data. Specifically, low-level data fusion of GC-IMS and FGC-ENose datasets produces an optimal model for classifying SROO and EVOO with an overall accuracy of 0.96 and the advantage of the rapid acquisition of the spectra.

This approach, combining GC-based techniques, data fusion and a PLS-DA-SVM strategy, likely provides a new framework for the authentication of EVOO in a possible industrial quality assurance setting.

References

Andrade, D. F., de Almeida, E., de Carvalho, H. W. P., Pereira-Filho, E. R. and Amarasiriwardena, D. (2021) 'Chemical inspection and elemental analysis of electronic waste using data fusion - Application of complementary spectroanalytical techniques', (1873-3573 (Electronic)).

Assis, C., Pereira, H. V., Amador, V. S., Augusti, R., de Oliveira, L. S. and Sena, M. M. (2019) 'Combining mid infrared spectroscopy and paper spray mass spectrometry in a data fusion model to predict the composition of coffee blends', *Food Chemistry,* 281, pp. 71-77.

Bajoub, A., Medina-Rodríguez, S., Gómez-Romero, M., Ajal, E. A., Bagur-González, M. G., Fernández-Gutiérrez, A. and Carrasco-Pancorbo, A. (2017) 'Assessing the varietal origin of extra-virgin olive oil using liquid chromatography fingerprints of phenolic compound, data fusion and chemometrics', *Food Chemistry,* 215, pp. 245-255.

Bevilacqua, M., Bucci, R., Magrì, A. D., Magrì, A. L. and Marini, F. (2013) 'Data Fusion for Food Authentication. Combining near and Mid Infrared to Trace the Origin of Extra Virgin Olive Oils', *NIR news,* 24(2), pp. 12-15.

Biancolillo, A., Bucci, R., Magrì, A. L., Magrì, A. D. and Marini, F. (2014) 'Data-fusion for multiplatform characterization of an italian craft beer aimed at its authentication', *Analytica Chimica Acta,* 820, pp. 23-31.

Borràs, E., Ferré, J., Boqué, R., Mestres, M., Aceña, L. and Busto, O. (2015) 'Data fusion methodologies for food and beverage authentication and quality assessment – A review', *Analytica Chimica Acta,* 891, pp. 1-14.

Borràs, E., Ferré, J., Boqué, R., Mestres, M., Aceña, L., Calvo, A. and Busto, O. (2016) 'Olive oil sensory defects classification with data fusion of instrumental techniques and multivariate analysis (PLS-DA)', *Food chemistry,* 203, pp. 314-322.

Bragolusi, M., Massaro, A., Tata, A. and Piro, R. (2021) 'A data fusion model of NIR and RAMAN techniques for the geographical screening of Italian extra virgin olive oil', *NIRItalia online 2021*.

Callao, M. P. and Ruisánchez, I. (2018) 'An overview of multivariate qualitative methods for food fraud detection', *Food Control,* 86, pp. 283-293.

Casadei, E., Valli, E., Panni, F., Donarski, J., Farrús Gubern, J., Lucci, P., Conte, L., Lacoste, F., Maquet, A., Brereton, P., Bendini, A. and Gallina Toschi, T. (2021) 'Emerging trends in olive oil fraud and possible countermeasures', *Food Control,* 124, pp. 107902.

Casale, M., Armanino, C., Casolino, C. and Forina, M. (2007) 'Combining information from headspace mass spectrometry and visible spectroscopy in the classification of the Ligurian olive oils', *Analytica Chimica Acta,* 589(1), pp. 89-95.

Casale, M., Casolino, C., Oliveri, P. and Forina, M. (2010a) 'The potential of coupling information using three analytical techniques for identifying the geographical origin of Liguria extra virgin olive oil', *Food Chemistry,* 118(1), pp. 163-170.

Casale, M., Oliveri, P., Casolino, C., Sinelli, N., Zunin, P., Armanino, C., Forina, M. and Lanteri, S. (2012) 'Characterisation of PDO olive oil Chianti Classico by non-selective (UV–visible, NIR and MIR spectroscopy) and selective (fatty acid composition) analytical techniques', *Analytica Chimica Acta,* 712, pp. 56-63.

Casale, M., Sinelli, N., Oliveri, P., Di Egidio, V. and Lanteri, S. (2010b) 'Chemometrical strategies for feature selection and data compression applied to NIR and MIR spectra of extra virgin olive oils for cultivar identification', *Talanta,* 80(5), pp. 1832-1837.

Cavanna, D., Hurkova, K., Džuman, Z., Serani, A., Serani, M., Dall'Asta, C., Tomaniova, M., Hajslova, J. and Suman, M. (2020) 'A Non-Targeted High-Resolution Mass Spectrometry Study for Extra Virgin Olive Oil Adulteration with Soft Refined Oils: Preliminary Findings from Two Different Laboratories', *ACS Omega,* 5(38), pp. 24169-24178.

Conte, L., Bendini, A., Valli, E., Lucci, P., Moret, S., Maquet, A., Lacoste, F., Brereton, P., García-González, D. L., Moreda, W. and Gallina Toschi, T. (2020) 'Olive oil quality and authenticity: A review of current EU legislation, standards, relevant methods of analyses, their drawbacks and recommendations for the future', *Trends in Food Science & Technology,* 105, pp. 483-493.
Damiani, T., Cavanna, D., Serani, A., Dall'Asta, C. and Suman, M. (2020) 'GC-IMS and FGC-Enose fingerprint as screening tools for revealing extra virgin olive oil blending with soft-refined olive oils: A feasibility study', *Microchemical Journal,* 159, pp. 105374.
Gertz, C., Matthäus, B. and Willenberg, I. (2020) 'Detection of Soft-Deodorized Olive Oil and Refined Vegetable Oils in Virgin Olive Oil Using Near Infrared Spectroscopy and Traditional Analytical Parameters', *European Journal of Lipid Science and Technology,* 122(6), pp. 1900355.
Gómez-Coca, R. B., Pérez-Camino, M. d. C., Bendini, A., Gallina Toschi, T. and Moreda, W. (2020) 'Olive oil mixtures. Part two: Detection of soft deodorized oil in extra virgin olive oil through diacylglycerol determination. Relationship with free acidity', *Food Chemistry,* 330, pp. 127226.
Hu, O., Chen, J., Gao, P., Li, G., Du, S., Fu, H., Shi, Q. and Xu, L. (2019) 'Fusion of near-infrared and fluorescence spectroscopy for untargeted fraud detection of Chinese tea seed oil using chemometric methods', *Journal of the Science of Food and Agriculture,* 99(5), pp. 2285-2291.
Jandric, Z., Tchaikovsky, A., Zitek, A., Causon, T., Stursa, V., Prohaska, T. and Hann, S. (2021) 'Multivariate modelling techniques applied to metabolomic, elemental and isotopic fingerprints for the verification of regional geographical origin of Austrian carrots', *Food Chemistry,* 338, pp. 127924.
Jiménez-Carvelo, A. M., Lozano, V. A. and Olivieri, A. C. (2019) 'Comparative chemometric analysis of fluorescence and near infrared spectroscopies for authenticity confirmation and geographical origin of Argentinean extra virgin olive oils', *Food Control,* 96, pp. 22-28.
Li, Y., Xiong, Y. and Min, S. (2019) 'Data fusion strategy in quantitative analysis of spectroscopy relevant to olive oil adulteration', *Vibrational Spectroscopy,* 101, pp. 20-27.
Massaro, A., Stella, R., Negro, A., Bragolusi, M., Miano, B., Arcangeli, G., Biancotto, G., Piro, R. and Tata, A. (2021) 'New strategies for the differentiation of fresh and frozen/thawed fish: A rapid and accurate non-targeted method by ambient mass spectrometry and data fusion (part A)', *Food Control,* 130, pp. 108364.
Márquez, C., López, M. I., Ruisánchez, I. and Callao, M. P. (2016) 'FT-Raman and NIR spectroscopy data fusion strategy for multivariate qualitative analysis of food fraud', (1873-3573 (Electronic)).
Nescatelli, R., Bonanni, R. C., Bucci, R., Magrì, A. L., Magrì, A. D. and Marini, F. (2014) 'Geographical traceability of extra virgin olive oils from Sabina PDO by chromatographic fingerprinting of the phenolic fraction coupled to chemometrics', *Chemometrics and Intelligent Laboratory Systems,* 139, pp. 175-180.
Pirro, V., Oliveri, P., Ferreira, C. R., González-Serrano, A. F., Machaty, Z. and Cooks, R. G. (2014) 'Lipid characterization of individual porcine oocytes by dual mode DESI-MS and data fusion', *Analytica chimica acta,* 848, pp. 51-60.
Pizarro, C., Rodríguez-Tecedor, S., Pérez-del-Notario, N., Esteban-Díez, I. and González-Sáiz, J. M. (2013) 'Classification of Spanish extra virgin olive oils by data fusion of visible spectroscopic fingerprints and chemical descriptors', *Food Chemistry,* 138(2), pp. 915-922.
Regulation, E. (2016) '2095 (2016). Amending Regulation of EEC No: 2568/91', *Official Journal of the European Communities,* 326, pp. 1-6.
Riuzzi, G., Tata, A., Massaro, A., Bisutti, V., Lanza, I., Contiero, B., Bragolusi, M., Miano, B., Negro, A., Gottardo, F., Piro, R. and Segato, S.

(2021) 'Authentication of forage-based milk by mid-level data fusion of (+/−) DART-HRMS signatures', *International Dairy Journal,* 112, pp. 104859.
Schwolow, S., Gerhardt, N., Rohn, S. and Weller, P. (2019) 'Data fusion of GC-IMS data and FT-MIR spectra for the authentication of olive oils and honeys-is it worth to go the extra mile?', *Anal Bioanal Chem,* 411(23), pp. 6005-6019.
Tata, A., Pallante, I., Massaro, A., Miano, B., Bottazzari, M., Fiorini, P., Dal Prà, M., Paganini, L., Stefani, A., De Buck, J., Piro, R. and Pozzato, N. (2021) 'Serum Metabolomic Profiles of Paratuberculosis Infected and Infectious Dairy Cattle by Ambient Mass Spectrometry', *Frontiers in Veterinary Science,* 7(1214).

Detection of soft-refined oils in extra virgin olive oil using data fusion approaches for LC-MS, GC-IMS and FGC-Enose techniques: the winning synergy of GC-IMS and FGC-Enose.

Alessandra Tata, Andrea Massaro, Tito Damiani, Roberto Piro, Chiara Dall'Asta, Michele Suman

Supplementary material

Table S1. Classification independent adulterated (OTHER) and authentic extra virgin olive oil (EVOO) by the support vector machine (SVM) model, built using informative variables from low-level data fusion of multimodal high performance liquid chromatography-high resolution mass spectrometry (HPLC-HRMS) datasets.

SAMPLES	ACTUAL	PREDICTED	AVAREGED PROBABILITY
CP_30	AUTHENTIC	EVOO	0.96514
CP_31	AUTHENTIC	EVOO	0.92409
CP_32	AUTHENTIC	EVOO	0.99294
DEO3	ADULTERATED	OTHER	0.98983
DEO_DEA_2	ADULTERATED	OTHER	0.9958
MIX_D	ADULTERATED	OTHER	0.91528

Table S2. Classification independent adulterated (OTHER) and authentic extra virgin olive oil (EVOO) by the support vector machine (SVM) model, built using informative variables from low-level data fusion of gas chromatography coupled with ion mobility spectrometry (GC-IMS) and flash gas chromatography electronic nose (FGC-Enose) datasets.

SAMPLES	ACTUAL	PREDICTED	AVAREGED PROBABILITY
CP_30	AUTHENTIC	EVOO	0.99903
CP_31	AUTHENTIC	EVOO	0.95723
CP_32	AUTHENTIC	EVOO	0.98761
DEO3	ADULTERATED	OTHER	0.99879
DEO_DEA_2	ADULTERATED	OTHER	0.99938
MIX_D	ADULTERATED	OTHER	0.67833

Table S3. Classification independent adulterated (OTHER) and authentic extra virgin olive oil (EVOO) by the support vector machine (SVM) model, built using informative variables from low-level data fusion of gas chromatography coupled with ion mobility spectrometry (GC-IMS), flash gas chromatography electronic nose (FGC-Enose) and multimodal high performance liquid chromatography-high resolution mass spectrometry (HPLC-HRMS) datasets.

SAMPLES	ACTUAL	PREDICTED	AVAREGED PROBABILITY
CP_30	AUTHENTIC	EVOO	0.99753
CP_31	AUTHENTIC	EVOO	0.97704
CP_32	AUTHENTIC	EVOO	0.99371
DEO3	ADULTERATED	OTHER	0.99936
DEO_DEA_2	ADULTERATED	OTHER	0.99969
MIX_D	ADULTERATED	OTHER	0.65853

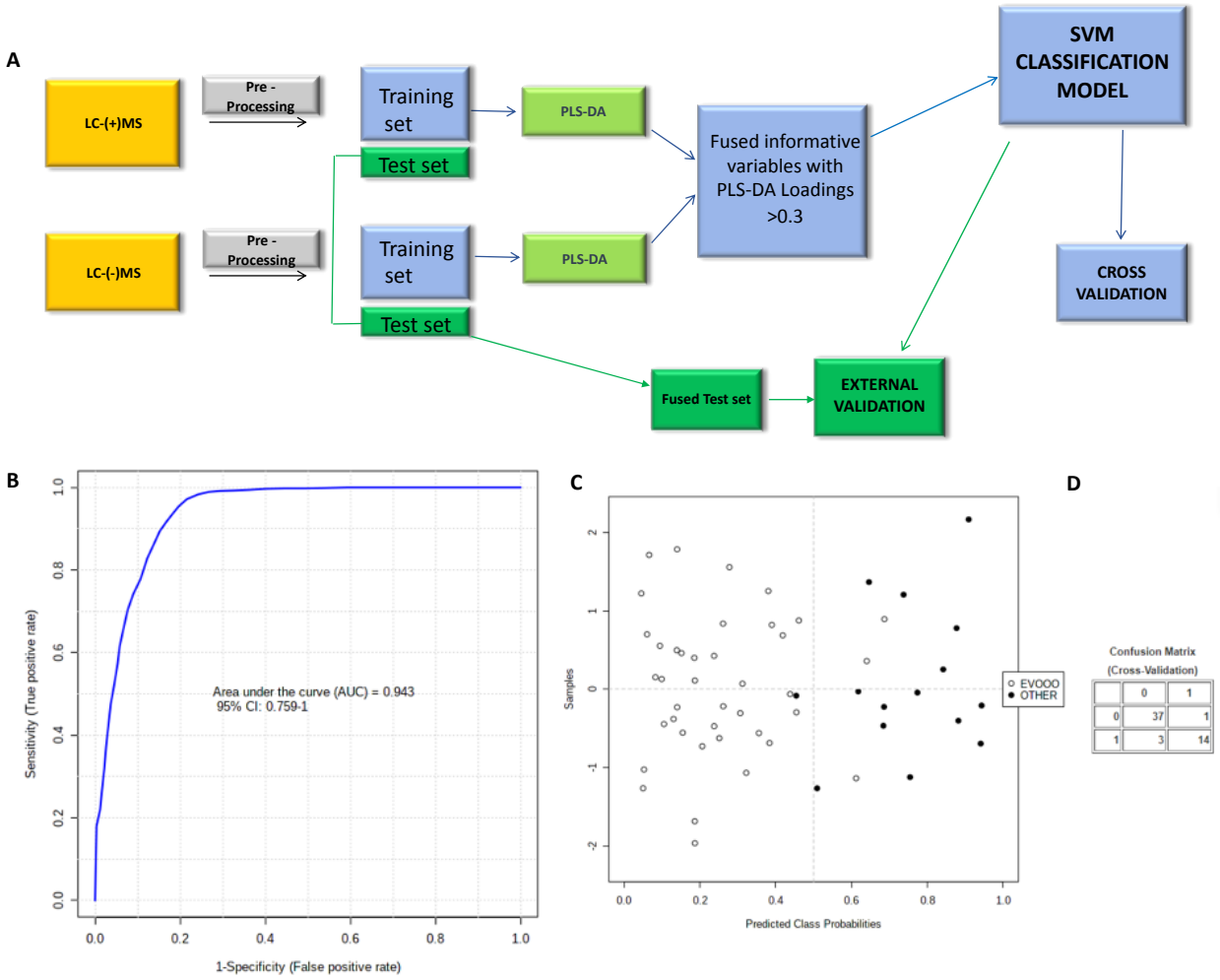


Figure S1. **Mid-level data fusion of multimodal high performance liquid chromatography-high resolution mass spectrometry (HPLC-HRMS) datasets and multivariate statistical analysis**. **aimed at the classification of extra virgin olive oil (EVOO)** A) Flow-chart of the mid level data fusion of multimodal HPLC-HRMS with extraction of the most informative variables by Partial least squares-discriminant analysis (PLS-DA)of the single datasets and built-in support vector machine (SVM). B) Receiver operating characteristic (ROC) the performance of a classification model in cross-validation on training set. C) The predictions of SVM model in the cross-validation with D) the resulting confusion matrix.

Table S4. *Accuracy Sensitivity specificity obtained by mid-level data fusion of GC-IMS and FGC-Enose and (+/)HPLC-HRMS datasets and built-in SVM model.*

Merged technique	Sensitivity on training set	Specificity on training set	Accuracy on training set	AUC of the ROC	Samples correctly classified in validation on test set	Probability of predictions in validation on test set
HPLC-(+/-)HRMS	0.93	0.93	0.93	0.94	6/6	0.93
GC-IMS FGC-Enose	0.86	0.98	0.95	0.90	6/6	0.88
GC-IMS FGC-Enose HPLC-(+/-)HRMS	0.93	0.98	0.96	0.88	6/6	0.94

Table S5. Classification independent adulterated (OTHER) and authentic extra virgin olive oil (EVOO) by the support vector machine (SVM) model, built using informative variables from mid-level data fusion of multimodal high performance liquid chromatography-high resolution mass spectrometry (HPLC-HRMS) datasets.

SAMPLES	ACTUAL	PREDICTED	AVAREGED PROBABILITY
CP_30	AUTHENTIC	EVOO	0.99749
CP_31	AUTHENTIC	EVOO	0.97675
CP_32	AUTHENTIC	EVOO	0.99858
DEO3	ADULTERATED	OTHER	0.73324
DEO_DEA_2	ADULTERATED	OTHER	0.96799
MIX_D	ADULTERATED	OTHER	0.94743

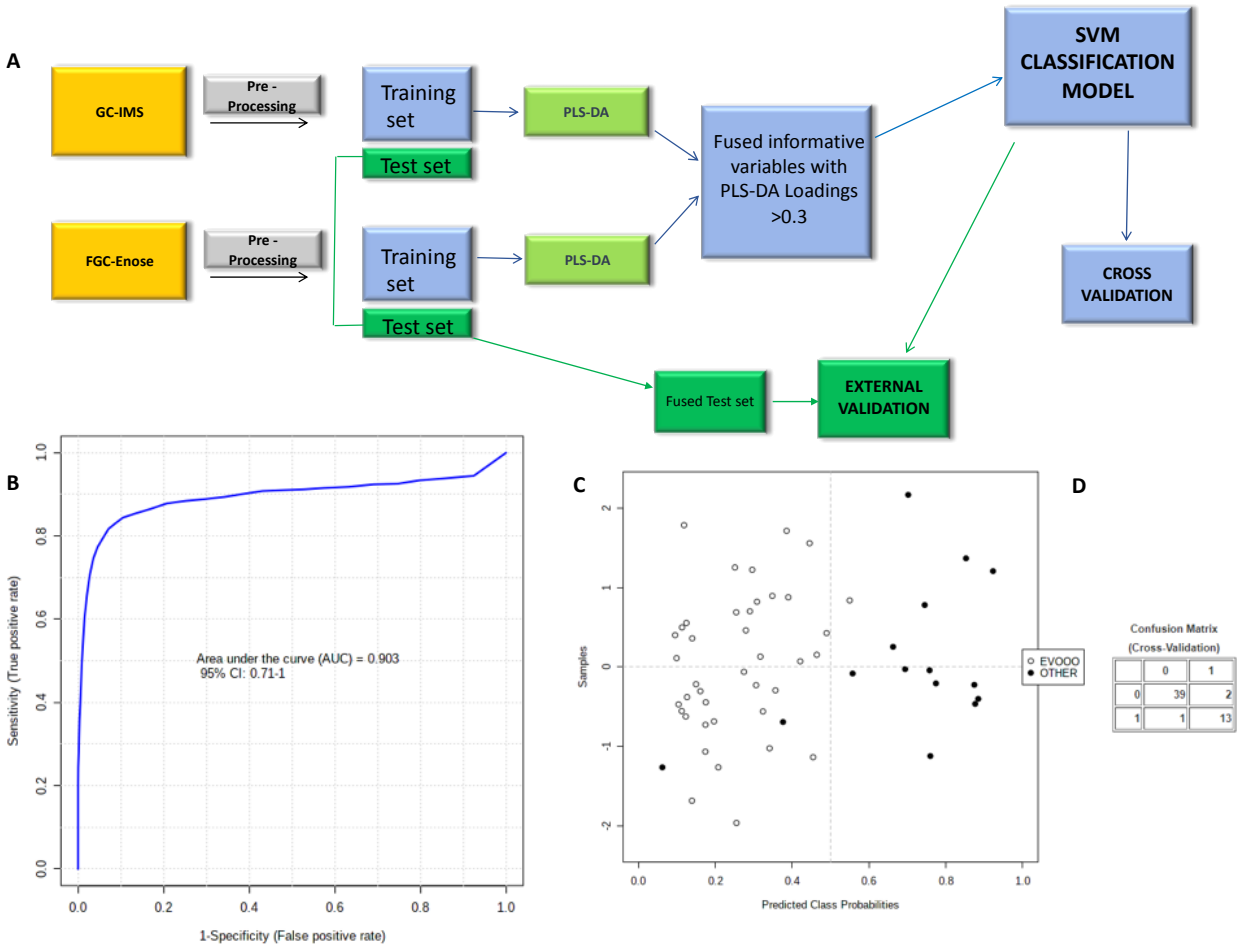


Figure S2. Mid-level data fusion of gas chromatography coupled with ion mobility spectrometry (GC-IMS) and flash gas chromatography electronic nose (FGC-Enose) datasets and multivariate statistical analysis aimed at the classification of extra virgin olive oil (EVOO). A) Flow-chart of the mid-level data fusion of GC-IMS and FGC-Enose with extraction of the most informative variables by Partial least squares-discriminant analysis (PLS-DA) of the single datasets and built-in support vector machine (SVM). B) Receiver operating characteristic (ROC) the performance of a classification model in cross-validation on training set. C) The predictions of SVM model in the cross-validation with D) the resulting confusion matrix.

Table S6. Classification independent adulterated (OTHER) and authentic extra virgin olive oil (EVOO) by the support vector machine (SVM) model, built using informative variables from mid-level data fusion of gas chromatography coupled with ion mobility spectrometry (GC-IMS) and flash gas chromatography electronic nose (FGC-Enose) datasets.

SAMPLES	ACTUAL	PREDICTED	AVAREGED PROBABILITY
CP_30	AUTHENTIC	EVOO	0.9968
CP_31	AUTHENTIC	EVOO	0.87557
CP_32	AUTHENTIC	EVOO	0.89127
DEO3	ADULTERATED	OTHER	0.99869
DEO_DEA_2	ADULTERATED	OTHER	0.95555
MIX_D	ADULTERATED	OTHER	0.60246

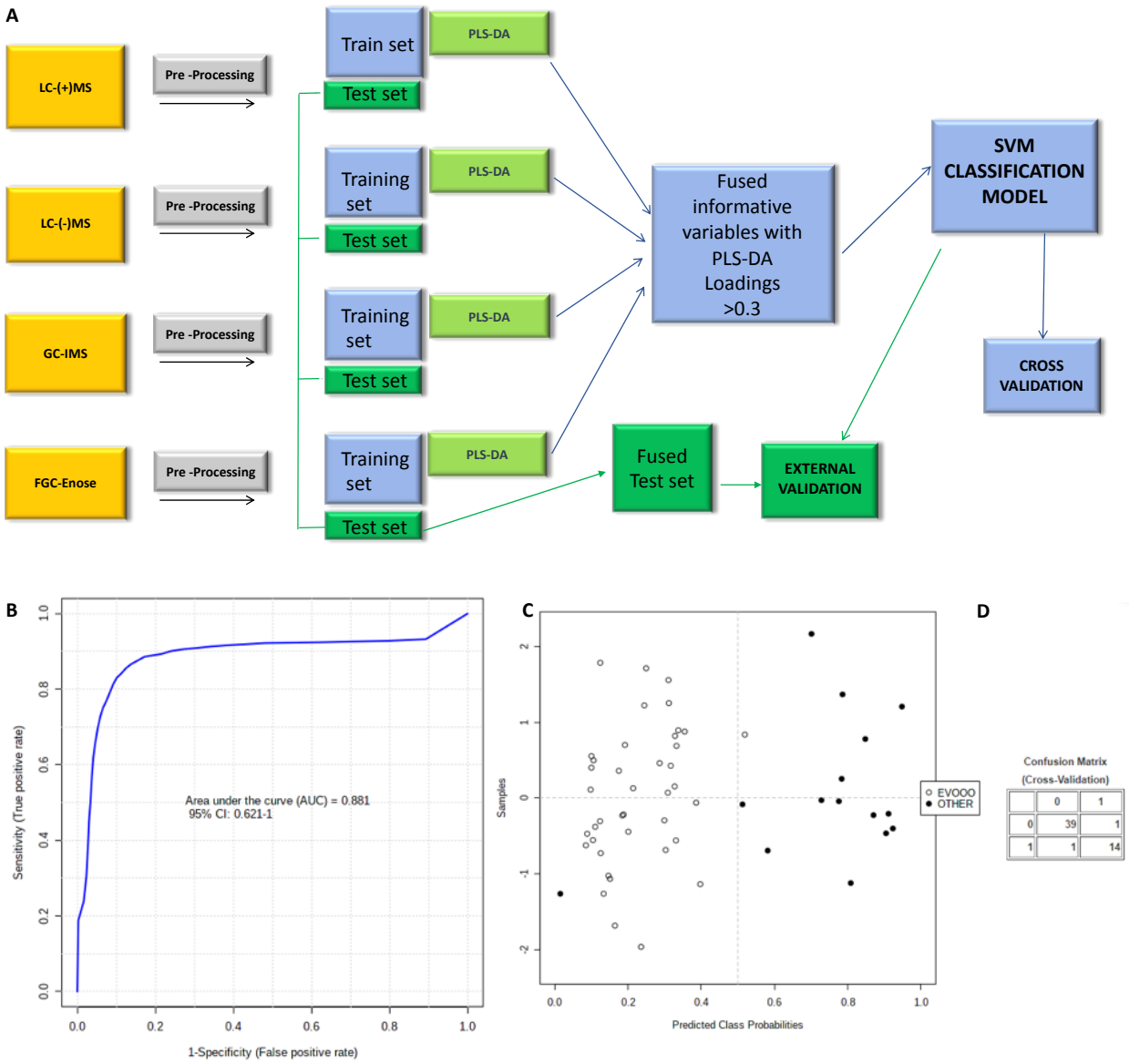


Figure S3. Mid-level data fusion of multimodal high performance liquid chromatography-high resolution mass spectrometry (HPLC-HRMS), gas chromatography coupled with ion mobility spectrometry (GC-IMS) and flash gas chromatography electronic nose (FGC-Enose) datasets coupled to multivariate statistical analysis aimed at the classification of extra virgin olive oil (EVOO). A) Flow-chart of the mid-level data fusion of the four datasets with extraction of the most informative variables by Partial least squares-discriminant analysis (PLS-DA) from each datasets and built-in support vector machine (SVM). B) Receiver operating characteristic (ROC) the performance of a classification model in cross-validation on training set. C) The predictions of SVM model in the cross-validation with D) the resulting confusion matrix.

Table S7. Classification independent adulterated (OTHER) and authentic extra virgin olive oil (EVOO) by the support vector machine (SVM) model, built using informative variables from low-level data fusion of gas chromatography coupled with ion mobility spectrometry (GC-IMS), flash gas chromatography electronic nose (FGC-Enose) and multimodal high performance liquid chromatography-high resolution mass spectrometry (HPLC-HRMS) datasets.

SAMPLE	ACTUAL	PREDICTED	AVAREGED PROBABILITY
CP_30	AUTHENTIC	EVOO	0.985
CP_31	AUTHENTIC	EVOO	0.9778
CP_32	AUTHENTIC	EVOO	0.96445
DEO3	ADULTERATED	OTHER	0.99454
DEO_DEA_2	ADULTERATED	OTHER	0.98428
MIX_D	ADULTERATED	OTHER	0.74301

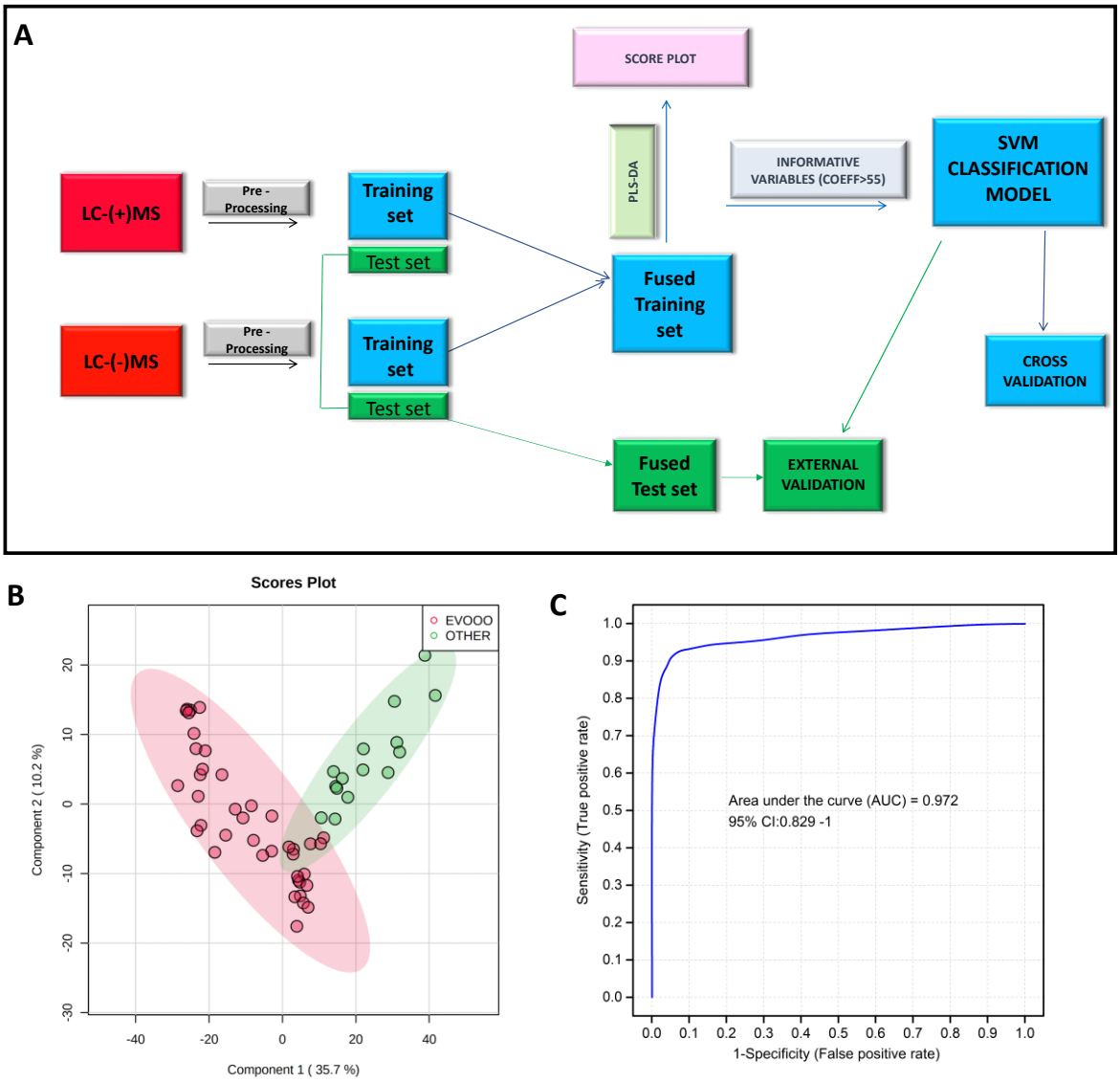


Figure 1. Flow chart of the low-level data fusion and multivariate statistical analysis of multimodality high pressure liquid chromatography-high resolution mass spectrometry (LC-(+/-) MS) datasets. *A) The flow chart showing the combination of LC(+/-)MS datasets after low-level data fusion. B) PLS-DA score plot that allowed visualization of the discrimination of the two groups in the study. C) The prediction power of the SVM model was estimated based on the area under the curve (AUC) of the receiver operating characteristic (ROC) curve.*

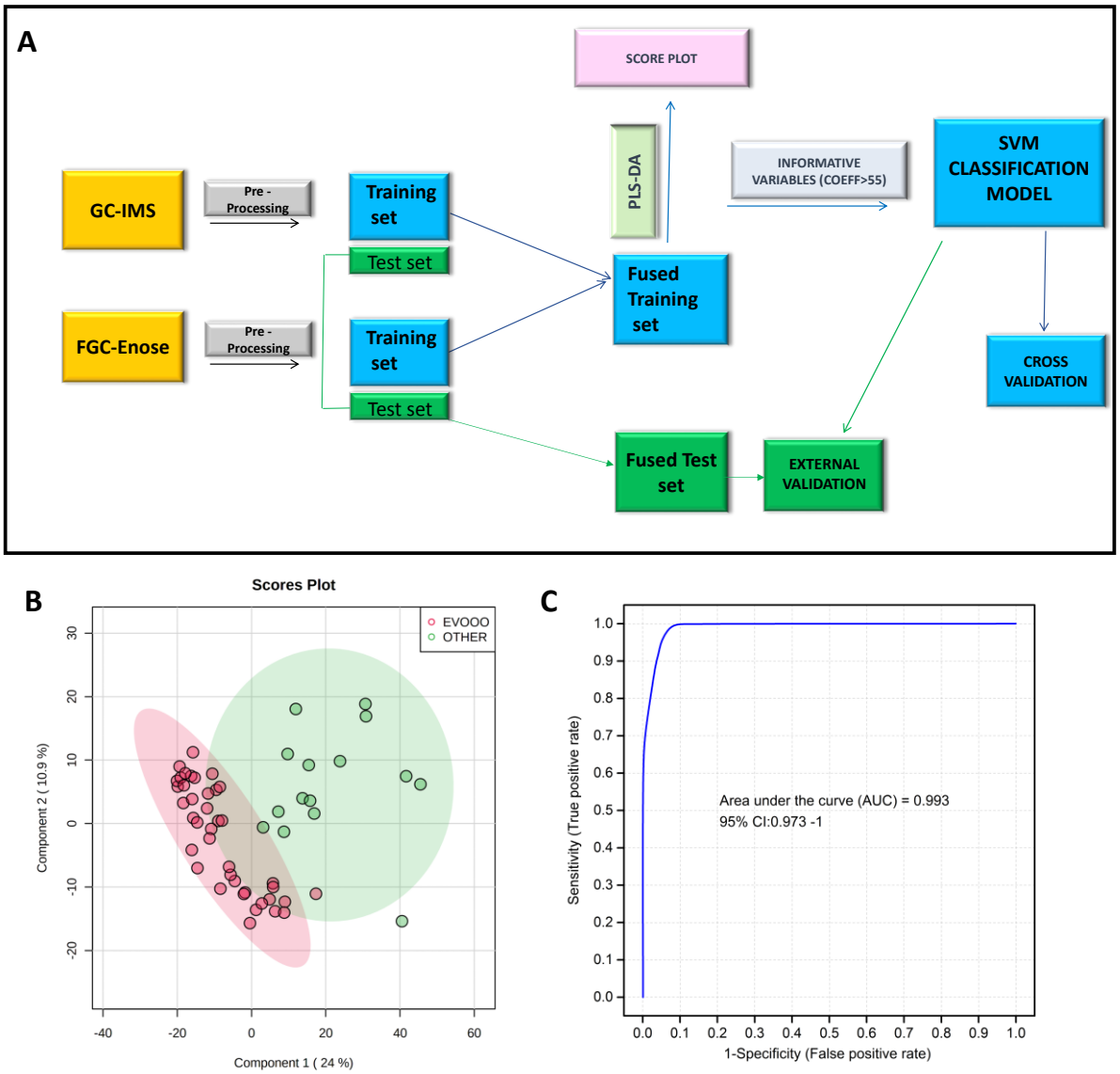


Figure 2. Flow chart of the low-level data fusion and multivariate statistical analysis of gas-chromatography ion mobility spectrometry (GC-IMS) and flash gas-chromatography electronic nose (FGC-Enose) datasets. *A) The flow chart showing the combination of GC-IMS and FGC-Enose datasets after low-level data fusion. B) PLS-DA score plot that allowed visualization of the discrimination of the two groups in the study. C) The prediction power of the SVM model was estimated based on the area under the curve (AUC) of the receiver operating characteristic (ROC) curve.*

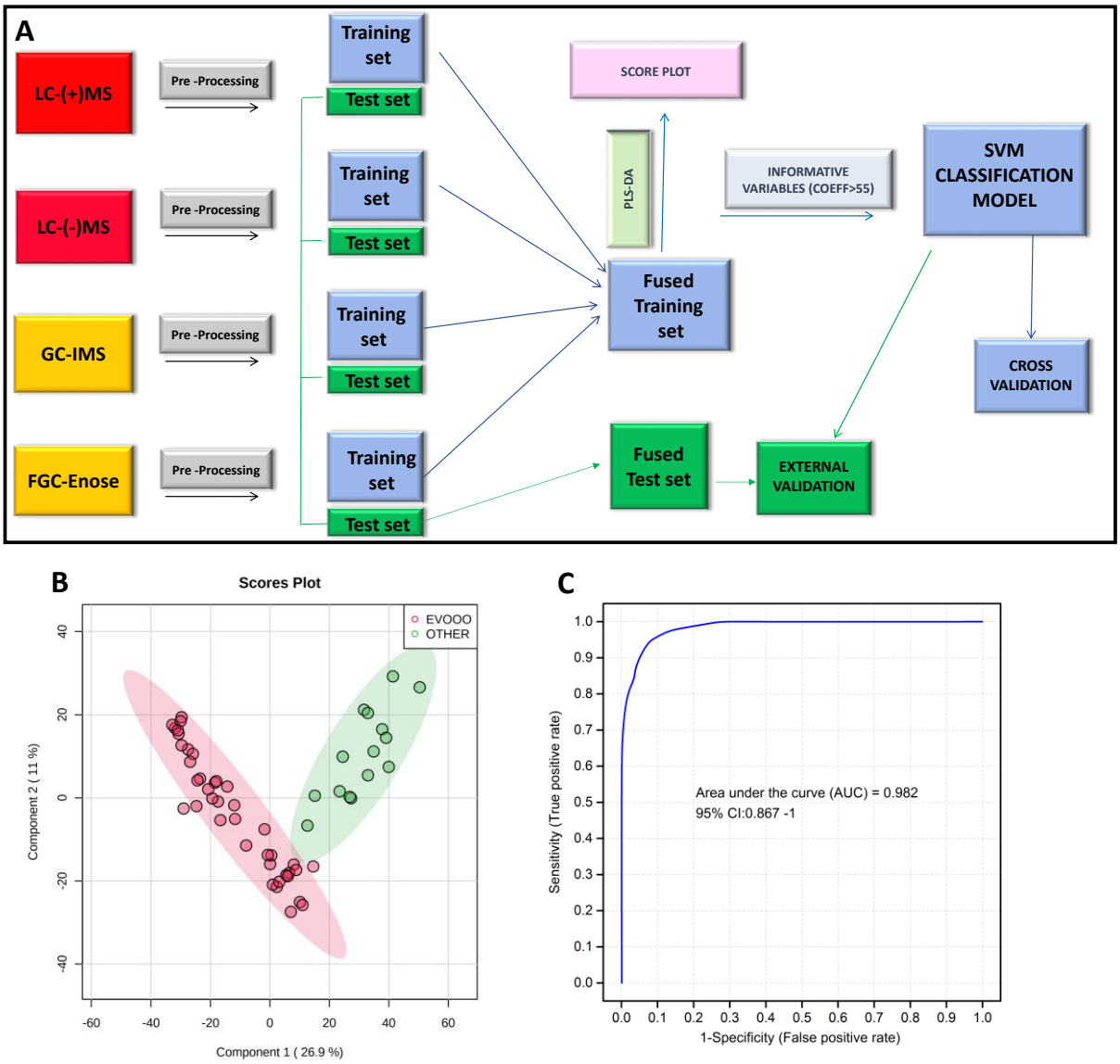


Figure 3. Flow chart of the low-level data fusion and multivariate statistical analysis of multimodality high pressure liquid chromatography-high resolution mass spectrometry (LC-(+/-) MS), gas-chromatography ion mobility spectrometry (GC-IMS) and flash gas-chromatography electronic nose (FGC-Enose) datasets*. A) The flow chart showing the combination of the four datasets after low-level data fusion. B) PLS-DA score plot that allowed visualization of the discrimination of the two groups in the study. C) The prediction power of the SVM model was estimated based on the area under the curve (AUC) of the receiver operating characteristic (ROC) curve.*

Table 1. *Statistical figures of merit of Support Vector Machine (SVM) models obtained in cross-validation on training set (sensitivity, specificity and accuracy) after combining the three analytical approaches by low-level data fusion. Number of samples correctly classified and probability of predictions during in validation on test set are also reported.*

Merged technique	Sensitivity on training set	Specificity on training set	Accuracy on training set	AUC of the ROC	Samples correctly classified in validation of the test set	Average probability of the predictions on test set
LC-(+/-)MS	0.93	0.95	0.94	0.97	6/6	0.96
GC-IMS FGC-Enose	0.93	0.97	0.96	0.99	6/6	0.93
GC-IMS FGC-Enose LC-(+/-)MS	0.93	0.96	0.96	0.98	6/6	0.94