Analytical Letters

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/lanl20

# Classification of Cereal Flour by Gas Chromatography – Mass Spectrometry (GC-MS) Liposoluble Fingerprints and Automated Machine Learning

Kristian Pastor, Marko Ilić, Jovana Kojić, Marijana Ačanski & Djura Vujić

**To cite this article:** Kristian Pastor, Marko Ilić, Jovana Kojić, Marijana Ačanski & Djura Vujić (2022): Classification of Cereal Flour by Gas Chromatography – Mass Spectrometry (GC-MS) Liposoluble Fingerprints and Automated Machine Learning, Analytical Letters, DOI: 10.1080/00032719.2022.2050921

**To link to this article:** https://doi.org/10.1080/00032719.2022.2050921

Published online: 21 Mar 2022.

ANALYTICAL LETTERS
https://doi.org/10.1080/00032719.2022.2050921

FOOD ANALYSIS

# Classification of Cereal Flour by Gas Chromatography – Mass Spectrometry (GC-MS) Liposoluble Fingerprints and Automated Machine Learning

Kristian Pastor[a], Marko Ilić[a], Jovana Kojić[b], Marijana Ačanski[a], and Djura Vujić[c]

[a]Faculty of Technology, University of Novi Sad, Novi Sad, Serbia; [b]Institute of Food Technology Novi Sad (FINS), University of Novi Sad, Novi Sad, Serbia; [c]Independent scientist, Novi Sad, Serbia

**ABSTRACT**

An innovative and rapid approach is described for classifying common types of gluten and non-gluten cereal flour (wheat, rye, triticale, barley, oats, and corn) into the groups defined by their botanical origin. Liposoluble compounds were extracted from flour samples, derivatized, and analyzed using gas chromatography – mass spectrometry (GC-MS). Raw signals used for data processing consisted of mass spectra scans of full chromatograms. These represented unique fingerprints for each class. An automated machine learning framework was applied for classification. The algorithm automatically explored each of the 39 classifiers provided by the software. Using 10-fold cross-validation, a simple logistic classifier was recommended to be optimal. The constructed model resulted in 85.71% correctly classification according to the botanical origin. Furthermore, it unequivocally discriminated samples of non-gluten corn flour. This non-targeted strategy supports the use of artificial intelligence in developing methods for flour authentication.

**ARTICLE HISTORY**
Received 7 February 2022
Accepted 4 March 2022

**KEYWORDS**
Automated machine learning; cereal flour; gas chromatography – mass spectrometry (GC-MS)

## Introduction

Food authentication and fraud detection still remain challenging issues worldwide (Böhme et al. 2019). In general, all food products are susceptible to fraud and adulteration, including flour and bakery products (Pastor, Ačanski, and Vujić 2019).

Cereal grains and cereal flours represent crucial staple foods in human daily nutrition. They are being incorporated into various bakery products, such as bread, cakes, biscuits, and many others (Pastor et al. 2018). There is an obvious trend to include other agricultural crops, other than wheat, into bakery products, including barley, rye, oats, and corn. In addition, non-wheat flour has been widely used in development of a wide range of non-gluten products, which are demanded by increasing incidence of celiac disease (Békés, Schoenlechner, and Tömösközi 2017). Consequently, there are two essential aspects for developing authentication protocols of cereal-based products. Economical concern is the first due to the high prices for alternative grains and consumer health is the second. The consumption of products containing undeclared

**CONTACT** Kristian Pastor kristian.pastor@uns.ac.rs Faculty of Technology, University of Novi Sad, Bulevar cara Lazara 1, 21000 Novi Sad, Serbia.

Table 1. Analyzed cereal grains, cultivars, corn hybrids, and their abbreviations.

| Number of samples | Cereal | Abbreviation | Cultivar – hybrid name |
|---|---|---|---|
| 19 | **Corn** *Zea mays* L. | C | NS-5, NS-7, NS-9, NS-10, NS-11, NS-12, NS-13, NS-14, NS-17, NS-18, NS-19, NS-20, NS-21, NS-22, NS-23, NS-24, NS-26, NS-27, NS-28 |
| 9 | **Wheat** *Triticum aestivum* L. | W | Renesansa, Rapsodija, Evropa 90, Pesma, Milijana, Nataša, Venera, Durumko, NS Dur |
| 8 | **Barley** *Hordeum vulgare* L | B | Novosadski 525, NS Pinon, NS Zitos, Atlas, Somborac, Rudnik, NS Marko, Golijat |
| 3 | **Oats** *Avena sativa* L. | O | Dunav, Jadar, Sedef |
| 2 | **Triticale** *Triticosecale* Wittm. | T | NS Karnak, NS Trifun |
| 1 | **Rye** *Secale cereale* L. | R | NS Savo |

constituents may cause intoxication or allergy in sensitized individuals, e.g., gluten intolerance (Cubero-Leon, Peñalver, and Maquet 2014; Borras et al. 2015; Manning 2016).

Flour and bakery products are generally not expensive, but they are definitely consumed in large quantities on a global scale. Therefore, a simple and rapid approach for flour authentication on the food market presents a significant challenge. There is a need for developing new methodologies that are more accurate and efficient (Cubero-Leon, Peñalver, and Maquet 2014; Danezis et al. 2016). However, although this trend is being strongly stimulated by consumers, regulatory bodies and the food industry, unfortunately little effort has been put into the analysis of flour and bakery products. Many authors believe that the future of food authentication lies in the synergistic fusion of sophisticated analytical instruments and novel data treatment tools for processing enormous amounts of complex datasets (Borras et al. 2015). The scientific literature provides studies aiming to authenticate flour by applying gas chromatography - mass spectrometry (GC-MS) (Pastor et al. 2016, 2018; Hammann et al. 2019; Pastor et al. 2020; Bodroža-Solarov et al. 2021). These involve the identification step, thereby obtaining qualitative information about lipid composition of analyzed flour samples. This work describes a novel, non-targeted approach, aimed at classifying various types of cereal flours: corn (non-gluten), wheat, rye, triticale, barley, and oats (gluten-containing), into the groups defined by botanical origin. The method employs GC-MS coupled to an automated machine learning algorithm for the rapid and efficient processing of obtained chemical information without the need to identify eluting compounds. According to the authors' knowledge, this work presents the first application of artificial intelligence in flour authentication.

## Materials and methods

Grain samples of various cereal species, Table 1, were collected from the Institute of Field and Vegetable Crops in Novi Sad, Serbia. Grains were ground into flour using a laboratory mill. Liposoluble compounds were extracted with *n*-hexane and derivatized into corresponding volatile compounds using 0.2 M trimethylsulfonium hydroxide (TMSH, Macherey-Nagel, Germany).

Derivatized extracts were analyzed by GC-MS (Agilent Technologies 7890 gas chromatograph with 5975 MS detector, Palo Alto, CA, USA) coupled to the Agilent MSD ChemStation software. The injector temperature was 250 °C. A split ratio of 1:50 was

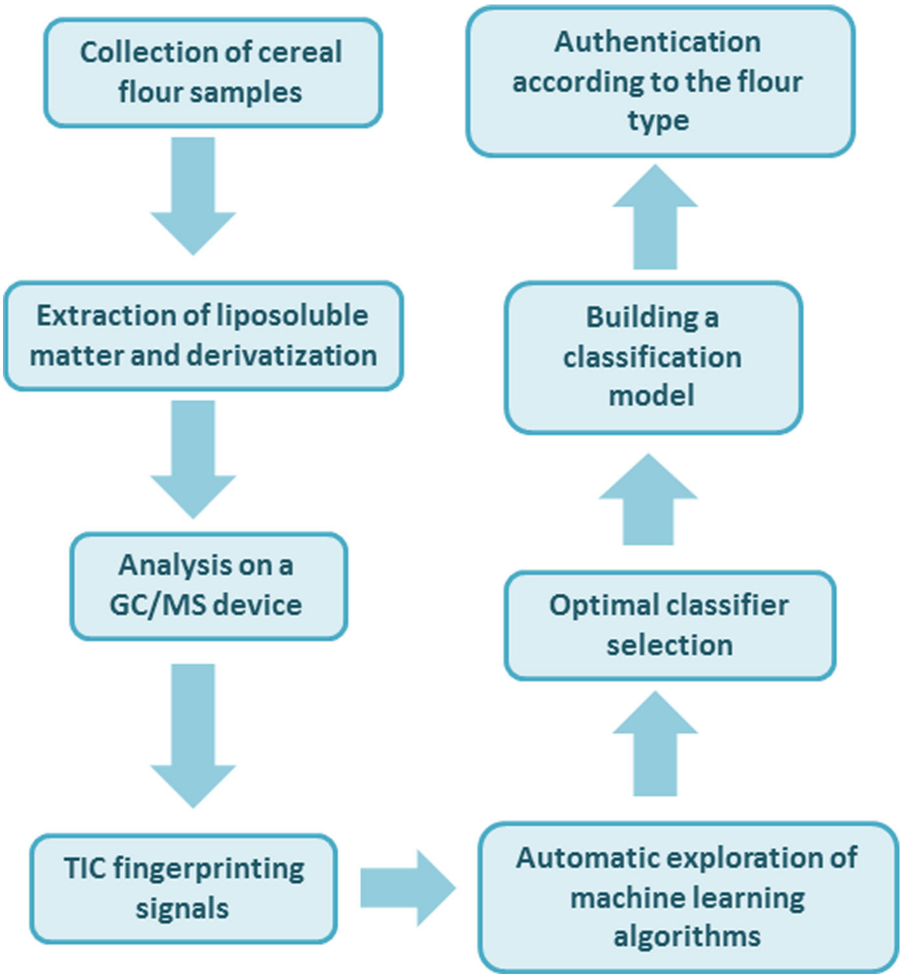


**Figure 1.** Workflow of studying the potential of automated machine learning by GC-MS fingerprinting for rapid flour authentication.

used for injecting 1 $\mu$l of sample. Lipid compounds were separated on a DB-5MS column (30 m $\times$ 0.25 mm $\times$ 25 $\mu$m) with the following temperature program: 50 to 130 °C at 30 °C/min and 130 to 300 °C at 10 °C/min. The flow rate of helium gas was 0.8 mL/min. A quadrupole MSD was utilized in electron ionization mode at 70 eV. The obtained total ion current chromatograms (TIC) consisted of mass spectra scans of various intensities. In raw format, they served as inputs for further data processing.

An automated machine learning algorithm was applied using the open-source Auto-WEKA package (https://www.cs.waikato.ac.nz/ml/weka/), employing a state-of-the-art Bayesian optimization method, thereby solving the combined algorithm selection and hyperparameter optimization (CASH) problem (Hall et al. 2009; Thornton et al. 2013; Frank, Hall, and Witten 2016; Kotthoff, Thornton, and Hutter 2017). The time budgets carried out by the computer unattended were 120, 60, 30, 15, 10 and 5 min, with a memory limit of 1024 MB, batch size 100, and 123 seeds. The methodological workflow scheme of the presented study is shown in Figure 1. Linear regression, such as simple

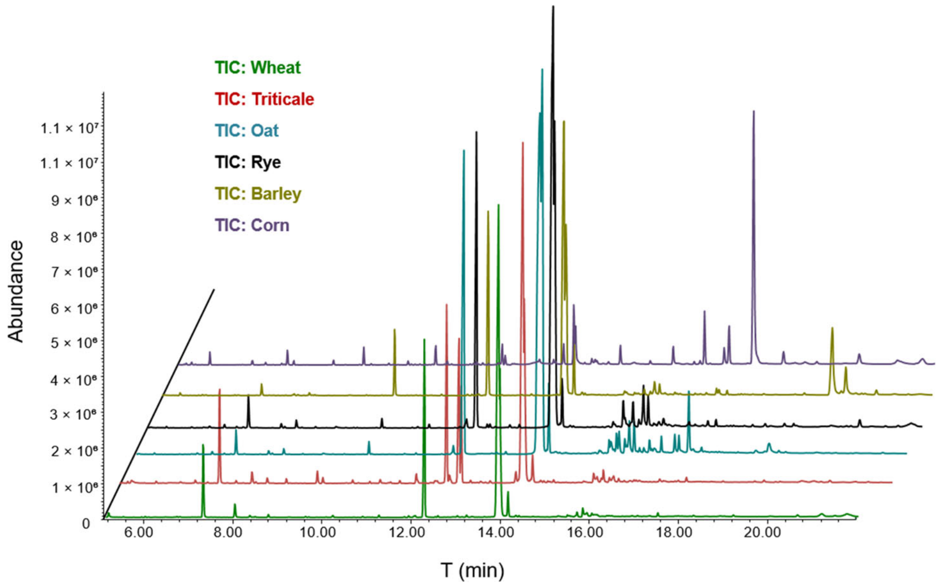


**Figure 2.** Total ion chromatograms of each flour type: wheat, triticale, oat, rye, barley, and corn.

logistic, can easily be used for classification in domains with numeric attributes. Any regression technique can be used for classification by performing a regression for each class, setting the output equal to one for training instances that belong to the class, and zero for those that do not. The result is a linear expression for the class. Next, given a test example of unknown class, the value of each linear expression is calculated and the largest is selected (Witten et al. 2017).

## Results and discussion

Example TIC chromatograms, consisting of 1666 mass spectra scans as data points, representative of each flour type, are shown in Figure 2. The previous work shows that, when hexane is used as the extraction solvent, fatty acids, phytosterols and tocopherols are extracted from the flour (Pastor et al. 2018). On the contrary, the presented study uses raw signals as fingerprinting variables for defining groups of botanical origin of corn, wheat, barley, triticale, oat, and rye flour by omitting the peak identification step. Thus, it is important to note that this approach does not require identification of the eluting compounds for classification purposes. In this case, the information about the specific compounds extracted with hexane is unimportant for applying machine learning algorithm.

Furthermore, the potential of lipid profiling in discriminating flour samples produced from a wide range of cereal and pseudocereal grains has already been demonstrated. For these purposes, the exploratory data analysis tools were applied, which have been frequently used in chemometrics and multivariate analysis of chemical data, including hierarchical cluster analysis and principal component analysis (Pastor et al. 2016, 2018, 2020; Hammann et al. 2019; Bodroža-Solarov et al. 2021). In these cases, the authors used common software packages, such as Statistica 10.0 (StatSoft, Tulsa, OK, USA),

**Table 2.** Confusion matrix obtained by the simple logistic classifier using Auto-WEKA.

| *Corn* | *Wheat* | *Barley* | *Triticale* | *Oats* | *Rye* | Class |
|---|---|---|---|---|---|---|
| **19** | 0 | 0 | 0 | 0 | 0 | *Corn* |
| 0 | **9** | 0 | 0 | 0 | 0 | *Wheat* |
| 0 | **2** | **5** | **1** | 0 | 0 | *Barley* |
| 0 | **1** | 0 | **1** | 0 | 0 | *Triticale* |
| 0 | 0 | **1** | 0 | **2** | 0 | *Oats* |
| 0 | 0 | **1** | 0 | 0 | **0** | *Rye* |

MetaboAnalyst 4.0 R-package (R foundation for statistical computing, Vienna, Austria), and PAST (Natural History Museum, Oslo, Norway). These tools were applied in order to observe the correlations of analyzed samples and variables, not necessarily having prior information about the samples. However, a machine learning approach employs classification algorithms, providing detailed information about classification accuracy, errors and misclassified samples, providing a confusion matrix.

The presented work describes a premiere application of a machine learning approach in an automated mode with the goal of flour authentication. Auto-WEKA is a system automatically searching through the joint space of WEKA's learning algorithms and their respective hyperparameter settings in order to maximize performance using a state-of-the-art Bayesian optimization method (Kotthoff et al. 2016). In this specific case, 120, 60, 30, 15, 10 and 5 minute time-budgets were carried out by the unattended computer in order to select the most appropriate algorithm among the 39 classifiers provided by the software that included 27 base learners, 10 meta-methods and 2 ensemble methods. A simple logistic classifier (under WEKA classifier functions) was recommended to be the most appropriate using a 10-fold cross-validation to exploit the performance gains on a given dataset. The constructed model resulted in 85.71% of correctly classified instances (36 out of 42 flour samples) (Table 2). Each row of the matrix represents the instances in an actual class, while each column represents the instances in a predicted class. As shown by the confusion matrix, the algorithm classified 19 corn, 9 wheat, 5 barley (out of 8), 2 oats (out of 3), and 1 triticale (out of 2) correctly. Furthermore, all non-gluten corn flour samples were unambiguously separated from gluten-containing small grain flour: wheat, barley, oat, triticale and rye. However, 2 barley samples and 1 triticale sample were misclassified as wheat, 1 oat and 1 rye sample as barley, and 1 barley sample as oats. Misclassified gluten-containing flour samples are in red in Table 2. The classification performance is quite reasonable, taking into account botanical similarity of small grains, which contributes to a similar composition. The mean absolute error obtained by the model was 0.0574 and the root mean squared error 0.2083. The error visualization of the recommended classifier is presented in Figure 3.

## Conclusions

The presented approach directly supports the application of artificial intelligence onto chemical information, aiming to develop innovative methods for food authentication. This rapid and non-targeted strategy successfully omitted time-consuming quantifications and occasionally non-reliable identification of chemical compounds from

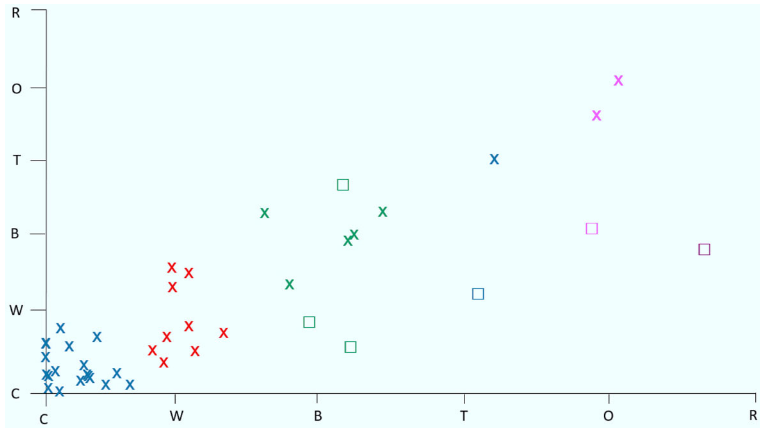


Figure 3. Visualization of classification errors obtained by the recommended model. Class colors: corn (C) – blue; wheat (W) – red; barley (B) – green; triticale (T) – blue; oat (O) – rose; rye (R) – pink. The *x*-symbols represent correctly classified instances and the squares represent misclassifications.

chromatographic systems. Furthermore, an applied graphical user interface software did not require extensive coding knowledge and in-depth mathematical background. A simple logistic classifier, suggested by the system as the optimal for analyzed datasets provided a classification accuracy of 85.71%. According to the authors' knowledge, no previous approach of this type has been employed to classify or authenticate cereal flour.

## Acknowledgement

Dr. Kristian Pastor also acknowledge the support from COST Action CA18101 – Sourdomics, and CA19145 - SensorFINT.

## Declaration of interest statement

The authors declare that they have no known competing financial interests or personal relationships that influenced the work reported in this paper.

## Funding

This study was funded by the Ministry of Education, Science and Technological Development of the Republic of Serbia (Program number 451-03-9/2021-14/200134 and 451-03-9/2021-14/200222).

## References

Békés, F., R. Schoenlechner, and S. Tömösközi. 2017. Chapter 14 - Ancient wheats and pseudocereals for possible use in cereal-grain dietary intolerances. In *Cereal grains - assessing and managing quality*, ed. C. Wrigley, I. Batey, and D. Miskelly, 2nd ed., 353–89. Cambridge, UK: Woodhead Publishing Series in Food Science, Technology and Nutrition. doi:10.1016/B978-0-08-100719-8.00028-0.

Bodroža-Solarov, M., S. Grobelnik-Mlakar, L. Pezo, S. Keleman, S. Ilin, B. Marić, and B. Filipčev. 2021. Identification of biomarkers in hydrosoluble extracts from spelt and wheat cultivated in different production systems. *Journal of the Science of Food and Agriculture* 101 (8):3413–21. doi:10.1002/jsfa.10971.

Böhme, K., P. Calo-Mata, J. Barros-Velázquez, and I. Ortea. 2019. Recent applications of omics-based technologies to main topics in food authentication. *Trac Trends in Analytical Chemistry* 110:221–32. doi:10.1016/j.trac.2018.11.005.

Borras, E., J. Ferre, R. Boque, M. Mestres, L. Acena, and O. Busto. 2015. Data fusion methodologies for food and beverage authentication and quality assessment - A review. *Analytica Chimica Acta* 891:1–14. doi:10.1016/j.aca.2015.04.042.

Cubero-Leon, E., R. Peñalver, and A. Maquet. 2014. Review on metabolomics for food authentication. *Food Research International* 60:95–107. doi:10.1016/j.foodres.2013.11.041.

Danezis, G. P., A. S. Tsagkaris, V. Brusic, and C. A. Georgiou. 2016. Food authentication: State of the art and prospects. *Current Opinion in Food Science* 10:22–31. doi:10.1016/j.cofs.2016.07.003.

Frank, E., M. A. Hall, and I. H. Witten. 2016. The WEKA workbench. Online appendix for data mining: Practical machine learning tools and techniques. 4th ed. Burlington, VT: Morgan Kaufmann.

Hall, M., E. Frank, G. Holmes, B. Pfahringer, B. Reutemann, and I. H. Witten. 2009. The WEKA data mining software: An update. *ACM SIGKDD Explorations Newsletter* 11 (1):10–8. doi:10.1145/1656274.1656278.

Hammann, S., A. Korf, I. D. Bull, H. Hayen, and L. J. E. Cramp. 2019. Lipid profiling and analytical discrimination of seven cereals using high temperature gas chromatography coupled to high resolution quadrupole time-of-flight mass spectrometry. *Food Chemistry* 282:27–35. doi:10.1016/j.foodchem.2018.12.109.

Kotthoff, L., C. Thornton, H. Hoos, F. Hutter, and K. Leyton-Brown. 2016. Auto-WEKA 2.0: Automatic model selection and hyperparameter optimization in WEKA. *Journal of Machine Learning Research* 18 (25):1–5. https://jmlr.org/papers/volume18/16-261/16-261.pdf.

Kotthoff, L., C. Thornton, and F. Hutter. 2017. User guide for Auto-WEKA version 2.6. https://www.cs.ubc.ca/labs/beta/Projects/autoweka/manual.pdf

Manning, L. 2016. Food fraud: Policy and food chain. *Current Opinion in Food Science* 10:16–21. doi:10.1016/j.cofs.2016.07.001.

Pastor, K., M. Ačanski, Đ. Vujić, G. Bekavac, S. Milovac, and S. Kravić. 2016. Rapid method for small grain and corn flour authentication using GC/EI-MS and multivariate analysis. *Food Analytical Methods* 9 (2):443–50. doi:10.1007/s12161-015-0215-6.

Pastor, K., M. Ilić, Đ. Vujić, Đ. Jovanović, and M. Ačanski. 2020. Characterization of fatty acids in cereals and oilseeds from the Republic of Serbia by gas chromatography – mass spectrometry (GC/MS) with Chemometrics. *Analytical Letters* 53 (8):1177–89. doi:10.1080/00032719.2019.1700270.

Pastor, K., L. Pezo, D. Vujic, D. Jovanovic, and M. Acanski. 2018. Discriminating cereal and pseudocereal speciesusing binary system of GC/MS data – Pattern recognition approach. *Journal of the Serbian Chemical Society* 83 (3):317–29. doi:10.2298/JSC170926014P.

Pastor, K., M. Ačanski, and Đ. Vujić. 2019. Chapter 3: A review of adulteration versus authentication of flour. In *Flour and breads and their fortification in health and disease prevention*, ed. V. R. Preedy, and R. R. Watson, 2nd ed., 21–36. London, UK: Academic Press, Elsevier Inc. doi:10.1016/B978-0-12-814639-2.00003-4.

Thornton, C., F. Hutter, H. Hoos, and K. Leyton-Brown. 2013. Auto-WEKA: Combined selection and hyperparameter optimization of classification algorithms. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, University of British Columbia, Vancouver, Canada. https://www.cs.ubc.ca/labs/beta/Projects/autoweka/papers/autoweka.pdf. doi:10.1145/2487575.2487629.

Witten, I. H., E. Frank, M. A. Hall, and C. J. Pal. 2017. *Data mining: Practical machine learning tools and techniques*. 4th ed. Cambridge, MA: Elsevier.