



ALLIANCE

A hoListic framework in the quality Labelled
food supply chain systems' management
towards enhanced data Integrity and verAcity,
interoperability, traNsparenCy, and tracEability



TOOLS & DIGITAL KNOWLEDGE DATABASE FOR DETECTING FOOD FRAUD USING NOVEL PORTABLE RAPID TESTING FOR ON-SITE INSPECTION

GRANT AGREEMENT NUMBER: 101070141



This project has received funding from the European Union's HE research and innovation programme under grant agreement No 101084188



Lead Beneficiary: Netcompany - Intrasoft [INTRA]

Type of Deliverable: R — Document, report

Dissemination Level: Public

Submission Date: 30.04.2024 (Month 18)

Version: 0.6

Versioning and contribution history

Version	Description	Contributions
0.1	Preparation of Table of Contents, Content Responsibility Assignments	INTRA
0.2	Contributions from partners to sections 2, 3, 4, 5 and 6 (1st round).	BIOC, ASINCAR, FINS, INTRA, and UNIBO
0.3	Contributions from partners to sections 2, 3, 4, 5 and 6 (2nd round).	BIOC, LGL, ASINCAR, FINS, INTRA, and UNIBO
0.4	Ready for internal review	INTRA
0.5	Comments received by IRs	ASINCAR, and UNIZG
0.6	Comments from IRs addressed, and deliverable sent to the PC	INTRA
1.0	Final QA'ed version and submission to the EC Portal	UTH

Authors

Author	Partner
Amalia Ntemou	Netcompany-Intrasoft (INTRA)
Athanasia-Maria Dourou, Evangelia Lampropoulou, Georgia Kesisoglou, Stylianos Arhondakis	BioCoS (BIOC)
Martín Hervello, Noemí Quintanal, Bárbara Álvarez, Marta Mier, Roberto Morán, Armando Menéndez	Association for Research on Meat Industry of the Principality of Asturias (ASINCAR)
Nikola Maravić, Predrag Ikonić	Institute For Food Technology of Novi Sad (FINS)
Giulia Maesano, Alessandra Castellini, Maurizio Canavari	Alma Mater Studiorum -University of Bologna (UNIBO)
Ingrid Huber, Frederic Schedel, Patrick Gürtler	Bavarian Health and Food Safety Authority (LGL)

Reviewers

Name	Organisation
Pelayo González González	Association for Research on Meat Industry of the Principality of Asturias (ASINCAR)
Marija Cerjak	University of Zagreb - Faculty of Agronomy (UNIZG)





Disclaimer

The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Commission. The Commission does not guarantee the accuracy of the data included in this study. Neither the Commission nor any person acting on the Commission's behalf may be held responsible for the use, which may be made of the information contained therein.



Contents

1	Introduction.....	10
1.1	Document purpose and scope.....	10
1.2	Relationship to project work.....	10
1.3	Document Structure	11
2	Next Generation Portable DNA Sequencing for Food Analysis	12
2.1	Introduction.....	12
2.1.1	Background	12
2.1.2	Principles of DNA-based methods	12
2.1.3	The importance of portability	13
2.1.4	DNA-based Authentication and Traceability.....	13
2.2	Relevance with the EVOO pilot.....	13
2.3	Results.....	14
2.3.1	Experimental design.....	14
2.3.2	Sample collection and analysis	15
2.3.3	Machine Learning for DNA-data classification.....	16
2.3.4	Key Results	16
2.4	Next steps.....	16
3	Enhanced Food Fraud Detection with Advanced Spectroscopy.....	18
3.1	Introduction.....	18
3.1.1	Portable NIR for food applications	21
3.1.2	HSI for food applications.....	22
3.1.3	Portable NIR and HSI for food authenticity	23
3.1.4	Methodology for the model development	24
3.2	Overall description of the use case	24
3.3	Achieved results	26
3.3.1	Experimental design.....	26
3.3.2	Data collection.....	30
3.3.3	Intelligent data processing.....	32
3.3.4	Results visualization.....	38
3.4	Conclusions.....	40
3.5	Next steps.....	40
4	Digital Knowledge Base for Food Fraud Mitigation	42
4.1	Introduction.....	42
4.2	Background	42
4.3	Digital Knowledge Base Overview	42



4.4	Development of Database	43
4.5	Future Directions and Improvements	44
5	Food Fraud Prevention with Predictive Analytics	45
5.1	Introduction.....	45
5.2	Conceptual Architecture of the Predictive Analytics Module	45
5.2.1	Data Acquisition and Ingestion	46
5.2.2	Data Processing	46
5.2.3	Data Mining	47
5.2.4	Machine Learning Model Development	47
5.2.5	Model Deployment and Inference.....	47
5.2.6	Predictive Analytics	47
5.2.7	Monitoring and Explainability.....	47
5.3	Data Value Chain	48
5.3.1	Data Collection and Integration	48
5.3.2	Data Cleaning and Preprocessing	48
5.3.3	Feature Engineering	49
5.3.4	Model Training Data Preparation	50
5.3.5	Model Evaluation	50
5.3.6	Model Deployment	50
5.3.7	Continuous Improvement	50
5.4	Food Fraud Detection Modelling.....	52
5.4.1	Data Acquisition and Cleansing.....	52
5.4.2	Model Training in the Champion-Challenger Framework	52
5.4.3	Deployment and Explainability	53
5.5	Visualisations and predictive Analytics	54
5.6	Concepts towards Explainable AI and rational decision-making	56
5.7	Next Steps	58
6	Consumer Demand Assessment and Strengthening.....	60
6.1	Introduction.....	60
6.2	Literature review	61
6.3	Theoretical Framework	62
6.4	Method.....	63
6.4.1	Sample Design and Data Collection	63
6.4.2	Questionnaire	63
6.4.3	Data Analysis.....	64
6.5	Assessment of consumer perception and behaviour	65





6.6	Conclusion and policy implication.....	65
7	Conclusions and Future Directions.....	67
7.1	Recap of Achievements.....	67
7.2	Addressing Project Objectives.....	67
7.3	Future Directions	67
8	Annex.....	68



List of figures

Figure 1 Spatial distribution of olive orchards in the area of Umbria, Italy.....	15
Figure 2 Olive leaf samples collection form Template.	15
Figure 3 Image acquisition modes: (a) reflectance mode, (b) transmittance mode and (c) transfectance mode	18
Figure 4 Representative image of a hyperspectral cube.....	20
Figure 5 Methodological approach scheme.	26
Figure 6 Map of the municipalities in the Principality of Asturias.	27
Figure 7 Measuring formats for beans.	27
Figure 8 Caliper and colorimeter used to obtain morphological and colour data of faba beans in all formats.	28
Figure 9 NIR measurements with a) portable NIR-S-G1 InnoSpectra equipment measuring whole beans and b) desktop NIR ASD LabSpec 4 equipment measuring milled beans.	28
Figure 10 Hyperspectral imaging configuration for longitudinal cut beans (same as for whole beans) and milled beans.	29
Figure 11 Protocol measurement NIR and HSI schema.	29
Figure 12 Protocol measurement physico-chemical and spectra schema.	30
Figure 13 Example for obtaining an average spectrum of each faba bean from the HSI image.	31
Figure 14 Sample data showcasing the length and width measurements extracted from an image captured using the FX10 camera.	32
Figure 15 Box and whisker diagrams (left) and violin diagrams (right) for colour (parameters L*a*b). L* in the top, a* middle, b* down.	33
Figure 16 Box and whisker diagrams (left) and violin diagrams (right) for water absorption. ..	33
Figure 17 Spectra obtained with the NIR-S-G1 portable device for whole beans.....	34
Figure 18 Accuracy, precision, recall and confusion matrix obtained for the models a) Random Forest, b) XGBoost.	36
Figure 19 Accuracy, precision, recall and confusion matrix obtained for SVM model for cut beans.	36
Figure 20 Spectra obtained by HSI with the FX17 camera for a sample of whole beans.	37
Figure 21 Accuracy, precision, recall and confusion matrix obtained for PLS-DA+XGB model for whole beans with FX17 camera.	38
Figure 22 Example of a results display window for end-consumers.	39
Figure 23 Internal functioning diagram of the results visualisation programme.....	39
Figure 24 Visual overview: key components and features of a database.	43
Figure 25 Digital knowledge base architecture.....	44
Figure 26 Conceptual architecture of the food fraud prevention system with predictive analytics.	46
Figure 27 Exploratory data analysis for an illustrative example on feta cheese use case.	49
Figure 28 Managing the data-driven solution lifecycle through MLflow.	51
Figure 29 Training pipeline with respect to multiple performance measures.....	54
Figure 30 SHAP values for an illustrative example on feta cheese use case.....	57
Figure 31 LIME approach for an illustrative example on feta cheese use case.....	58
Figure 32 Probabilities (predictions) distributions for an illustrative example on feta cheese use case.	58



List of tables

Table 1 Number of Distinct Olive Orchards per variety	14
Table 2 Class Description	16
Table 3 Performance metrics	16
Table 4 Physico-chemical parameters collected to date by classical laboratory techniques..	30
Table 5 Information gathered to date by spectroscopy techniques.	30

List of Abbreviations

Abbreviation	Description
FSC	Food Supply Chain
EVOO	Extra Virgin Olive Oil
qPCR	Quantitative Polymerase Chain Reaction
PCR	Polymerase Chain Reaction
HRM	High Resolution Melting Analysis
ML	Machine Learning
PDO	Protected Designation of Origin
PGI	Protected Geographical Indication
AI	Artificial Intelligence
Ismea	Institute of Services for the Agricultural Food Market
WP	Work Package
NIR	Near-infrared
HSI	Hyperspectral imaging
BFTS	Blockchain Food Traceability System
TAM	Technology Acceptance Model
TPB	Theory of Planned Behaviour
SHAP	SHapley Additive exPlanations
LIME	Local Interpretable Model-agnostic Explanations
AUC	Area Under Curve
EDA	Exploratory Data Analysis
RNN	Recurrent Neural Network



Executive Summary

The deliverable at hand presents a comprehensive exploration of innovative technologies and methodologies aimed at combating food fraud and enhancing food safety and quality assurance. Across its various sections, the document delves into cutting-edge approaches, tools, and techniques implemented within the framework of the ALLIANCE EU project. From next-generation portable DNA sequencing for food analysis to enhanced food fraud detection with advanced spectroscopy, and from digital knowledge bases and predictive analytics for food fraud mitigation to consumer demand assessment and strengthening, each section offers valuable insights and solutions, along with preliminary results, to address the multifaceted challenges posed by food fraud and adulteration.

Following the introductory section, subsequent sections delve into distinct technological approaches and solutions. The utilization of next-generation portable DNA sequencing is explored to identify and verify olive oil varieties, offering a robust tool for detecting fraud in premium olive oils. Advanced spectroscopy technologies are then harnessed to develop low-cost, portable devices capable of detecting fraudulent practices in specific food products, such as the PGI Asturian Faba bean. Additionally, a Digital Knowledge Base for Food Fraud is introduced, providing data-driven tools and analysis to regulatory bodies and stakeholders. The application of predictive analytics in food fraud prevention is discussed, emphasizing the integration of deep neural networks and explainable AI.

Furthermore, addressing the growing consumer demand for transparent and traceable food supply chains, the subsequent section underscores the pivotal role of blockchain technology in enhancing traceability and promoting consumer confidence. By monitoring consumer opinion and perceptions, this section provides valuable insights for policymakers and producers in the agri-food sector, guiding the development of strategies to promote safety, quality, transparency, and sustainability. Concluding the deliverable, a recap of achievements aligned with the project's objectives is provided, alongside plans for future activities.

Overall, this deliverable serves as a comprehensive resource for stakeholders across the food industry, offering innovative solutions and strategies to combat food fraud, enhance food safety, and meet consumer demand for transparent and trustworthy food products.

1 INTRODUCTION

1.1 Document purpose and scope

The goal of the Horizon Europe ALLIANCE project is to provide a holistic framework that safeguards data integrity and veracity, enhances traceability and transparency, and reinforces interoperability in quality labelled food supply chain through innovative technology solutions and validated approaches that fosters evidence-based decision making.

Among its objectives are the following: (a) to provide novel rapid and portable test technologies for identifying authenticity and detecting fraud on-site; (b) to create a digital knowledge base; (c) to apply novel Artificial Intelligence (AI) and Machine Learning (ML) techniques to prevent food fraud; and (d) to use portable devices for on-site rapid testing for the identification of adulteration and counterfeit in quality-labelled food products.

The progress that has been made so far (Month 18) in attaining the afore-mentioned objectives as well as preliminary results from each activity are comprehensively documented in the current deliverable, named D3.2 "Interim AI-enabled tools & Digital Knowledge Database for Detecting Food Fraud using novel portable rapid testing for on-site inspection".

1.2 Relationship to project work

The objectives listed in Sec. 1.1 which are the subject of the present deliverable are implemented in Work Package 3 (WP3). More specifically, this deliverable represents the first iteration and outcome concerning the first active period of all WP3 tasks except T3.1 (which has been included in D2.1).

The DNA-based olive oil authentication method used to verify its authenticity and quality along with its advantages as a portable technique and preliminary results are detailed as an outcome of T3.2 led by BIOC. The work that is being performed in T3.3 by ASCINCAR is also thoroughly presented in this deliverable. The characteristics of portable NIR and HSI technologies as revolutionary tools to detect the main fraudulent practices reported for the PGI Asturian faba bean as well as the methodologies for ML models' development and associated results are also documented. D2.3 also provides an initial description of the main modules and features of the ALLIANCE digital knowledge base, which will incorporate functionalities and results derived from the Vulnerability Risk Assessment Framework (T2.3), and the Early Warning and Decision Support System (T2.4) and will interact with the ALLIANCE marketplace (T5.4). Furthermore, the food fraud prevention system with predictive analytics is part of this deliverable. From the conceptual architecture and modelling to visualisations and x-AI concept and results, detailed as output of the ongoing work within T3.5. Last but not least, a literature review, methods and analysis of the assessment of consumer perception and behaviour towards new solutions that is performed in T3.6 are also provided.

Most of the said tasks accompanied by their solutions are related to the pilot demonstrations carried out within WP4. Moreover, it is worth mentioning that solutions/tools presented here (i.e., next-generation portable DNA sequencing for food analysis, enhanced food fraud detection with advances spectroscopy, digital knowledge base, and food prevention with



predictive analytics) are main modules of the ALLIANCE platform, mentioned in the architecture prepared in the context of the D2.3.

D3.2 follows the deliverable D2.1 "Food Fraud Landscape, Strategic Gap Analysis, User Needs and Requirements", which serves as the outcome of T2.1 and T3.1. It is also followed by D3.3 "Final AI-enabled tools & Digital Knowledge Database for Detecting Food Fraud using novel portable rapid testing for on-site inspection", which will be submitted in M30 with the aim to report the final versions and results of the ALLIANCE novel tools that use advanced portable qPCR DNA sequencing, NIR and HSI Spectroscopy, leverage AI and predictive analytics to detect adulteration.

1.3 Document Structure

The document is structured as follows: Executive summary provides summary of the whole document. Section 1 introduces the main scope, and structure of this deliverable as well as its relation to the project work. Section 2 presents the next generation portable DNA sequencing for food analysis. Section 3 introduces the methodologies used for enhanced food fraud detection with advanced spectroscopy. The ALLIANCE Digital Knowledge Base for food fraud mitigation is described in Section 4 while section 5 documents food fraud prevention with predictive analytics. Section 6 report activities related to consumer demand assessment and strengthening. Lastly, section 7 serves as the final and concluding section of the document.



2 NEXT GENERATION PORTABLE DNA SEQUENCING FOR FOOD ANALYSIS

2.1 Introduction

2.1.1 Background

Fraudulent practices and adulteration within the olive oil sector remain persistent challenges for regulatory bodies and food safety authorities with very high impact. The newly developed qPCR High-Resolution Melting Analysis (HRM) fingerprinting tool will allow reliable identification of selected Italian and Greek varieties and provide a new tool for verification of the declaration of selected mono varietal olive oils. This has the potential of detecting food fraud in high price olive oils from Italy and Greece. From an operational standpoint for food safety authorities the development of additional markers for the detection of adulteration of olive oils with other vegetable oils (as already in development by BioCoS) will also be very useful as large fraction of routine samples concern the testing of authenticity and integrity of olive oil products for adulteration and mislabelling. For public health authorities it is important to uphold the highest standards of food safety and regulatory compliance enabled by DNA- based authentication and traceability.

2.1.2 Principles of DNA-based methods

While there are several methods based on analytical chemistry that can identify adulterants to a certain extent and with certain limitations, there are only few methods, such as isotopes, that are able to detect different origination of olive oil.¹ Herein, we employ the powerful genetic information of the olive genome, and we couple a novel approach for the olive oil industry, yet very common in other raw material identification, the real-time PCR coupled with the HRM)² Real-time PCR, also known as quantitative PCR (qPCR), is a molecular biology technique used to amplify and quantify specific DNA sequences in real-time during the PCR process.³ Unlike traditional PCR, where DNA amplification is measured at the end of the reaction, real-time PCR allows for the continuous monitoring of DNA amplification as it occurs. This is achieved by incorporating fluorescent dyes or probes into the PCR reaction, which emit fluorescence as DNA is amplified.⁴ The amount of fluorescence generated correlates with the amount of DNA present in the sample, enabling quantitative analysis. Moreover, HRM is a molecular method used to detect sequence variations in nucleic acids with high sensitivity and speed.⁵ It involves PCR amplification of the target DNA region followed by a gradual increase in temperature to induce DNA melting. Fluorescent dyes, like SYBR Green, are utilized to monitor the melting

Hashempour-baltork, F., Zade, S. V., Mazaheri, Y., Alizadeh, A. M., Rastegar, H., Abdian, Z., ... & Damirchi, S. A. (2024). Recent methods in detection of olive oil adulteration: State-of-the-Art. *Journal of Agriculture and Food Research*, 16, 101123.

² Kajan, M. (2021). High resolution melt (HRM) based quantitative real-time polymerase chain reaction (qPCR) primer design tool for point mutation detection (Doctoral dissertation).

³ Raeymaekers, L. (2000). Basic principles of quantitative PCR. *Molecular biotechnology*, 15, 115-122.

⁴ Navarro, E., Serrano-Heras, G., Castaño, M. J., & Solera, J. J. C. A. (2015). Real-time PCR detection chemistry. *Clinica chimica acta*, 439, 231-250.

⁵ Druml, B., & Cichna-Markl, M. (2014). High resolution melting (HRM) analysis of DNA—its role and potential in food analysis. *Food chemistry*, 158, 245-254.





process, with changes in fluorescence indicating DNA denaturation.⁶ HRM distinguishes subtle sequence differences by analysing the shape of the melting curve, which is influenced by variations such as SNPs or mutations.⁷ Specialized software is then employed to interpret melting curve profiles, enabling the detection of genetic variations in samples. HRM offers a cost-effective and efficient approach for genotyping, mutation detection, and sequence analysis across various molecular biology applications.

2.1.3 The importance of portability

A portable DNA-based sensor offers several benefits, particularly in settings where rapid, on-site DNA analysis is crucial. Firstly, portability enables HRM analysis to be conducted directly at the point of sample collection, reducing the time and cost associated with transporting samples to centralized laboratories. Having an automated DNA extractor along the device transforms – even revolutionise – the DNA testing as it was described above. Taking into account that, coupled with the portable device will provide real-time results, enabling immediate decision-making when a potential disruption the EVOO supply chain is about to happen. Overall, the goal through a portable device is to improve efficiency, accessibility, and applicability of molecular analysis in diverse settings of the EVOO supply chain, while keeping the process as simple as possible.

2.1.4 DNA-based Authentication and Traceability

DNA-based authentication of olive oil is a sophisticated method used to verify its authenticity and quality. This technique involves analysing the unique genetic material of olive trees and detecting any potential adulteration by comparing DNA profiles with reference databases. By leveraging PCR and DNA sequencing, this method offers high specificity and sensitivity, enabling the identification of adulterants e.g. other lower quality vegetable oils. DNA-based authentication is increasingly adopted by regulatory bodies and producers to combat fraud and maintain consumer trust in the olive oil industry. On the other hand, DNA-based traceability revolutionizes the olive oil industry by tracing the product's journey from olive groves to consumers through genetic characterization. This innovative approach involves analysing the DNA of olive trees in groves to establish a unique genetic fingerprint. As the olives progress through each stage of the supply chain, from harvesting to processing and packaging, their DNA profiles are tracked and documented. By the time the olive oil reaches consumers, its authenticity and origin can be verified with precision, ensuring transparency and quality assurance. This comprehensive traceability system enhances consumer confidence, facilitates regulatory compliance, and mitigates the risk of fraud, ultimately bolstering the integrity of the olive oil market.

2.2 Relevance with the EVOO pilot

Task 3.2 is closely linked to the ALLIANCE pilot demonstration focusing on PDI/PGO EVOO (Task 4.2). The DNA-based authentication and traceability system aims to rapid detection and

⁶ Zipper, H., Brunner, H., Bernhagen, J., & Vitzthum, F. (2004). Investigations on DNA intercalation and surface binding by SYBR Green I, its structure determination and methodological implications. *Nucleic acids research*, 32(12), e103-e103.

⁷ Grazina, L., Costa, J., Amaral, J. S., & Mafra, I. (2021). High-resolution melting analysis as a tool for plant species authentication. *Crop breeding: Genetic improvement methods*, 55-73.



localisation of food fraud in olive oil (OO) and EVOO. Throughout the pilot demonstration, the efforts made in this task will be used to establish an end-to-end closed system for the CIA partners in Umbria. This includes a DNA extraction workflow harmonised between two partners (BioCos and LGL) and a validated qPCR-based detection method with HRM fingerprinting. The DNA data generated in Task 3.2 will be used to train a machine learning/artificial intelligence (ML/AI) algorithm to automate the classification of reported olive varieties. ML/AI post-processing is integrated into all DNA-testing during the pilot, involving relevant stakeholders in the EVOO supply chain, from producers to retailers, without human intervention. Subsequently, processed DNA data are integrated into a DNA blockchain system, ensuring complete DNA traceability from field to store for EVOO. BioCoS focuses on DNA traceability for Italian varieties, particularly those dedicated to organic PDO/PGI EVOO production, such as Moraiolo, Frantoio, Leccino, Dolce d’Agogia, Rajo, and San Felice. Additionally, other Italian and Greek varieties will be included. Task 3.2 involves discovering novel biomarkers for Italian varieties through sequencing and bioinformatics, alongside exhaustive DNA authentication tests to calibrate the ML/AI pipeline for these varieties. Finally, a Blockchain system (DNA Blockchain) will be implemented to incorporate DNA profiles assessed across different supply chain stages.

2.3 Results

2.3.1 Experimental design

Olive leaf samples were collected from 34 different producers in Italy's Umbria region, representing six olive varieties: Leccino, Moraiolo, Frantoio, San Felice, Rajo, and Dolce d’Agogia. The number of distinct orchards (107 in total) for each variety is depicted in Table 1 below, while their spacial distribution is presented in Figure 1.

Table 1 Number of Distinct Olive Orchards per variety

Olive Variety	Number of Orchards
Frantoio	17
Moraiolo	16
Leccino	17
San Felice	17
Dolce d’Agogia	20
Rajo	20

The objective was to distinguish at least four out of the six varieties using proprietary genetic markers and ML DNA-data analysis. Real-time PCR coupled with HRM analysis was conducted on each sample using six distinct primer sets targeting specific regions within the olive genome.



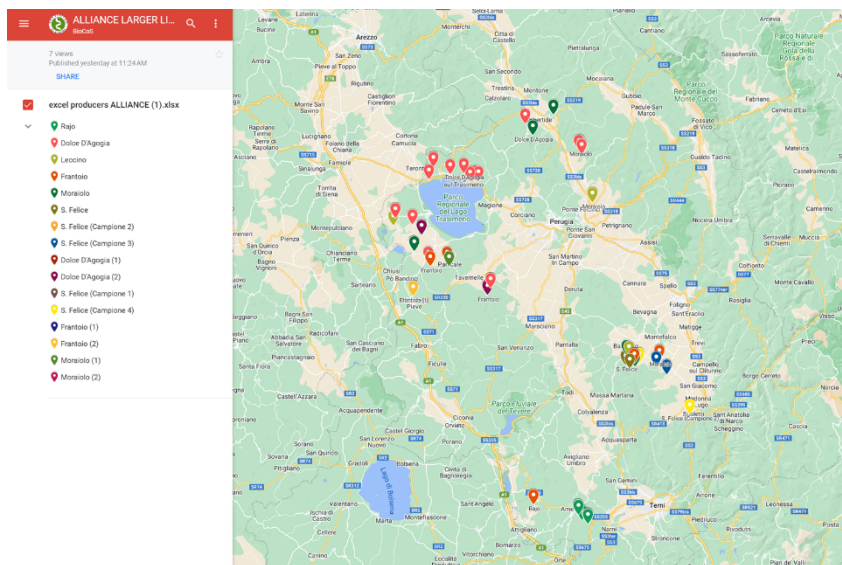


Figure 1 Spatial distribution of olive orchards in the area of Umbria, Italy.

A considerable number of experiments were performed (793 reactions), resulting in a substantial dataset for ML training. The data obtained were processed using bioinformatics tools to design an ML pipeline, ensuring comprehensive coverage of genetic variability within the olive leaf samples. By employing multiple primer sets and experiments, the study aimed to improve the accuracy and reliability of the analysis, enabling detailed characterization

of the genetic profiles of the olive samples.

2.3.2 Sample collection and analysis

Leaf Sample Identification Form


Producer/Company	
Name/Surname	
Company name	
CUAA	
Address/City	
Phone	
E-mail	
Name/Surname of who collected the sample(s)	

LEAF SAMPLE – INFO COLLECTOR

Sample Number					
Collection Date					
Geo-Locus					
Irrigation (Y/N)					
Intensive Cultivation (Y/N)					
Variety					
Pollinators (Y/N)					
Which variety(ies) Pollinators					
Age (tree(s))					
Biological (Y/N)					
Annual Yield (estimation)					
Soil characteristics					
Fertilizers (Y/N)					
Type of Fertilizers (if Yes)					
Altitude (approximative)					
Field Direction (N/S/E/W/NW/NE/SW/SE)					

* BioCoS or other partners of ALLIANCE will not communicate or use personal data without a written consent.
 ** It is recommended to collect leaf samples from different trees of the same field/orchard.
 *** BioCoS will not use personal data of companies or people for commercial purposes without written consent and vice versa.
 **** BioCoS is responsible safety for the provided samples.

Address: OSIAS ERINS Str. 4, Chania 71500, Crete
 Email: info@biocos.it, sales@biocos.it
 Website: www.biocos.it
 Phone: +39 01220 48751



Safety Private and Confidential

Figure 2 Olive leaf samples collection form Template.

At the onset of the T3.2 (M5), a sample collection form was meticulously crafted to gather comprehensive information from olive producers, extending beyond olive grove details. This form, illustrated in Figure 2, aimed to extract relevant insights from producers through extensive discussions. The same form was distributed for the collection of olive oil samples as well (form not shown). The collected samples represented six distinct Italian olive varieties (Leccino, Frantoio, Moraiolo, Dolce Agogia, San Felice, Rajo) sourced from various regions within Umbria, Italy, as depicted in the provided image. Subsequent analysis involved DNA extraction from the collected olive leaf samples, followed by the assessment of proprietary molecular markers. Real-time PCR coupled with HRM was employed to assess the discriminatory capacity of these markers among the six olive varieties mentioned. To uncover potential molecular markers with enhanced discriminatory capabilities, novel primers targeting specific regions within the olive genome were developed. These primers underwent rigorous testing across all six olive varieties using real-time PCR coupled with HRM, enabling a thorough



2.3.3 Machine Learning for DNA-data classification

The objective of the ML approach was to create a model capable of accurately performing varietal classification among the six Italian varieties. To achieve this, we trained an ensemble classifier using features extracted from HRM curves. For each sample, two markers were utilized, and we aimed to train a model capable of classifying each sample into one of three classes. Table 2 illustrates these three classes and their corresponding varieties.

Table 2 Class Description

Class	Variety1	Variety2
1	San Felice	Moraiolo
2	Leccino	Frantoio
3	Dolce d'agogia	Rajo

The performance of the final model on the test dataset is summarized in Table 3. In brief, similar performances were observed in classes 1 and 3, while the metrics for class 2 slightly decreased. Overall, the accuracy of the model was calculated to be 0.8.

Table 3 Performance metrics

	Class: 1	Class: 2	Class: 3
Sensitivity	0.85	0.7	0.85
Specificity	0.95	0.875	0.875
Pos Pred Value	0.8947	0.7368	0.7727
Neg Pred Value	0.9268	0.8537	0.9211
Prevalence	0.3333	0.3333	0.3333
Detection Rate	0.2833	0.2333	0.2833
Detection Prevalence	0.3167	0.3167	0.3667
Balanced Accuracy	0.9	0.7875	0.8625

2.3.4 Key Results

Our experiments led to the discovery of four novel genetic molecular markers, which were tested across the six olive varieties. After extensive validation and testing, two of them were selected for further utilization. Notably, the combination of these markers exhibited significant discriminatory power, revealing the presence of three distinct varietal clusters, as shown in Table 2. These findings could potentially provide valuable insights into the application of DNA fingerprinting technology, reinforcing DNA testing as a reliable source of information, especially for economically significant varieties. This ensures the genetic authenticity of products throughout the entire supply chain, from field to store.

2.4 Next steps

The next steps for the remaining months of Task 3.2 are focused on the analysis of the olive oil samples, along with the continuous training of our ML model to increase further its power. Moreover, we aim to incorporate the algorithm of the ML model into a Blockchain system. In collaboration with Task 2.3. LGL will perform a cross-validation of all the herein presented





results while BioCoS will standardize and finalize the DNA kits for the EVOO authentication process that will be applied during the pilot in Task 4.2.



3 ENHANCED FOOD FRAUD DETECTION WITH ADVANCED SPECTROSCOPY

3.1 Introduction

Near-infrared (NIR) spectroscopy technology, which uses the 700nm-2500nm range of the electromagnetic spectrum, is a powerful analytical tool that can be employed in reflection, transmission, or transreflectance modes (Figure 3) to extract detailed chemical and physical information based on the molecular vibrational behaviour. This technology offers significant advantages across various applications due to its non-destructive, rapid, and versatile nature, such as agriculture, pharmaceuticals, food safety and environmental monitoring⁸.

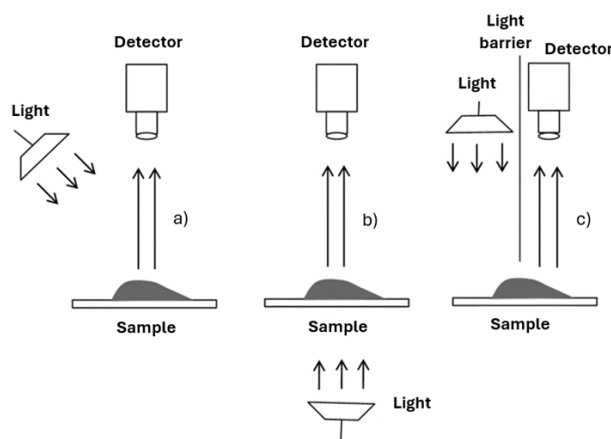


Figure 3 Image acquisition modes: (a) reflectance mode, (b) transmittance mode and (c) transreflectance mode⁹.

In agriculture, NIR technology facilitates accurate crop monitoring and soil analysis by measuring parameters such as moisture content and nutrient levels, thus optimizing the use of resources and improving crop yields^{10,11}. On the other hand, for pharmaceuticals, NIR spectroscopy provides a rapid and non-destructive measurement of different solid-state pharmaceuticals, supporting conventional quality control. This speeds up the manufacturing process and ensures product quality^{12,13}.

In the field of food analysis, NIR technology has evolved to meet the industry's changing demands. Various NIR devices have been developed, as advanced laboratory NIR

⁸ Grossmann, Luiz & Borges, Marco. (2017). Pharmaceutical applications using NIR technology in the cloud. 102100S. 10.1117/12.2264239. Cao, Nanning and Baijing Cao. "NIRS in the contemporary world for food and agriculture." NIR News 31 (2020): 23 – 24

⁹ Lodhi, Vaibhav & Chakravarty, Debashish & Mitra, Pabitra. (2019). Hyperspectral Imaging System: Development Aspects and Recent Trends. Sensing and Imaging An International Journal.

¹⁰ Dale, Laura Monica et al. "Chemometric Tools for NIRS and NIR Hyperspectral Imaging." Bulletin of University of Agricultural Sciences and Veterinary Medicine Cluj-Napoca. Agriculture (2012).

¹¹ Verma, Badal et al. "Enhancing Precision Agriculture and Environmental Monitoring Using Proximal Remote Sensing." Journal of Experimental Agriculture International (2023).

¹² Johansson, Jonas et al. "Time-Resolved NIR Spectroscopy for Analysis of Solid Pharmaceuticals." NIR News 17 (2006): 7-9.

¹³ Grossmann, Luiz and Marco Antonio Costa Borges. "Pharmaceutical applications using NIR technology in the cloud." Commercial + Scientific Sensing and Imaging (2017).

spectrometers with high accuracy and spectral resolution for detailed compositional studies and quality control assessments in fundamental research, like NIR ASD LabSpec 4. Additionally, portable NIR devices have gained popularity for on-site measurements due to their compact size, ease of use and ability to provide real-time results. These portable instruments will empower non-scientists to meet the challenges of quality control and authentication in everyday situations. They have become essential tools in the agri-food industry, with notable applications such as fruit maturity monitoring, meat quality assessment, and dairy analysis^{14,15}. In addition, the field of environmental monitoring is benefiting from NIR technologies thanks to the development of sensitive broadband photodetectors to detect environmental contaminants. These detectors offer high sensitivity and specificity over a wide wavelength range, which facilitates air and water quality monitoring and contributes to pollution control.¹⁶

One of the most prominent examples is its application in wheat flour production. During the milling process, wheat samples are taken directly from silos or transport trucks. These samples represent different batches of wheat and are used to assess the quality of the raw material. In this case, an initial calibration is performed to relate the NIR spectra (Infratec™ NOVA) to the desired quality parameters (such as protein, moisture and gluten content), and based on these results, process parameters such as blending of different types of wheat are adjusted.¹⁷

These examples highlight the continuing evolution and increasing usefulness of NIR instruments in a variety of applications, ranging from agriculture to food analysis and environmental monitoring.

Hyperspectral imaging (HSI) technology has also gained relevance as a tool in various applications, including food analysis, by combining spatial and spectral information. In that way, detailed spectral information (400nm-2500nm) from each pixel of the image is provided. Each pixel is analysed across selected wavelengths to capture detailed spectral information, allowing a comprehensive breakdown of the radiation received by each pixel into multiple spectral bands (Armin Schneider, 2017).¹⁸ Summarizing, HSI involves capturing and processing data from the entire electromagnetic spectrum, allowing objects to be accurately identified and classified based on their unique spectral signatures.¹⁹ Figure 4 presents a hyperspectral cube in which each slice of the cube represents images captured at different wavelengths.

¹⁴ Dos Santos CAT, Lopo M, Páscoa RNMJ, Lopes JA. A Review on the Applications of Portable Near-Infrared Spectrometers in the Agro-Food Industry. *Applied Spectroscopy*. 67 (2013):1215-1233

¹⁵ Wang, Jingyao et al. "Self-Powered and Broadband Photodetectors Based on High-performance Mixed Dimensional Sb₂O₃/PdTe₂/Si Heterojunction for Multiplex Environmental Monitoring." (2023).

¹⁶ Mesquita, Pedro et al. "Low-cost microfluidics: Towards affordable environmental monitoring and assessment." *Frontiers in Lab on a Chip Technologies* (2022).

¹⁷ <https://www.fossanalytics.com>

¹⁸ Schneider, Armin & Feussner, Hubertus. (2017). *Biomedical Engineering in Gastrointestinal Surgery*.

¹⁹ Zahra, Anam et al. "Current advances in imaging spectroscopy and its state-of-the-art applications." *Expert Syst. Appl.* 238 (2023): 122172.



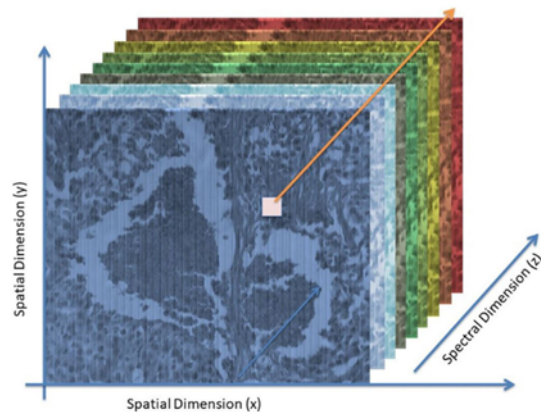


Figure 4 Representative image of a hyperspectral cube²⁰

This HSI technology offers a versatile and powerful tool for a wide range of applications. In the biomedical field, HSI systems using broadband LED light sources have revolutionised research by providing a detailed view of samples without causing heat damage. Meticulous calibration and verification processes ensure the reliability and standardisation of these systems, paving the way for advances in biomedical imaging and analysis.²¹

In addition, HSI technology also plays a crucial role in agricultural quality analysis, as it enables rapid and non-destructive evaluation of produce for disease detection, grading and chemical attribute assessment. Its effectiveness in improving quality analysis processes underlines its importance in agriculture, contributing to the development of accurate and efficient detection methods.²²

HSI devices can also come in a variety of forms, from laboratory systems, with high spectral resolution for detailed research and analysis, to portable devices for on-site inspections and field applications. These devices, as in the case of portable NIR devices, offer flexibility and versatility across several applications, such as biomedical, agriculture and food industry, enabling real-time monitoring, ensuring product quality and process optimisation in the different sectors.

The distinct characteristics of HSI and NIR spectroscopy present unique advantages and disadvantages across a range of applications. HSI provides high spatial resolution and rich chemical information, enabling precise feature localization and detailed chemical analysis. However, it involves complex data processing, limited penetration depth, and sophisticated equipment. On the other hand, NIR spectroscopy offers non-destructive analysis, rapid results and versatility across industries. However, it may have limited chemical specificity, required sample homogeneity and involved complex calibrations. The choice between HSI and NIR depends on specific application needs, considering factors like spatial resolution, chemical analysis requirements, and cost-effectiveness.

²⁰ Khouj, Yasser & Dawson, J. & Coad, James & Von-Davis, Linda. (2018). Hyperspectral Imaging and K-Means Classification for Histologic Evaluation of Ductal Carcinoma In Situ. *Frontiers in Oncology*.

²¹ Stergar, Jošt et al. "Design and Validation of a Custom-Made Laboratory Hyperspectral Imaging System for Biomedical Applications Using a Broadband LED Light Source." *Sensors (Basel, Switzerland)* vol. 22,16 6274 (2022).

²² Wang, Baodi et al. "The Applications of Hyperspectral Imaging Technology for Agricultural Products Quality Analysis: A Review." *Food Reviews International* 39 (2021): 1043 - 1062.



By following these steps, developers can create effective NIR and HSI applications tailored to specific needs, whether in agriculture, environmental monitoring or other fields requiring detailed spectral information.

A clear example is the need to detect the fraudulent practices carried out in PGI Asturian Faba beans (mixing PGI certified beans with imported ones from Argentina, Bolivia or Mexico, much cheaper, and selling them as certified by the PGI). The most common practice to characterize the PGI faba during routine fraud inspections is through visual inspections, which is based only on the expert knowledge of the person in charge. Other characterization techniques, based on physio-chemical, microbiology and sensory analysis, follow expensive and time-consuming lab protocols, limiting the number of inspections. Moreover, with the results of these analyses it is not possible to differentiate the origin of the faba bean and, as result, the development of much more sophisticated analytical protocols is highly demanded.

Therefore, main demand of end-users and stakeholders responsible for combating fraud in PGI Asturian Faba beans is the development of analytical tools that are low-cost, portable, simple to be used by non-experienced operators and able to produce the outputs in real-time, or quasi, with minimal sample pre-treatments. NIR and HSI based tools previously described and tailored to detect fraud in PGI Asturian Faba beans are perfectly aligned with the demands and needs of end-users and stakeholders in charge of the inspections for avoiding the fraud of PGI faba due to its mixture with beans from other cheaper origins. Moreover, the PGI is also really interested in advancing the development of a specific fingerprint for each PGI plot, avoiding also a common fraud practice: selling under a specific PGI license number beans from other producers. The development of NIR and HSI based tools could also provide respond to this need.

3.1.1 Portable NIR for food applications

The most decisive breakthrough that took place in relation to this technology in terms of instrumental development and progress occurred with the appearance of Fourier transform NIR spectrometers (FT-NIR) in the early 1990s, and the second technological breakthrough can be associated with the introduction of portable instruments in the 2000s.²³ However, from a conceptual point of view, the move towards portability formed a much more decisive change in the application horizon of this technique, as it was instrumental in moving the analysis from the laboratory to on-site analysis, which brought particular benefits to the agri-food sector, as the real-time obtention of the results, the potential to monitor the whole production, or the possibility to use the device for multiple applications (including the required calibrations) with the same device. Therefore, being a versatile and simple technology, a rapid acceleration in the development and optimization of these portable devices has been observed by significantly increasing the flexibility of the analysis.

Some commercial portable NIR instruments have promoted significant advances in various agrifood applications. A relevant example is *Mmeter*, introduced by Fred McClure, a portable NIR used for manure assessment. In addition, multi-purpose spectrophotometers such as the *MMS1* and the *Corona*, manufactured by Zeiss, have been widely used for real-time forage composition analysis and other applications.²⁴

²³ <https://www.bruker.com>

²⁴ Yan, Hui et al. "Handheld Near-Infrared Spectroscopy: State-of-the-Art Instrumentation and Applications in Material Identification, Food Authentication, and Environmental Investigations." *Chemosensors* (2023)



For instance, NIR spectroscopy has been utilized in cheese-making processes to monitor milk coagulation, with applications including determining curd cutting time and assessing milk coagulation dynamics. The use of commercial probes equipped with LEDs and photodetectors has facilitated real-time monitoring of physicochemical changes during cheese production. These innovative applications demonstrate the versatility and effectiveness of portable NIR technology in optimizing food processing procedures and ensuring product quality.²⁵

Despite the aforementioned characteristics, portable configuration still presents a number of challenges that must be overcome such as ongoing maintenance of NIR models, variability in chemical composition and environmental factors, as well as data security and protection in cloud-based solutions. These factors may affect their performance depending on the different scenarios in which these devices are used, as well as the chemical complexity, granularity and heterogeneity of food matrices, being crucial the use of different chemometric methods to achieve an accurate understanding of the data and solve food authentication problems.²⁶

3.1.2 HSI for food applications

Portable hyperspectral imaging (HSI) devices are increasingly utilized in the food industry for various applications related to food quality and safety control. One example is the *SOC-710 Instrument*, which is employed for monitoring changes in plant health, nutrient levels, water stress, disease and insect infestations, enabling precise agricultural management practices. Another example includes *Imec Snapshot Cameras*, which offer a cost-effective solution for collecting high-resolution spectral data in real-time, simplifying the scanning process and enhancing efficiency in agricultural applications. These examples highlight the ongoing advancements in portable HSI technology, making it more accessible and practical for a wide range of applications, from remote sensing to environmental monitoring.²⁷⁻²⁹

HSI technology in food applications also presents certain challenges, such as the complexity of data analysis due to the large amount of information generated. This requires the use of sophisticated machine learning techniques to extract meaningful data efficiently,³⁰ the need to establish uniform calibration protocols and methods to ensure consistent and reliable results depending on the processing environment and the variability of food samples, and also to ensure that HSI complies with regulatory standards for food safety and quality assessment, which is challenging due to constantly evolving and changing regulations.³¹ Another important aspect that needs to be considered is the implementation of this technology into existing food processing systems due to some technical challenges in terms of compatibility and integration within the workflow.

²⁵ De Bock, Maarten et al. "Miniature NIR spectrometer for mobile applications." OPTO (2022).

²⁶ Beć, K.B.; Grabska, J.; Huck, C.W. Miniaturized NIR Spectroscopy in Food Analysis and Quality Control: Promises, Challenges, and Perspectives. *Foods* (2022), 11, 1465

²⁷ <https://surfaceoptics.com/applications/precision-agriculture-hyperspectral-sensors/>

²⁸ https://www.photonics.com/Articles/The_Food_Industrys_Appetite_for_Hyperspectral/a66946

²⁹ <https://www.portableas.com/news/hyperspectral-imaging-in-the-food-industry/>

³⁰ Feng, Chaohui et al. "Hyperspectral imaging and multispectral imaging as the novel techniques for detecting defects in raw and processed meat products: Current state-of-the-art research advances." *Food Control* 84 (2018): 165-176.

³¹ Ayaz, H.; Ahmad, M.; Mazzara, M.; Sohaib, A. Hyperspectral Imaging for Minced Meat Classification Using Nonlinear Deep Features. *Appl. Sci.* 2020, 10, 7783



3.1.3 Portable NIR and HSI for food authenticity

The recent evolution of portable NIR and HSI technologies for food purposes has seen substantial advances, especially in the area of food authentication. These advanced technologies are now being used more frequently to detect food adulteration, confirm geographic origins, and discern specific agricultural practices, among other critical applications.^{32,33} In addition, as mentioned above, the portability of these devices allows for on-site measurements throughout the supply chain, significantly reducing the time and costs associated with food quality and authenticity assessments. An example is their application in the study of Iberian hams.³⁵ These technologies, in particular portable NIR, have been instrumental in categorizing pork carcasses according to their fatty acid compositions.³⁴ In this case, the profiles obtained serve as markers of the diet and breeding conditions of the pigs, parameters of great importance in products such as Iberian ham, where the diet and breed of the pigs significantly influence the quality and authenticity of the final product.³⁵ Other scientific research has demonstrated the effectiveness of using these technologies for seed origin and variety authentication, as is the case with HSI technology to discriminate between three Iranian (Shiroudi, Khazar, and Hashemi) rice varieties (Edris, M. et al., 2024), a study that demonstrated the ability of HSI to capture detailed spatial and spectral information to identify unique spectral signatures associated with specific Iranian rice varieties³⁶. Other examples of interest are the use of portable NIR spectroscopy to classify soybeans according to their geographical origin (Li et al., 2024)³⁷. The researchers developed a model using the spectral data to differentiate soybeans from different regions with high accuracy, or the use of portable HSI technology to determine the geographical origin of mung beans.³⁸

While these technologies have demonstrated immense potential in assuring the quality and authenticity of food products, their application is not without challenges and limitations. A primary concern is the precision and reliability of the calibrations used to interpret spectral data. Portable NIR instruments can classify various parameters with high probabilities, but there is a need for further refinement to enhance the accuracy and robustness of these calibrations. The miniaturization of NIR instruments, a necessity for portability, can diminish the wavelength range and resolution, potentially impacting the accuracy of some calibrations.³⁹ Another constraint is the requirement for a substantial number of samples to build robust predictive models. This requirement poses a hurdle to the widespread adoption of these technologies as it necessitates more effort.

³² Ayaz, H.; Ahmad, M.; Mazzara, M.; Sohaib, A. Hyperspectral Imaging for Minced Meat Classification Using Nonlinear Deep Features. *Appl. Sci.* 2020, 10, 7783

³³ Kharbach, M.; Alaoui Mansouri, M.; Taabouz, M.; Yu, H. Current Application of Advancing Spectroscopy Techniques in Food Analysis: Data Handling with Chemometric Approaches. *Foods* 2023, 12, 2753

³⁴ Feng, Lei et al. "Application of Visible/Infrared Spectroscopy and Hyperspectral Imaging with Machine Learning Techniques for Identifying Food Varieties and Geographical Origins." *Frontiers in nutrition* vol. 8 680357, (2021).

³⁵ Piotrowski, C et al. "Short Communication: The potential of portable near infrared spectroscopy for assuring quality and authenticity in the food chain, using Iberian hams as an example." *Animal: an international journal of animal bioscience* vol. 13,12 (2019), 3018-3021

³⁶ Edris, Mahsa et al. "Identifying the Authenticity and Geographical Origin of Rice by Analyzing Hyperspectral Images Using Unsupervised Clustering Algorithms." *Journal of Food Composition and Analysis* (2023).

³⁷ Li, X., Wang, D., Yu, L., Ma, F., Wang, X., Pérez-Marín, D., Li, P., & Zhang, L. Origin traceability and adulteration detection of soybean using near infrared hyperspectral imaging. *Food Frontiers*, (2024) 00, 1–8

³⁸ Kaewkarn Phuangsombut, Te Ma, Tetsuya Inagaki, Satoru Tsuchikawa & Anupun Terdwongworakul. Near-infrared hyperspectral imaging for classification of mung bean seeds, *International Journal of Food Properties*, (2018) 21:1, 799-807

³⁹ Mendez, Jeffrey et al. Trends in application of NIR and hyperspectral imaging for food authentication. *Scientia Agropecuaria*. 2019, vol.10, n.1, pp.143-161



Therefore, future advances in portable NIR and HSI technologies are expected to focus on increasing instrument accuracy, robustness, and ease of use. This could include advances in sensor technology, data processing algorithms, and intuitive user interfaces. In addition, as these technologies evolve, broader commercial adoption is anticipated, which will consequently validate the expense of collecting larger sample sets for calibration purposes.

3.1.4 Methodology for the model development

The core component of most NIR-based applications consists in the chemometric model able to correlate the sensor signal (spectra) with the target parameter. The common methodology for the development of this chemometrics involves the following steps:

- 1) Experimental design and optimization of NIR measurements
 - a. Define the number of samples and their variability depending on the matrix and the target parameter. Typically, about 100 samples are considered for the development of a robust model.
 - b. Critical elements to be considered include sample representativeness and consistency, as well as ensuring a diverse sample set that covers the expected variability in the target parameter.
- 2) Data acquisition
 - a. Acquire two types of data: NIR spectra and analysis of the target parameter using an established reference method. These data will form the training database for model development.
- 3) Data pre-processing
 - a. Pre-process the NIR spectra data to improve signal quality and remove noise. Typical methods include baseline correction, smoothing, normalisation and derivative transformations.
- 4) Intelligent data processing
 - a. Use of advanced statistical and mathematical techniques for model development. Traditional methods, generally linear, such as partial least squares (PLS) regression are often used. In recent years, non-linear methods including cutting edge Artificial Intelligence techniques, as machine learning, support vector machines or neural networks, have been extensively applied in this NIR studies with promising results.
 - b. At this point a relevant decision to be made is whether regression or classification techniques are more appropriate depending on the nature of the target parameter.
- 5) Results visualisation interface
 - a. Develop a user-friendly interface, easy to be used by non-experts, to visualise model results, predictions, and model performance metrics.

By following these steps, researchers and industry professionals can effectively develop robust NIR applications for quality assessment and process optimisation.

Therefore, as a conclusion of this introduction section, it can be stated that the versatility of NIR and HSI technologies plays a key role in several fields, ensuring the safety and quality of products, as well as improving process efficiency in different industrial environments.

3.2 Overall description of the use case

The main objective of this pilot is to develop and validate, at operational environment, a digital tool based on the use of low-cost and portable advanced optical sensors (NIR and in-line HSI





technologies), for the detection of main fraudulent practices in the PGI Asturian Faba bean (IGPFA, “IGP Faba de Asturias”). This can be summarised in these two main use cases:

- (1) Mix of PGI certified beans with cheaper foreign material (generally from South America).
- (2) Mixture of bean batches from different certified lots (proof of concept aiming to define a lot of fingerprints).

Main end-users will be:

- PGI control body; and
- competent public authority dealing with food quality and authenticity.

During the validation they will incorporate ALLIANCE tools within their control measures and protocols to evaluate its performance and compare it against control protocols.

PGI Asturian faba are dried, shelled beans from the *Phaseolus vulgaris* L. species, which comprises the traditional “Granja Asturiana” variety, it should be cropped in the region of Asturias, northwest Spain, and must comply with different minimum requirements as being within the Extra and First categories, whole grains, healthy, free from mold, botrytis and insects, and with a maximum moisture of a 15%. It has creamy white colour, kidney, long and flat shape, and a large size with about 100-110 beans/100g of seeds.⁴⁰ Common nutritional values are summarized below:

- Characteristics (gr/100 g):
- Carbohydrates: 50 - 70
- Protein: 20 - 30
- Moisture: 11-14
- Total Fiber: 4 - 5
- Ash: 3 - 5
- Magnesium: 0.09
- Calcium: 0.04
- Fat: 0.3 - 1.5
- Iron: 66.44 (expressed in mg/kg)

This PGI faba bean is the main ingredient of the most popular dish of Asturias, named “Fabada Asturiana”, a bean stew including cured meat (chorizo, black pudding and bacon).

Key figures for IGPFA during the 2022-2023 campaign are shown below:

- 135 producers
- 30 industrial companies
- 48 producers with own brand
- 404 lots
- 218 hectares

⁴⁰ https://www.mapa.gob.es/images/es/faba_asturiana1996_06_12_tcm30-210853.pdf



3.3 Achieved results

The works carried out during this first reporting period were focused on the first use case presented above: to develop a fast, non-destructive and robust method to authenticate the geographical origin of beans, as geographical origin significantly influences quality and price. The method is based on the use of NIR and HSI technologies, combined with advanced machine learning algorithms, to accurately classify beans based on their origin. This method aims to address the problem of food fraud, such as the mixing of Protected Geographical Indication (PGI) certified Asturian Faba beans with cheaper imports from South America, providing a reliable, cost-effective and easy-to-use solution for on-site verification.

3.3.1 Experimental design

Methodological approach: The methodological approach applied for this study could be divided in three common stages covered during the development of a NIR-based model: (i) data acquisition, followed by (ii) intelligent data analysis, and finally (iii) the results visualization. Regarding data acquisition, two main databases containing raw data are generated. On the one side, multiple physico-chemical parameters of the beans, from Asturias and other origins, are measured using conventional laboratory instruments. On the other, NIR spectra and HSI images of all this set of beans are acquired, generating the second database. Afterwards, these raw data are processed using advanced statistical, mathematical and Artificial Intelligence techniques. And finally, the analysis result is shown through a user interface. The main fundamental idea of this study is to find differences between the physico-chemical and/or spectral characterization beans and through them build a chemometric model able to classify a specific bean as Asturian or foreign, with the subsequent detection of frauds (Figure 5).

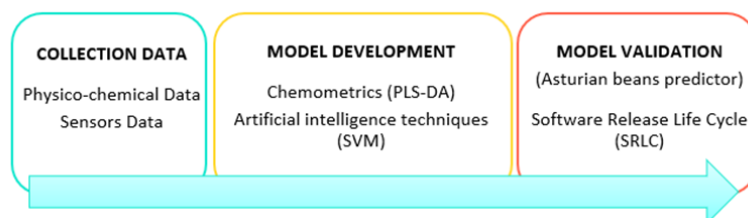


Figure 5 Methodological approach scheme.

Based on the historical data provided by IGPFA as well as related scientific literature,⁴¹ Galicia (adjacent Spanish region to Asturias to the west) and South America (specially Bolivia) are the main origins of the foreign beans that are mixed with PGI Asturian ones in the cases of fraud. So, to ensure a good coverage of the potential bean variability a large set of samples were collected from all different municipalities of Asturias producing faba), Galicia and Bolivia for the 2022 and 2023 campaigns.

Number of samples: For 2022 harvest, a total of thirty-six samples were available, with twenty-eight from Asturias (one per municipality producing faba) (see Figure 6), two from Galicia and six from Bolivia. For the 2023 harvest, a larger number of samples are available, with a total of

⁴¹ Reguera Galán, A.; Moldovan Feier, M.; García Alonso, J. I.. The combined measurement of ⁸⁷Sr/⁸⁶Sr isotope ratios and ⁸⁸Sr/⁸⁵Rb elemental ratios using laser ablation MC-ICP-MS and its application for food provenance studies: The case for Asturian beans. *Journal of Analytical Atomic Spectrometry*, 33(5), p. 867-875 (2018); doi:10.1039/c8ja00061a



forty-eight, thirty-three from Asturias (samples from different plots within the same municipality), one from Galicia and fourteen from Bolivia. IGPFPA is still actively collecting samples from this harvest, focusing on foreign beans from Bolivia and Galicia.



Figure 6 Map of the municipalities in the Principality of Asturias.

To ensure the quality of data collection and spectra for analysis, as well as to determine the optimal sample format for the analysis, three distinct sample configurations were considered: whole beans, longitudinal cut beans and milled bean (Figure 7).

Analysis protocol: From each sample 100 grains were initially weighed, following by the selection of 20 grains from there and they are used as material for the physico-chemical analysis and the acquisition of NIR spectra and HSI images.



Figure 7 Measuring formats for beans.

The **physicochemical analysis** covered a comprehensive evaluation of both whole and ground beans to assess several key parameters. In the case of whole beans, the analysis includes morphology (caliper), colour attributes (colourimeter), weight as well as water absorption capacity using common laboratory protocols (see sec 8).



Figure 8 Caliper and colorimeter used to obtain morphological and colour data of faba beans in all formats.

In the case of milled beans measurements were focused on specific parameters such as moisture levels (oven-drying method) and colour properties (colourimeter).

On the other hand, in order to carry out **NIR and HSI spectral measurements** on beans, it is essential to have a detailed protocol that guarantees accurate and reliable measurements. For NIR spectra, two different devices were used in diffuse reflectance mode: Portable NIR-S-G1 InnoSpectra with a wavelength range from 900 to 1700 nm, to measure whole and longitudinal cut beans on both sides (Figure 9a), and NIR ASD LabSpec 4, with a wavelength range from 350 to 2500 nm to obtain the spectra of milled beans (Figure 9b).

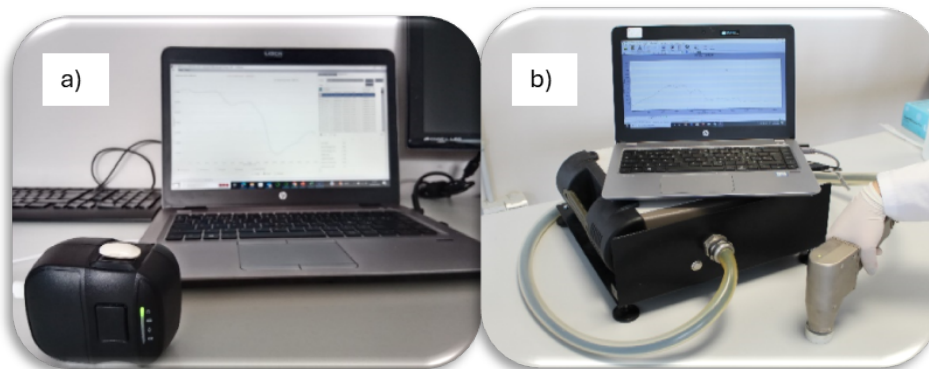


Figure 9 NIR measurements with a) portable NIR-S-G1 InnoSpectra equipment measuring whole beans and b) desktop NIR ASD LabSpec 4 equipment measuring milled beans.

In the case of obtaining the spectra by HSI, two types of hyperspectral cameras operating in diffuse reflectance mode were used to take the hyperspectral images, namely the FX-10 (Visible-Near Infrared) and FX-17 (Near Infrared) cameras from the company Specim (see specification in appendix).

These two hyperspectral cameras are intended to be used on conveyor belts or sliding trays, as they operate in "pushbroom line scanning technique" mode, which means that the cameras have a line of sensors that capture the light coming from the sample at a defined speed (FPS - frames per second). For both hyperspectral cameras, the configuration use has been the same, changing only the type of hyperspectral camera (Figure 10).

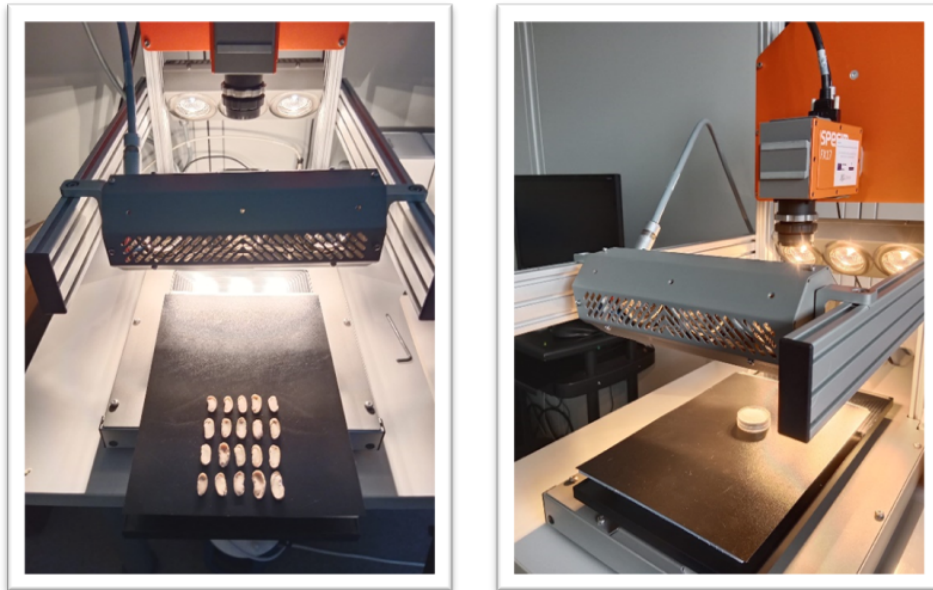


Figure 10 Hyperspectral imaging configuration for longitudinal cut beans (same as for whole beans) and milled beans.

After several optimization tests, the optimal geometrical configuration and optical parameters for the hyperspectral cameras were set as follows:

- Height of the cameras above the movable scanner tray: 15 cm
- Number of frames per second: FX-10: 50 Hz and FX-17: 50 Hz.
- Exposure times: FX-10: 13 ms and FX-17: 6 ms.
- Tray speed: FX-10: 8.6 mm/s and FX-17: 12.70 mm/s.
- Halogen lights placed at an angle to the camera of 45°, both at the same height above the moving tray and in line with the hyperspectral camera

The attached diagram summarises the NIR and HSI measurements acquired with the different bean formats (Figure 11):

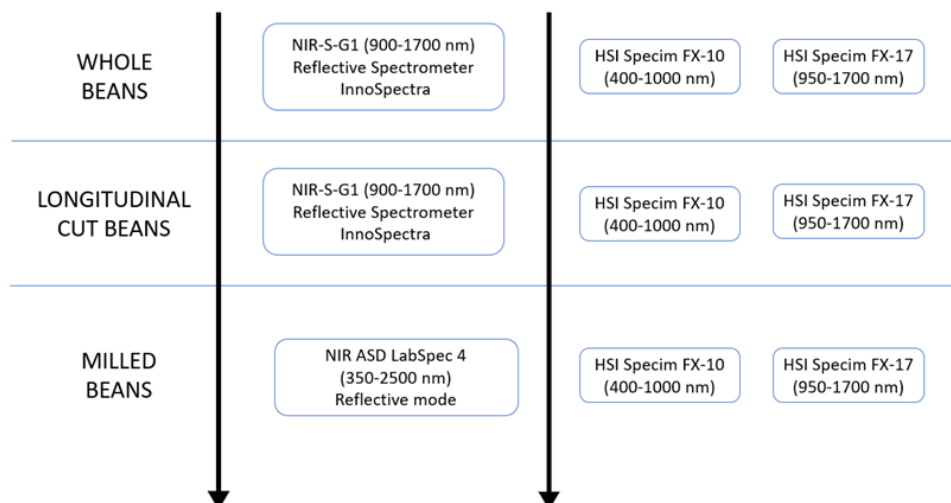


Figure 11 Protocol measurement NIR and HSI schema.



To finalise, below scheme represents visually the measurement workflow for both the physico-chemical properties and the spectra (Figure 12):

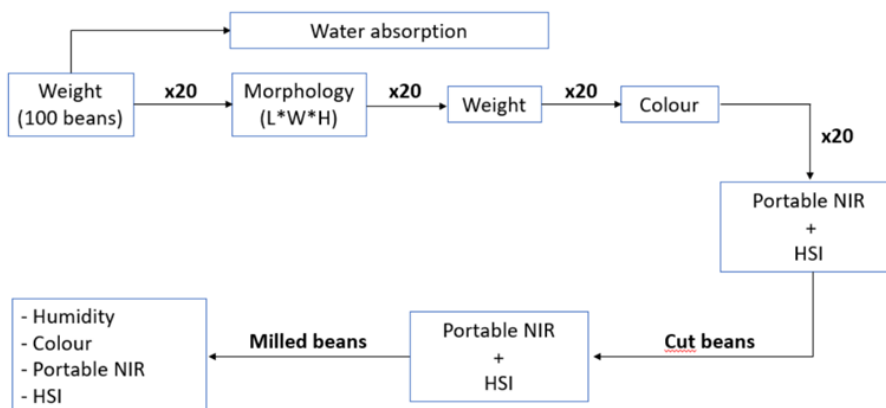


Figure 12 Protocol measurement physico-chemical and spectra schema.

3.3.2 Data collection

To date, all (36) bean samples from the 2022 harvest have been measured, compiling the data shown in the following tables (Table 4 and Table 5):

Table 4 Physico-chemical parameters collected to date by classical laboratory techniques.

	WHOLE bean	LONGITUDINALY CUT bean	MILLED bean
Morphology (length, width and thickness)	2160	-	-
Colour (L*, a*, b*)	720	-	108
Weight/100 beans	36		
Weight/ each bean	720	-	-
Humidity	-	-	36
Absorption	36	-	-

Table 5 Information gathered to date by spectroscopy techniques.

	WHOLE bean	LONGITUDINALY CUT bean	MILLED bean
Portable NIR Inno	1440 spectra (two spectra for each bean)	1440 spectra (two spectra for each bean)	0
HSI-FX10 camera	36 images (20 beans for each image)	36 images (20 beans for each image)	36 images (one foe each sample)
HSI-FX17 camera	36 images (20 beans for each image)	36 images (20 beans for each image)	36 images (one foe each sample)
NIR ASD	0	0	108 spectra (three spectra for each sample)

This data set collected through the experimental protocol defined in the previous section forms the basis for further data analysis and the development of models to take advantage of NIR and HSI technologies for food authentication purposes.

For HSI, it should be noted that in order to use the spectral information contained in the hyperspectral images it is necessary to treat them in order to obtain an average spectrum for



each bean. To achieve this Python language will be used and following steps are followed (Figure 13):

- Loading in memory the hyperspectral image of the beans, with its corresponding white (reference) and black (dark current).
- Separation of the area where the bean is located from the rest of the image. For this purpose, segmentation techniques based on thresholding were used, identifying the regions with different spectral characteristics and then extracting the spectra for each pixel of those regions.
- Finally, all the pixels in the area of this mask are averaged to get an average spectrum of the area that will be used as representative for this faba sample.

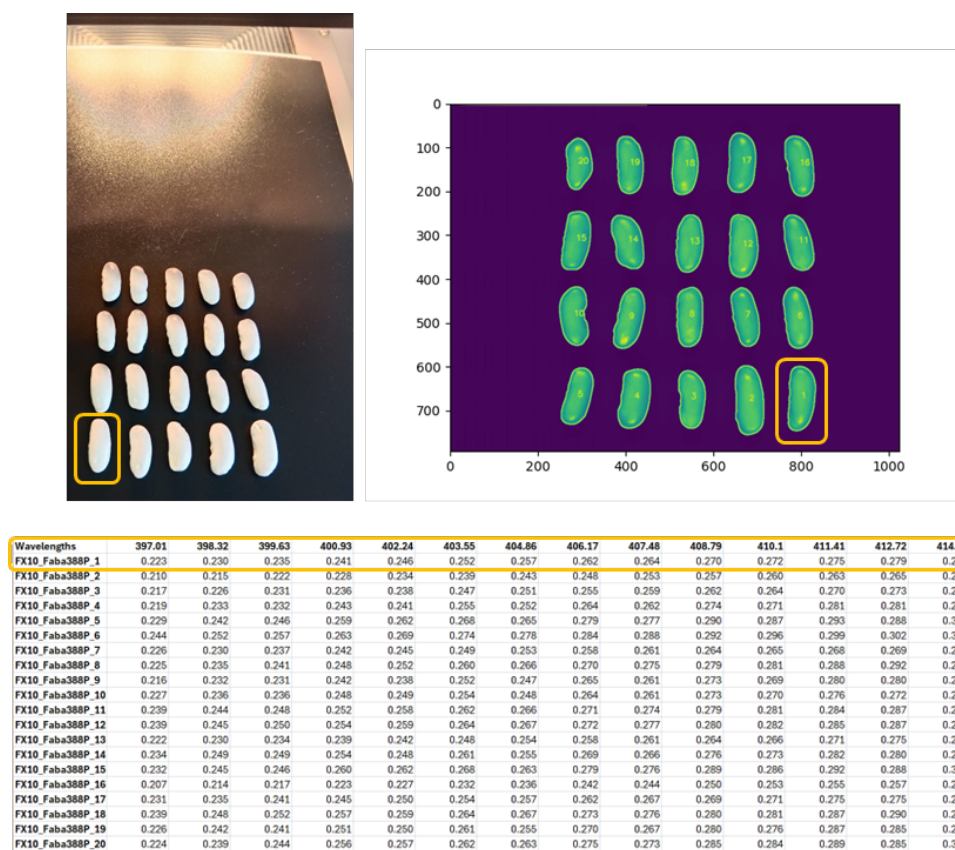


Figure 13 Example for obtaining an average spectrum of each faba bean from the HSI image.

Each average spectra are saved in a ".csv" file format, which will then be imported into Python for advanced data processing.

Moreover, to reduce the time needed in the laboratory to determine manually the morphological features of each faba (length and width), using the caliper, the development of a Python code able to extract these features automatically from the HSI images, containing the 20 beans, was explored. It involved the use of a reference template with known-sized basic shapes and upon detecting each faba bean, a rectangle was generated around its edges. The length of the bean was determined by the long side of the rectangle, while the width was calculated by drawing a line at the center of the bean (length/2) and considering only the pixels within the bean's boundaries (Figure 14).



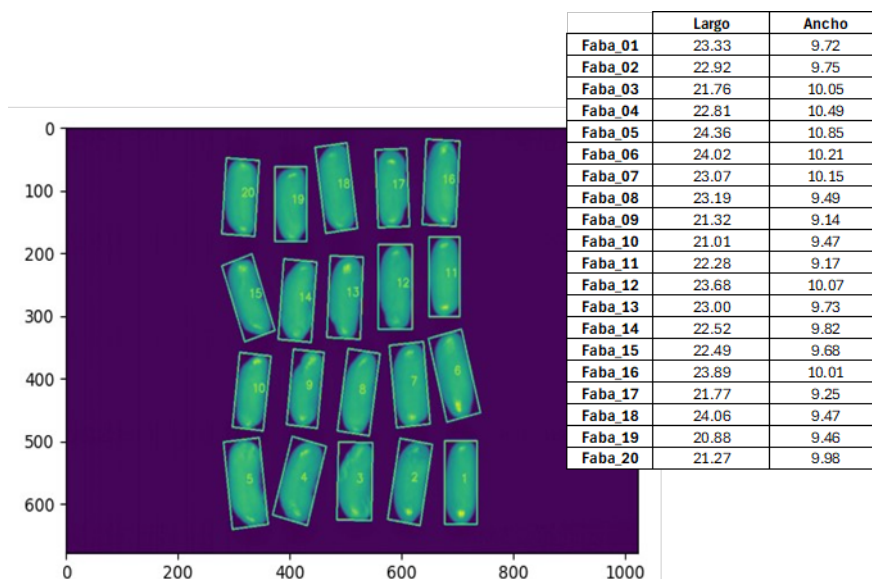


Figure 14 Sample data showcasing the length and width measurements extracted from an image captured using the FX10 camera.

The produced algorithm was tested using images captured with the FX10 camera, known for its higher resolution compared to the FX17 camera, which allows more accurate length and width determinations. However, when the results of the algorithm are compared with the caliper measurements, relevant discrepancies in the image-derived values were observed. These discrepancies are due to the inherent errors of the camera itself as well as variations in the thickness of the grains, which results in variations in the distance between the camera lens and the sample leading to inaccuracies in the measurements. Therefore, this approach was discarded.

3.3.3 Intelligent data processing

Physico-chemical properties

Initially, a statistical analysis was conducted utilizing both box and whisker plots as well as violin plots. Box and whisker plots visually represent data distribution by showing the median, quartiles, and potential outliers. On the other hand, violin plots offer a more detailed visualization of data distribution, enhancing the comparison among the various origins of the beans. Together, these graphical representations provide a comprehensive overview of the data's central tendency and spread, simplifying comparisons between different groups or datasets.

Based on the findings, it can be concluded that significant differences among faba beans from Asturias, Bolivia, and Galicia are observed in relation to colour and water absorption (Figure 15 and Figure 16), while parameters like morphology or weight, commonly remarked by producers as distinction features, do not exhibit notable changes, therefore these parameters will not be measured in the next round of physico-chemical measurements. Regarding moisture content in ground beans, there is considerable variability within the same group, indicating the need for additional data to draw a reliable conclusion. It is important to highlight at this point the limited sample size of faba beans from Galicia (two), being needed a reassessment when more material of this origin is available. This should be done also for the case of Bolivia to enhance the robustness of the results. All these is planned for the next months.



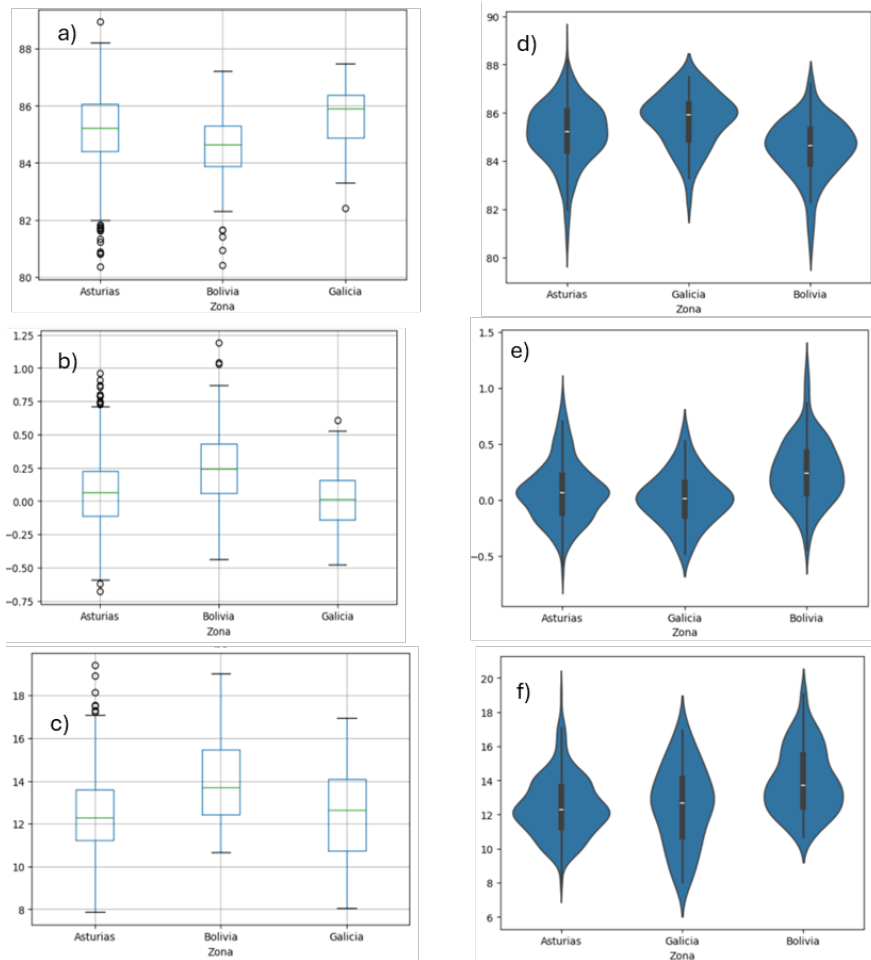


Figure 15 Box and whisker diagrams (left) and violin diagrams (right) for colour (parameters $L^*a^*b^*$). L^* in the top, a^* middle, b^* down.

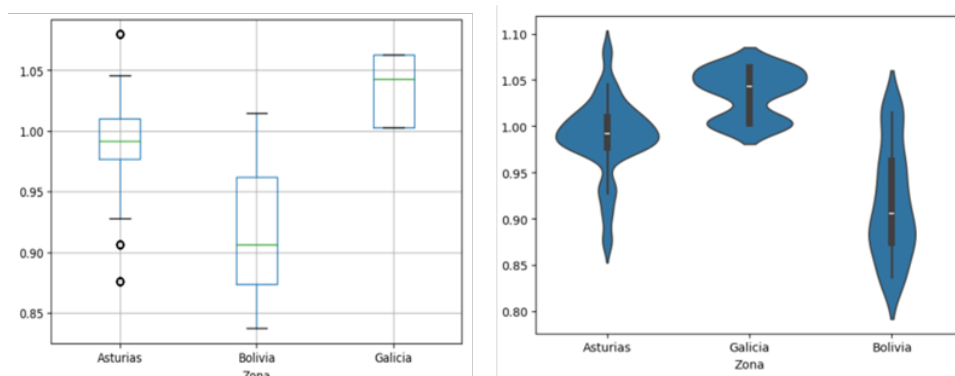


Figure 16 Box and whisker diagrams (left) and violin diagrams (right) for water absorption.

NIR and HSI spectra analysis

The spectroscopic data obtained by NIR and HSI are analysed using Python, a software widely used in data analysis, and the spectra are stored in a Python data frame together with their corresponding class labels. Figure 17 shows some of the reflection spectra used for the analyses (whole and longitudinal cut bean). It should be noted that spectrum ranges between 930 and 1670 nm for the portable device NIR-S-G1 and the hyperspectral camera FX17 and



430-970 nm for the FX10 camera are used, since measurements close to the limit of the devices present a lot of noise.

In the development of a classification model, three key consecutive steps could be defined: pre-treatment of NIR and HSI data, selection of the variables (wavelengths), and development of classification models.

Pre-treatment involves several techniques that aim to improve the quality of the acquired data before it undergoes further analysis. The main objectives are noise reduction, outlier detection and normalization. Although the classification of these methods is complicated, as many of them could fall into several categories, the following are considered:

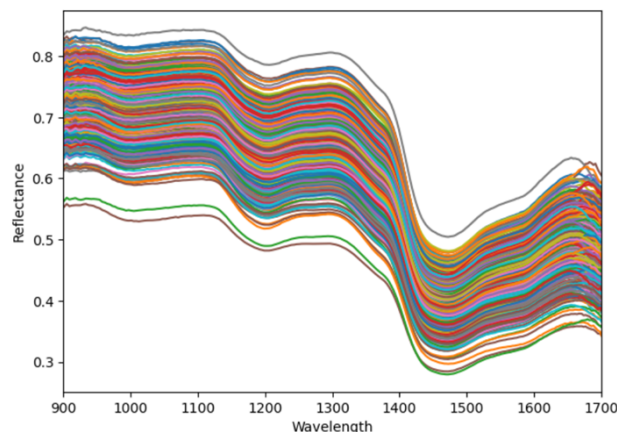


Figure 17 Spectra obtained with the NIR-S-G1 portable device for whole beans.

- **Smoothing:** are applied to reduce the noise in the spectral data; “Movil average”, “Median Filter” or SG “Savitzky Golay Smoother” are the most common.
- **Normalization:** adjust the scale of spectral data to reduce variability among samples that are not related to compositional differences. This category includes “Area Normalization”, MSC (“Multiplicative Signal Correction”), EMSC (“Extended Multiplicative Signal Correction”), SNV (“Standard Normal Variate”), “Autoscale Normalization”.
- **Baseline Correction and Derivation:** aimed at removing background effects and highlighting the spectral features of interest by correcting the baseline drift and enhancing the resolution of spectral features; some examples are “Savitzky-Golay”, “Gap-Derivative”, “Norris Gap” and “Detrend”.

These pre-treatment methods are designed for their application over the spectroscopic data and can be utilized individually or in combinations of up to three consecutive treatments. The selection of the appropriate pre-treatment methods is crucial and will be guided by their impact on the predictability and performance of the final classification models.

In the second step, a wavelength selection is performed. These methods play a vital role in determining which wavelengths are most relevant for discriminating between different materials or components through the spectroscopic data. By selecting the optimal set of wavelengths, classification models can achieve higher accuracy and robustness in identifying and categorizing various substances based on their spectral signatures. Some of the most well-known and widely used methods for wavelength selection are:

- **Stepwise selection methods:** involve iteratively adding or removing variables based on statistical criteria to improve the model performance (Forward Interval, Backward Interval or Stepwise Interval. Especially suitable for linear methods)
- **Recursive Feature Elimination (RFE):** works by recursively eliminating less important wavelengths and retraining the model. The importance of each wavelength is evaluated in terms of model performance.
- **Mutual information-based methods:** evaluates the relevance of each wavelength to the target variable. Wavelengths with high values of mutual information are considered important and are retained for further analysis.

- **Principal component analysis (PCA) loadings:** PCA is a dimensionality reduction technique that transforms the original spectral data into a new set of variables called principal components (PC).

Once the pre-treatment and variable selection stages have been completed, the classification model can be applied. For the first use case, several widely known classification models were tested, and their results were compared in order to select the one that yields the best statistical parameters based on the spectral data provided. One of the methods used was the Partial Least Squares Discriminant Analysis (PLS-DA), a powerful tool for classification tasks, especially with high-dimensional data sets, offering a robust approach to model and analyse complex data structures. From this model, combinations of PLS-DA with other algorithms that enhance multiclass prediction, such as softmax, Naive-Bayes, Random Forest, or Extreme Gradient Boosting (XGBoost), were produced and implemented. These complementing algorithms were trained using as input the predictions from the PLS-DA. This approach leverages the strengths of PLS-DA in handling high-dimensional datasets and discriminative variable selection, while also benefiting from the capabilities of other algorithms to improve the overall predictive performance in multiclass scenarios.

Alongside the PLS-DA model, the application of a supervised non-linear machine learning algorithm known as Support Vector Machine (SVM) is also explored in this study. It is highly utilized for classification and regression tasks. SVMs are robust classifiers adept at handling diverse classification challenges by identifying an optimal hyperplane that segregates distinct classes of data points. The algorithm's objective is to maximize the margin, representing the greatest distance between data points of varying classes, thereby enhancing the accuracy of future data point classification.

Galicia origin is excluded from this initial classification analysis due to the limited number of samples, two, that could potentially skew the model results. Consequently, following results are focused on two distinct origins: Asturias and Bolivia.

Analysis of spectral data using portable NIR

Spectral data analysis from NIR (NIR-SG1 Innospectra) (Figure 15) was conducted on whole and longitudinally cut beans. The raw dataset was split into training and test sets at an 80/20 ratio, followed by the application of various pre-treatments to the training set. These pre-treatments included SNV, SG, MSC, and DeTrend, applied in different combinations.

PCA was used as a method for wavelength selection. The goal was to reduce the dimensionality of the dataset by employing the support vector machine (SVM) method. Unlike other models that rely on the outputs of PLS-DA, where a specific number of principal components is already selected. We are also evaluating the feasibility of other methods, such as Stepwise selection, to further enhance feature selection.

Subsequently, the dimensionality of the spectroscopic data was reduced by means of a PLS-DA algorithm. The latent variables obtained with this method remained, thus reducing the dimensionality, and one of the multiclass classification models mentioned above was applied to the PLS-DA results. Finally, a cross validation with 10 partitions (K-folds=10) was performed on the training set and the predicted values were calculated using the test set, obtaining the model parameters for accuracy, precision and recall. The best results for both, whole and cut beans, were obtained with the combinations PLS-DA plus Random Forest and PLS-DA and XGBoost (see Figure 18 Figure 17), using as pretreatment the SNV method followed by DeTrend and a number of principal components equal to sixteen (PCN=16). Accuracies equal and higher to 0,9 were obtained with both methods.



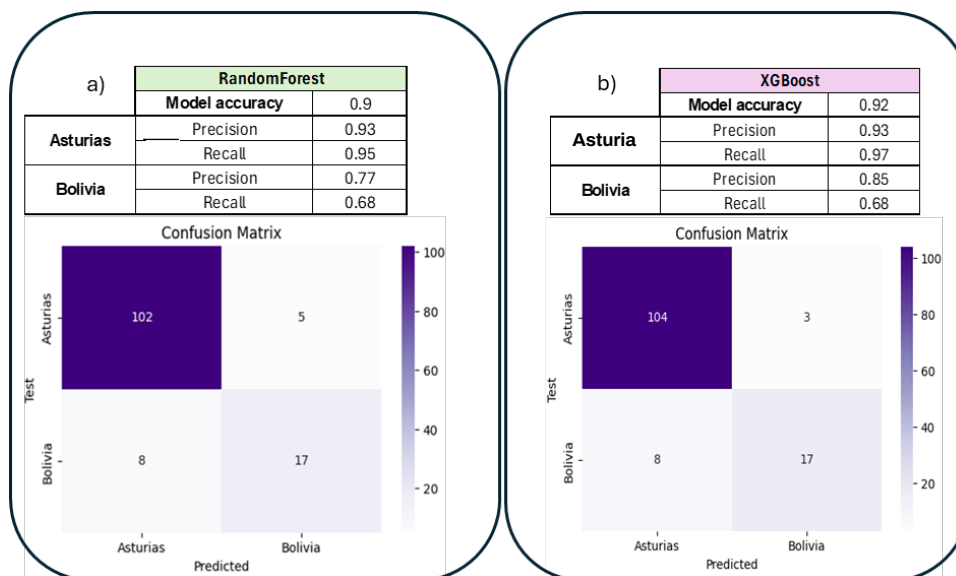


Figure 18 Accuracy, precision, recall and confusion matrix obtained for the models a) Random Forest, b) XGBoost.

When opting for the nonlinear SVM approach, an initial consideration was to employ PCA for dimensionality reduction. However, this decrease model accuracy, so PCA was discarded. Instead, the SNV method was utilized, due to its superior performance, coupled with hyperparameter tuning (C=100, gamma=0.1, kernel='poly') using GridSearchCV. Slightly better results were achieved with spectra from cut beans (Figure 17) compared to whole beans, getting an accuracy of 0.90 and 0.88 respectively.

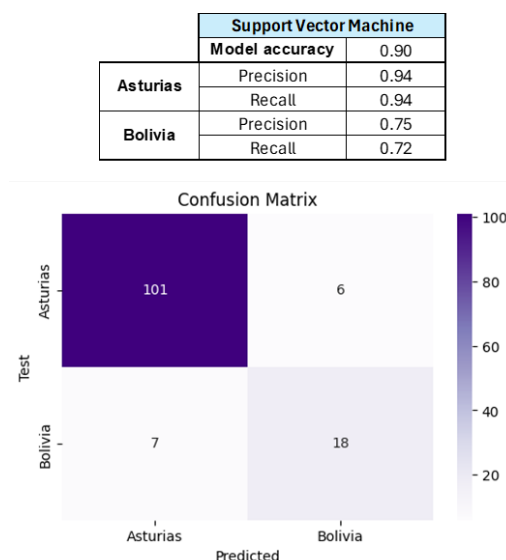


Figure 19 Accuracy, precision, recall and confusion matrix obtained for SVM model for cut beans.

Analysis of spectral data using HSI

The analysis of the spectra obtained by HSI, using the FX10 and FX17 cameras, is similar to that carried out with the spectra obtained by NIR. As explained before, the main difference



consists of extracting the spectrum of each bean pixel, performing the appropriate pretreatment (SNV, SG, MSC or DeTrend) and then obtaining an average spectrum for each bean (Figure 20).

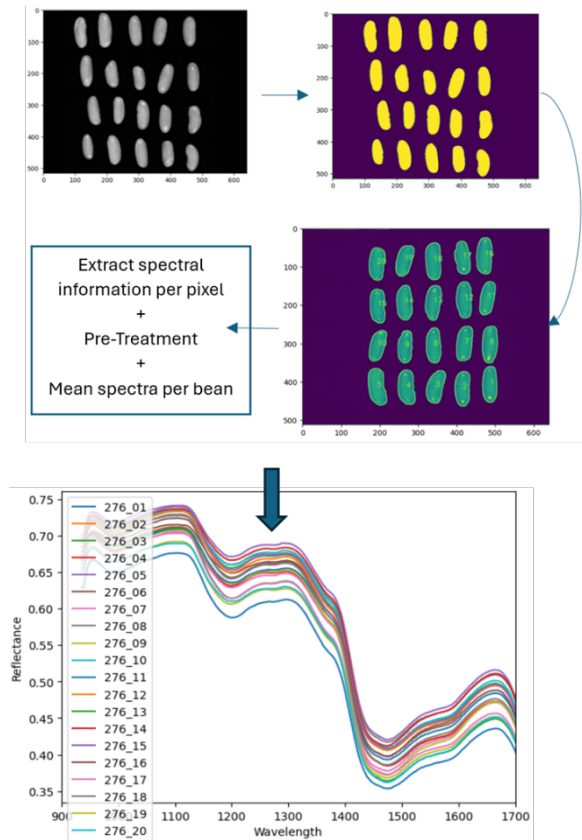


Figure 20 Spectra obtained by HSI with the FX17 camera for a sample of whole beans.

After obtaining an average spectrum for each bean, the data is split into training and test sets following an 80/20 ratio. Next, the PLS-DA method is applied in conjunction with the mentioned algorithms on the training set and a cross-validation with K-folds=10 is conducted on the same data. Predicted values are calculated from the test set.

For HSI, SNV method proves to be the optimal pretreatment across all cases, including both cameras (FX10 and FX17) and various faba bean formats. FX17 camera produces slightly superior results across most employed models but with minimal distinctions. Contrary to the case of portable NIR, the model giving the best results is PLS-DA together with SoftMax and SVM, resulting in accuracies above 0.95 for whole beans and reaching accuracy values equal to 1 for cut beans. In particular, the combination of PLS-DA and SoftMax with NPC=16 was the ideal model configuration, with superior precision and recovery values compared to the others (see Figure 21).



		Softmax	
		Model accuracy	1.00
Asturias	Precision	1.00	
	Recall	1.00	
Bolivia	Precision	1.00	
	Recall	1.00	

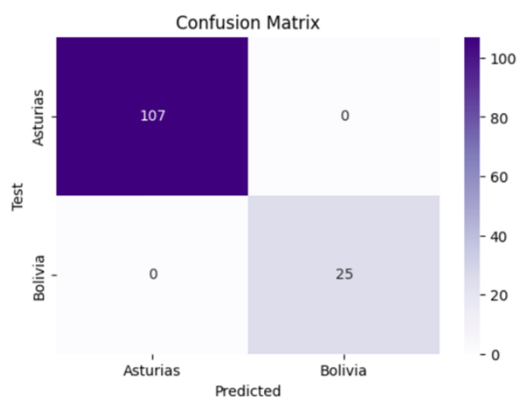


Figure 21 Accuracy, precision, recall and confusion matrix obtained for PLS-DA+XGB model for whole beans with FX17 camera.

The spectral data acquired by NIR and HSI for the milled beans are currently being analysed, so results are not yet available.

An evident conclusion of this initial data analysis is that portable NIR and in-line HSI technologies could play a crucial role in enhancing traceability and food authenticity, specifically in this case for the PGI Asturian faba bean. Moreover, these initial results also, provide valuable insights for determining the following actions points required to enhance current models, as detailed in Section 3.5.

3.3.4 Results visualization

Finally, an interface showing the result of the analysis, bean origin, to end-users was developed (Figure 22). The main idea is that this interface is easily usable by non-experts and, moreover, it minimizes at maximum the learning curve for being familiarize. Python was selected as the optimal language for this development.



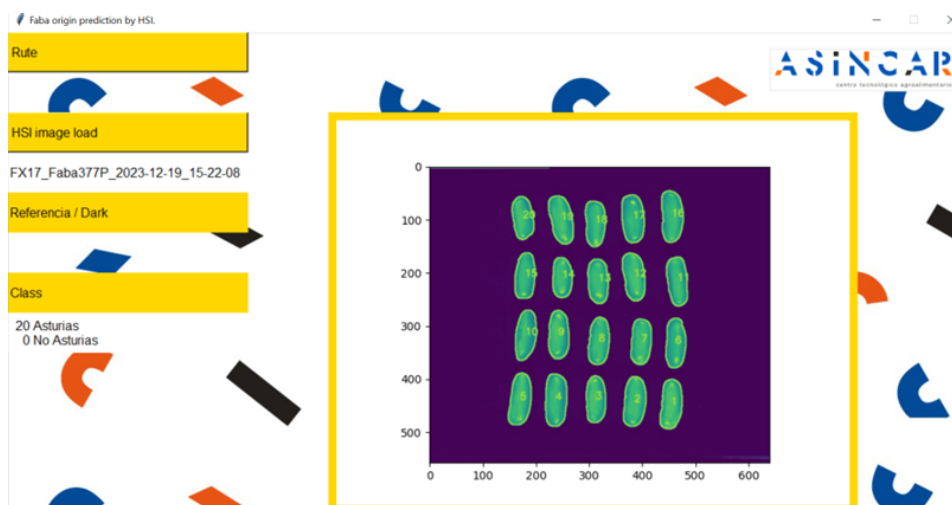


Figure 22 Example of a results display window for end-consumers.

The main reason of this choice is that Python packages for creating result presentation windows are easy to use and this language offer many additional possibilities. It should not be forgotten that Python is not primarily intended for data analysis but to other multitude of applications.

The developed program has the following characteristics and works as follows:

- The pre-treatments, wavelengths and algorithm parameters used in the development of the chemometric model are inside a Python class. When the program is executed, an object of this class is created and all of them can be used. This point is very important for the modularity of the solution. In the program it is enough to change the class to predict another parameter or to create objects of classes of several parameters at the same time to obtain predictions of all of them.
- The Python package used to create the results display window is TkInter, as it is an ideal package for such windowed applications.
- The internal functioning of the program (Figure 23) is as follows. First, the program creates an object of the class where pre-treatment, wavelengths and parameters are stored. Then, when the spectrometer generates the necessary spectra (in the case of the NIR on the faba bean, there are two, one on each side of the faba bean), the program collects all of them and creates an average spectrum representative of the sample. This average spectrum enters as input of the object initially created, the pre-treatments are executed, only the wavelengths are used if necessary, and the classification model is applied to make the prediction. Finally, the prediction is extracted from the object and displayed on screen.

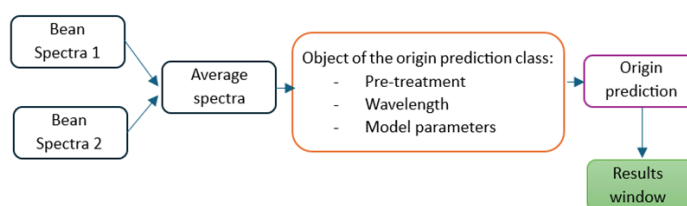


Figure 23 Internal functioning diagram of the results visualisation programme.

To use this programme for predicting the origin of the beans, following steps must be followed:

- Have the NIR spectrometer working, calibrated with the reference (Spectralon) and ready to take measurements.
- Run the faba bean origin prediction visualisation program.

- When the program is started, it is ready to take measurements. So, the two spectra are taken on both sides of the bean.
- The result of the origin prediction is automatically displayed on the screen for that sample.
- To take another measurement, press "Start". Then two new measurements are taken on opposite sides of the bean and the result is automatically displayed again.
- If an error is made during the measurement for whatever reason, just press the "Start" button to restart.
- The directory button is used to change the directory where the spectra used in the predictions are stored.
- To exit the program, just click on the cross in the upper left corner of the window.

3.4 Conclusions

Promising initial results of this study demonstrates that portable NIR as well as in-line HSI could be revolutionary tools for the identification of main fraudulent practices reported for the PGI Asturian faba bean (that could be extended to other PGI, PDO foods) by control agencies.

An initial prototype for the complete tool has been generated allowing to identify the origin of the faba bean and detecting in that way main detected fraud consisting in the mixture of PGI faba beans with cheaper ones coming from foreign origins. During this reporting period two main origins has been explored, Bolivia and Galicia, but additional samples from these locations are needed to improve the performance of the tool as well as its robustness. The complete development covers all main stages considered for the generation of a NIR-based application, meaning data acquisition, intelligent processing and result visualization.

NIR and HIS classification models showing accuracies equal or higher than 0,9 has been developed for this application, combining multiple pre-treatments, variable selection and classification methods, considering linear and non-linear models. In the case of portable NIR and HIS studies, PLS-DA paired with XGBoost provided the best results in terms of accuracy, but the non-linear SVM achieved very similar performances. When NIR and HSI developed models are compared, HSI provides slight better results.

The basis of the study, data collection, is a quite large and extensive characterization of the PGI faba beans as well the main adulterants used, faba from Bolivia and Galicia. The number of foreign samples needs to be increased in the next months to achieve comparable sizes to the one used for the PGI and is a work already ongoing (for example for Bolivia in 2023 campaign already 16 samples are collected). The physico-chemical characterization includes 3708 determinations and in the case of the spectral more than 8640. Also to highlight that multiple NIR, HIS devices and bean formats were considered in the experimental design.

Finally, it is important to highlight that during the whole itinerary the feedback of the main end-users was incorporated into the tool design thanks to the presence on the pilot of the two main control agencies, the one of the PGI (IGPFA) and the regional authority competent in food quality and safety (PDA).

3.5 Next steps

After completing this initial stage achieving considerable outcomes, it is imperative to outline the following actions aimed at broadening and fortifying the conducted research and developed



models. The following section delineates proposed measures to delve deeper into the findings and amplify the impact and significance of the results.

1. Augmenting sample size and addressing imbalance:

- Increase the sample size from Bolivia and Galicia to rectify the existing imbalance between Asturian and foreign bean data.
- Integrate data from the 2023 harvest samples to enrich the dataset further.
- Simultaneously collect additional foreign samples to enhance diversity and robustness in the dataset.

2. Model refinement and validation:

- Refine modeling techniques such as Random Forest and XGBoost by optimizing hyperparameters and feature selection based on the enriched dataset.
- Validate the updated models using rigorous testing methodologies to ensure reliability, accuracy, and generalizability of predictions.

3. Probability estimation integration:

- Incorporate methods for estimating response probabilities within the models to provide insights into prediction certainty or uncertainty.
- Explore techniques for probabilistic modelling to enhance predictive capabilities and decision-making processes based on calculated probabilities.

By implementing these next steps, the research efforts will advance significantly, leading to a more comprehensive understanding of bean data dynamics, improved predictive accuracy, and enhanced applicability of the study's outcomes in real-world scenarios.

Supplementary material can be found in the Annex.

4 DIGITAL KNOWLEDGE BASE FOR FOOD FRAUD MITIGATION

4.1 Introduction

Digital solutions have been developed with the goal of enhancing the integrity of the FSC and ensuring authenticity and quality in food products. The development of a Digital Knowledge Base for Food Fraud, as a result of ALLIANCE, marks as significant advancement in this domain. The project targets both regulatory bodies and other stakeholders in the food industry, equipping them with robust data-driven tools and analyses to streamline their decision processes and enhancing the understanding of potential food fraud risks. In addition to this, advanced data management strategies will be implemented to further enhance consumer safety and trust. The following sections outline the concept, development, current status, as well as future steps of this knowledge base, emphasizing its crucial role in the fight against food fraud.

4.2 Background

Food fraud is the deliberate adulteration, substitution or dilution of food, which leads to threats to consumer confidence, economic well-being and even public health threats to any food industry. Modern food supply chains face the risk of fraud due to complexity, combined with the large diversity of food products and ingredients. In response to these challenges, the need for an innovative, robust mechanism capable of identifying, assessing, and mitigating food fraud risks is evident. The proposed Digital Knowledge Base, therefore, comes as a guiding light of technological innovation meant to enable the enhancement of mechanisms for the discovery and prevention of food fraud.

4.3 Digital Knowledge Base Overview

The Digital Knowledge Base is conceptualized as an all-inclusive repository, well-designed with the assimilation of processed data, insights, and inferences derived from the analysis of food products along with their supply chains in an immaculate manner. The integration of external data (standards, certificates, PDO/PGI CoPs, scientific articles, links to related websites, etc.) with the data originating from the projects (Project results) makes it easy to take a thorough examination and extraction of valuable insights and reports by each product. Moreover, the knowledge base is bound together with the Vulnerability Risk Assessment Framework and the Early Warning and Decision Support System that systematically captures the relevant information for potential fraud or adulteration cases across Quality Labelled Food Supply Chains (QFS) and Food Supply Chains (FSCs). Furthermore, highlighting that this data source, as one of the Project results, will provide valuable information for the database, which in turn will offer critical insights, emphasizes its importance. This data will be seamlessly integrated within the database, enriching its capabilities. Additionally, it will significantly contribute to the functionalities of the dashboard, which is being developed as part of this task, enhancing its effectiveness in monitoring and decision-making. Perhaps search bar and menus would be an integral part of this system to afford a user the opportunity of gaining exact information

pertaining to the products and fraud associated with them along with possible prevention strategies.

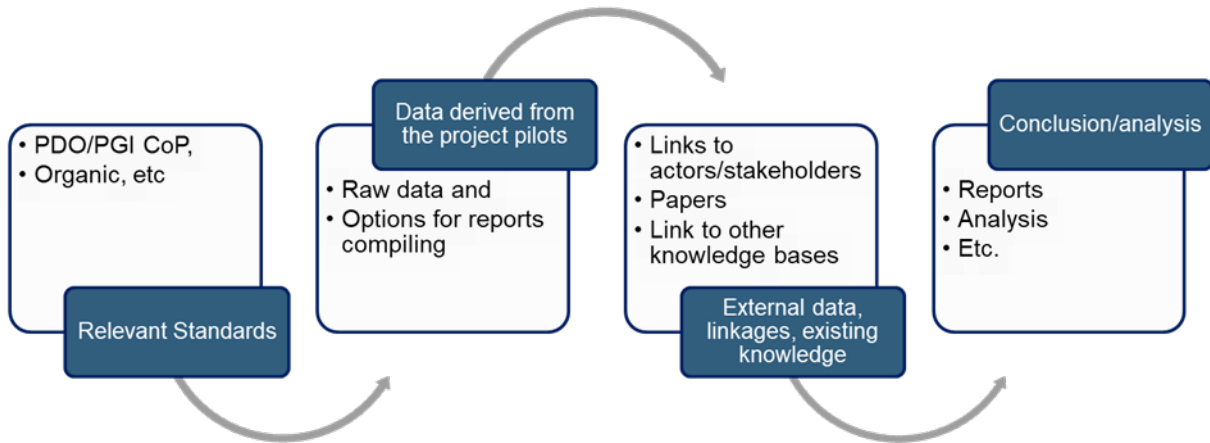


Figure 24 Visual overview: key components and features of a database.

4.4 Development of Database

The implementation of the Digital Knowledge Base is currently in the prototype stage. Meaning, the first considered implemented database solution for data management and data processing. It will consist of a number of layers in such a way that they work optimally under diversity of data types and high scalability of work. The proposed layers present the technologies and technical solutions with the functionalities to display, manage data, and process, and it is presented in the figure below (Figure 25).

Data Ingestion Layer: This layer houses all ingested data in its raw form as it arrives from various sources, and it automatically transforms these inputs using available tools prior to analysis. This process is transparent, allowing for ongoing evaluation and adjustments. *The storage layer* includes mechanisms for document, picture, and other file storage, facilitating effective data extraction and interrogation. Subsequently, the *Processing and Analysis Layer* undertakes responsibilities to prepare data for indexing and to enable powerful full-text search capabilities. Crucially, data are aggregated and analysed to allow for sophisticated statistical analysis and to generate visualizations that highlight trends and anomalies. *The Knowledge Representation and Management Layer* enhance this setup by supporting data representation in a knowledge graph, which increases the potential for more interconnected data points. Additionally, robust data security measures are implemented across all layers to ensure the confidentiality, integrity, and availability of the data.



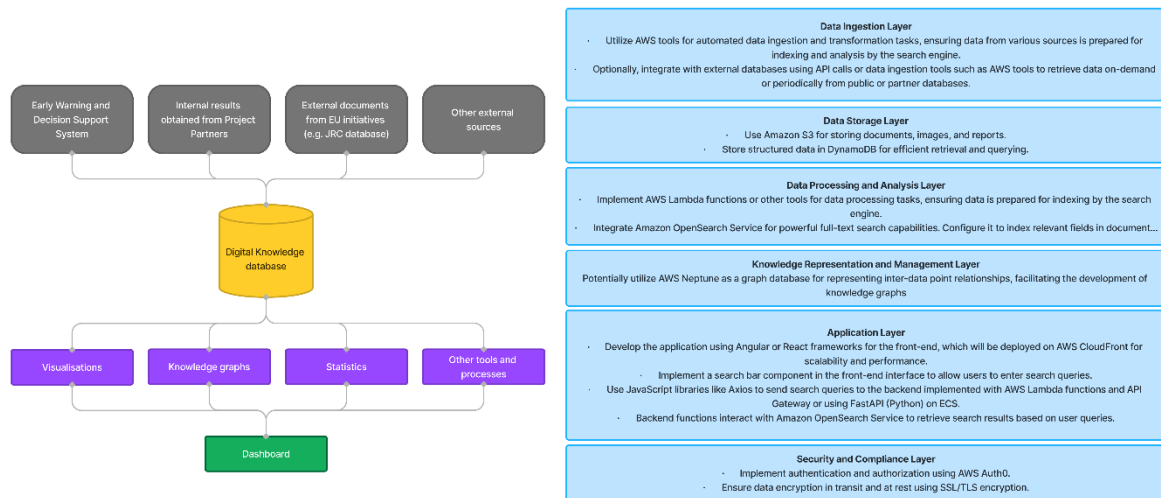


Figure 25 Digital knowledge base architecture

4.5 Future Directions and Improvements

As technology demands evolve and expand, addressing the scalability and upgradability of systems has become crucial in the strategic planning for knowledge database creation. While certain technologies are utilized in the initial stages of project development, the knowledge database remains flexible, open to modifications and enhancements in response to evolving technological needs and advancements. This could be through the adoption, in future developments, of technology with more or better abilities to assist in the processing of real-time data and the traceability needed. In sum, the program of the Digital Knowledge Base on Food Fraud represents a wider commitment to using technological innovation for food safety and integrity. This project is truly the guiding light to a food supply chain that, with collaborative and technological efforts to advance it, will surely promise higher security, transparency, and trustworthiness that will ultimately serve them as a barricade against food fraud threats, for consumers' well-being and confidence.



5 FOOD FRAUD PREVENTION WITH PREDICTIVE ANALYTICS

5.1 Introduction

In a time when safeguarding the integrity of our food supply chains is crucial, the application of predictive analytics emerges as a promising tool in the ongoing fight against food fraud. Section 5 delves into the complex landscape of food fraud prevention with predictive analytics, offering a comprehensive exploration of essential modules and components crucial for establishing an effective data value chain. Through the integration of cutting-edge technologies and advanced algorithms, such as deep neural networks, this section aims to empower stakeholders with the tools and insights necessary to effectively address the threat of food fraud. To elucidate the concepts and algorithms employed in the food fraud prevention system with predictive analytics, this section utilizes data collected by OLYMPOS. Hence, the figures presented herein are derived from data exploration and predictive modelling specific to the feta cheese use case.

Central to this section is the conceptual architecture of the predictive analytics module (subsection 5.2), which serves as the foundation for our exploration. By breaking down the fundamental components of this architecture, including data value chains (subsection 5.3) and predictive modelling techniques (subsection 5.4), we aim to explain the underlying principles guiding our approach. From data acquisition and exploration to the deployment of sophisticated algorithms, each subsection within this section is carefully crafted to provide a comprehensive understanding of the predictive analytics landscape in the context of food fraud prevention.

Moreover, this section delves into the vital role of explainable AI and rational decision-making processes (subsection 5.6) in ensuring the trustworthiness and transparency of predictive analytics solutions. Through insightful visualizations and intuitive explanations (subsection 5.5), stakeholders are equipped with the knowledge and tools to interpret model outputs and make informed decisions. Finally, as we look towards the future, the section outlines potential next steps, guiding stakeholders on the path towards continuous improvement and innovation in the realm of food fraud prevention with predictive analytics (subsection 5.7).

5.2 Conceptual Architecture of the Predictive Analytics Module

This section aims at presenting the conceptual architecture of the food fraud prevention system with predictive analytics (see Figure 26).



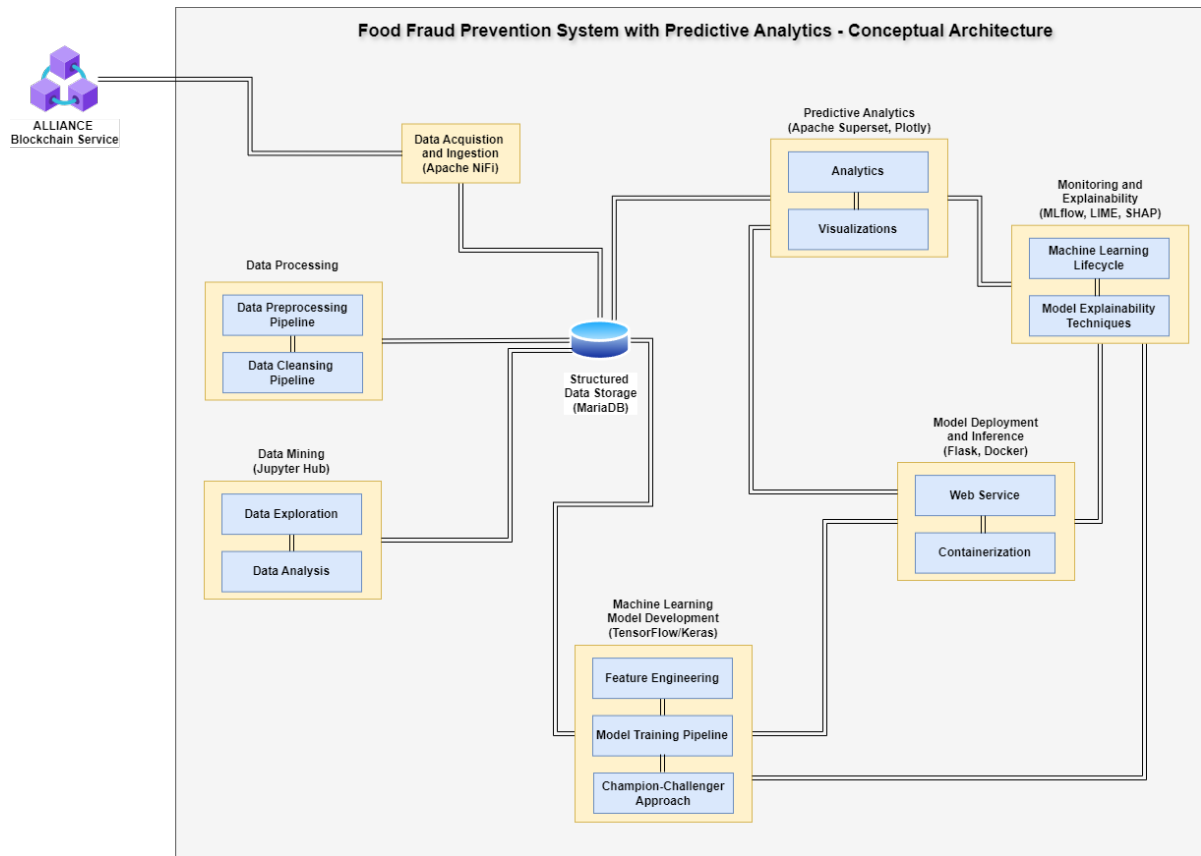


Figure 26 Conceptual architecture of the food fraud prevention system with predictive analytics.

The conceptual architecture provides a framework for building a scalable and robust predictive analytics system for detecting food fraud using deep learning models. Each component plays a critical role in the end-to-end data pipeline, from data acquisition and preprocessing to model development, deployment, and monitoring. Additionally, incorporating open-source tools like Apache NiFi, MariaDB, TensorFlow, MLflow, and Apache Superset ensures flexibility, cost-effectiveness, and community support.

5.2.1 Data Acquisition and Ingestion

Apache NiFi⁴² is used for data ingestion from the UTH blockchain. NiFi provides a user-friendly interface for designing data flows and handling various data formats and protocols. The ingested data are stored in a MariaDB⁴³ database for efficient retrieval and management.

5.2.2 Data Processing

The data processing module implements data preprocessing pipelines to clean and transform the raw data. This may involve handling missing values, outlier detection, normalization, and feature engineering.

⁴² <https://nifi.apache.org/>

⁴³ <https://mariadb.org/>



5.2.3 Data Mining

Jupyter Notebooks are utilized for exploratory data analysis (EDA)⁴⁴ and model prototyping. Specifically, Jupyter Hub⁴⁵ provides an interactive environment for data exploration, visualization, and experimentation with different machine learning algorithms. Pandas, NumPy, Matplotlib, and Seaborn Python libraries are essential for data manipulation, numerical computations, and visualization during the EDA phase.

5.2.4 Machine Learning Model Development

TensorFlow and Keras Python libraries⁴⁶ are used for building deep learning models. These libraries offer high-level APIs for constructing and training neural networks. Particularly, with this component we design a pipeline for model training and evaluation. This pipeline includes data preprocessing, model training, hyperparameter tuning, and model evaluation stages. Furthermore, we are going to implement a mechanism to compare multiple models in a champion-challenger mode.

5.2.5 Model Deployment and Inference

A web service using Flask⁴⁷ will be implemented for deploying trained models. APIs will be exposed for real-time inference and allowing integration of the deployed models with other systems or applications. In addition to this, Docker⁴⁸ will be used for containerizing model deployment environments, ensuring thus consistency and portability across different deployment environments.

5.2.6 Predictive Analytics

Apache Superset⁴⁹ is suggested for interactive data visualization and analytics. Superset provides a rich set of visualization options and allows users to create custom dashboards for monitoring and analysis. Moreover, other Python libraries (e.g., Dash or Plotly) can be used to create interactive dashboards and visualizations for displaying model performance metrics, data insights, and predictions.

5.2.7 Monitoring and Explainability

We utilize MLflow⁵⁰ for managing the machine learning lifecycle, including experiment tracking, model packaging, and deployment. MLflow offers built-in capabilities for model monitoring, versioning, and explainability. Approaches such as SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-agnostic Explanations) will be explored to interpret and explain model predictions. These techniques will increase the trustworthiness and transparency of the deployed models.

⁴⁴ Liao, S.-H., Chu, P.-H., & Hsiao, P.-Y. (2012). Data mining techniques and applications - A decade review from 2000 to 2011. *Expert Systems with Applications* 39(12), 11303-11311.

⁴⁵ <https://jupyter.org/hub>

⁴⁶ <https://www.tensorflow.org/guide/keras>

⁴⁷ <https://flask.palletsprojects.com/en/3.0.x/>

⁴⁸ <https://www.docker.com/>

⁴⁹ <https://superset.apache.org/>

⁵⁰ <https://mlflow.org/>



5.3 Data Value Chain

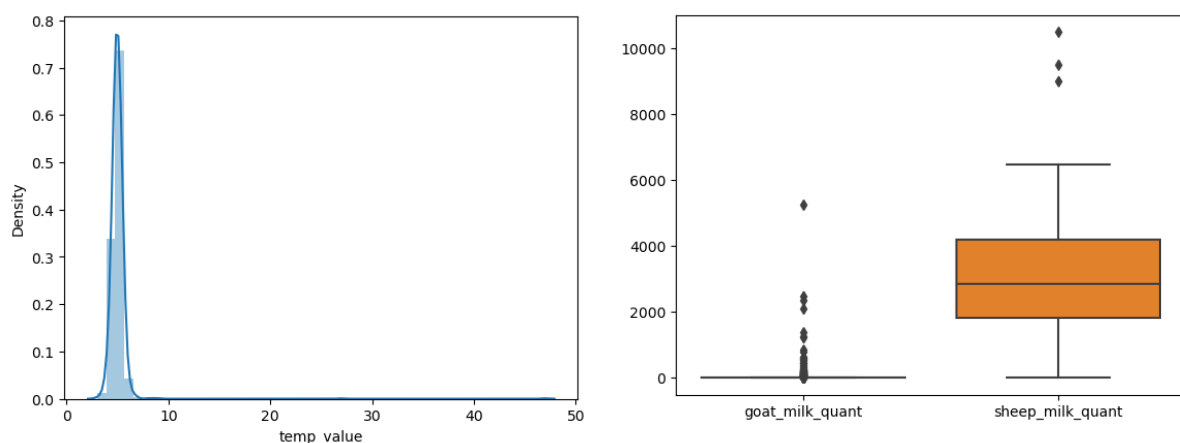
In the complex ecosystem of food fraud detection, the data value chain emerges as a foundational pillar, guiding the journey from raw data to actionable insights.⁵¹ This chain represents the process of collecting, integrating, cleansing, and analysing data, resulting in the development of robust predictive models. Understanding and optimizing each link in this chain is crucial, as it directly impacts the efficacy and reliability of fraud detection efforts. In this context, the data value chain acts as the core support system, enabling stakeholders to utilize state-of-the-art technologies and advanced analytics to address the challenge of food fraud. This sub-section briefly describes the main phases of the data value chain.

5.3.1 Data Collection and Integration

Datasets are collected from various sources across the food supply chain, including supplier information, transaction records, product specifications, and historical fraud incidents. Emphasis is placed on obtaining high-quality, relevant data to ensure the efficacy of predictive analytics models. Furthermore, collected datasets undergo integration to consolidate disparate sources into a unified format suitable for analysis.

5.3.2 Data Cleaning and Preprocessing

Data cleansing processes are applied to identify and rectify anomalies, handle missing values, and standardize data formats (see Figure 27). Preprocessing steps such as normalization, feature scaling, and encoding categorical variables are performed to prepare the data for model training.



⁵¹ Faroukhi, A. Z., Alaoui, I. E., Gahi, Y., & Amine, A. (2020). Big data monetization throughout big data value chain: A comprehensive review. *Journal of Big Data* 7(3), 619-634.



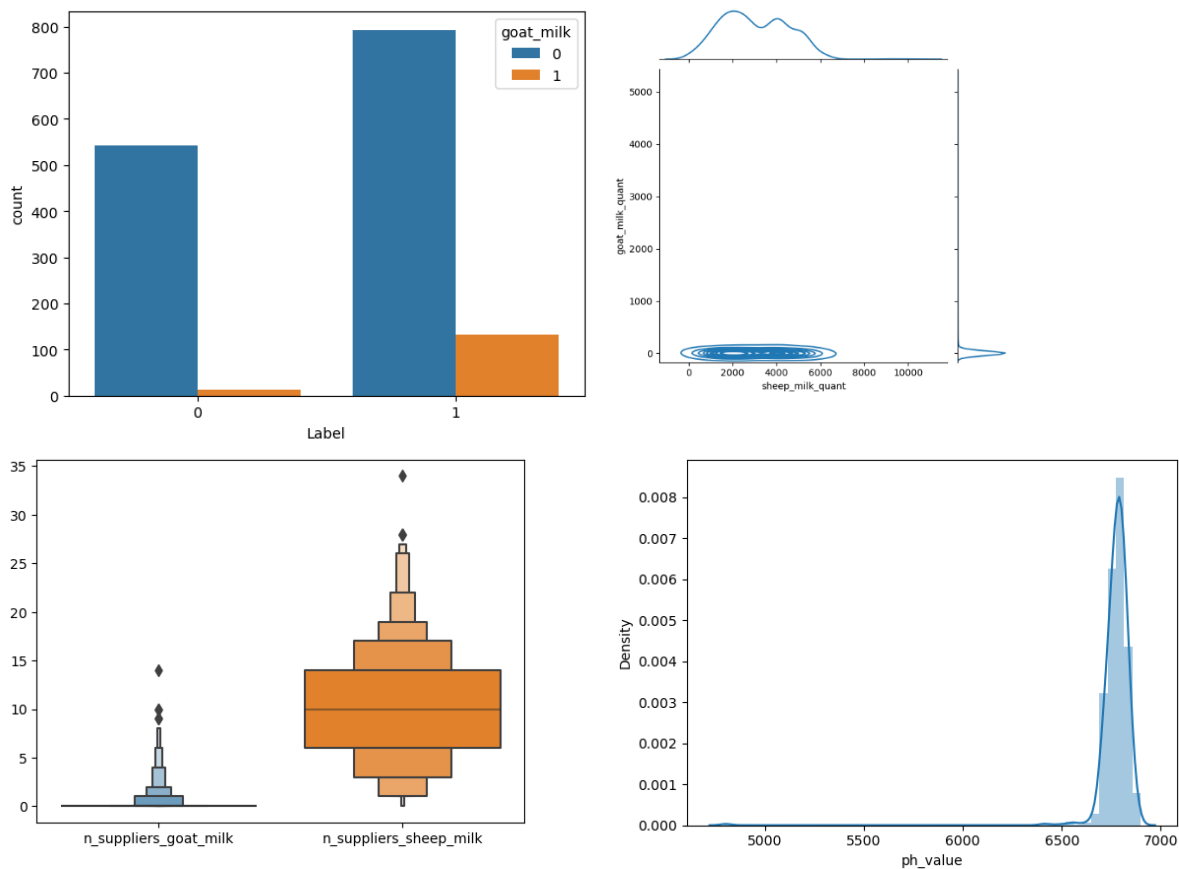


Figure 27 Exploratory data analysis for an illustrative example on feta cheese use case.

5.3.3 Feature Engineering

Feature engineering techniques are employed to extract meaningful insights from raw data and enhance model performance. Domain-specific knowledge is leveraged to create relevant features that capture key characteristics of the food supply chain and fraud vulnerabilities. Python tools and libraries will support the feature engineering process for the machine learning model development. Pandas library provides data structures like DataFrame which are excellent for handling structured data, including numerical, categorical and datetime variables, whereas NumPy supports arrays, which are essential for numerical operations on available data. Furthermore, scikit-learn provides a wide range of tools for preprocessing data, including methods for scaling, encoding categorical variables, handling missing values, etc. In addition to this, libraries like Feature-engine and Featuretools will be explored for handling common feature engineering tasks such as imputation, encoding, discretization, and more.



5.3.4 Model Training Data Preparation

The prepared data is partitioned into training, validation, and test sets for model development and evaluation. Techniques such as stratified sampling ensure representative distribution of data across different subsets to mitigate bias.⁵²

5.3.5 Model Evaluation

Models are evaluated using performance metrics such as accuracy, Area Under Curve (AUC) score, F1 score, precision, and recall assessing their effectiveness in detecting fraud vulnerabilities. Cross-validation techniques validate model robustness and generalization capabilities across diverse datasets.

5.3.6 Model Deployment

Deployed models integrate seamlessly into operational workflows, whereas real-time inference capabilities enable timely detection and mitigation of fraud risks within the food supply chain.

5.3.7 Continuous Improvement

Continuous monitoring of model performance and feedback loops inform iterative improvements to the predictive analytics framework. In this direction, we are using MLflow, an open-source platform, which focuses on the full lifecycle of the machine learning projects, ensuring that each phase is manageable, traceable, and reproducible (see Figure 28). Stakeholder engagement and collaboration facilitate ongoing refinement of data collection strategies, model architectures, and deployment practices.

By analysing each stage of the data value chain, stakeholders can identify opportunities for optimization and enhancement, ensuring the seamless integration of predictive analytics into food fraud detection efforts.

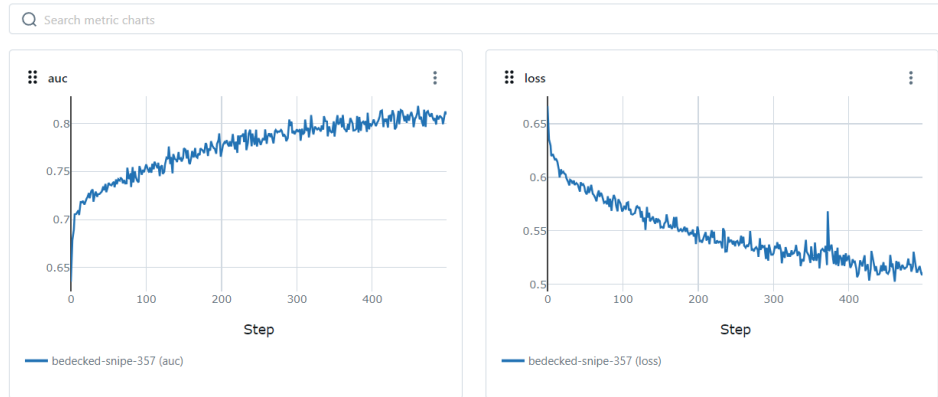
⁵² Corliss, D. J. (2024). Designing against bias: Identifying and mitigating bias in machine learning and AI. In: K. Arai (eds.) Intelligent systems and applications. IntelliSys 2023. Lecture Notes In Networks and Systems, vol. 824, Springer.



Alliance Food Fraud Detection Experiment >

bedecked-snipe-357

Overview **Model metrics** System metrics Artifacts

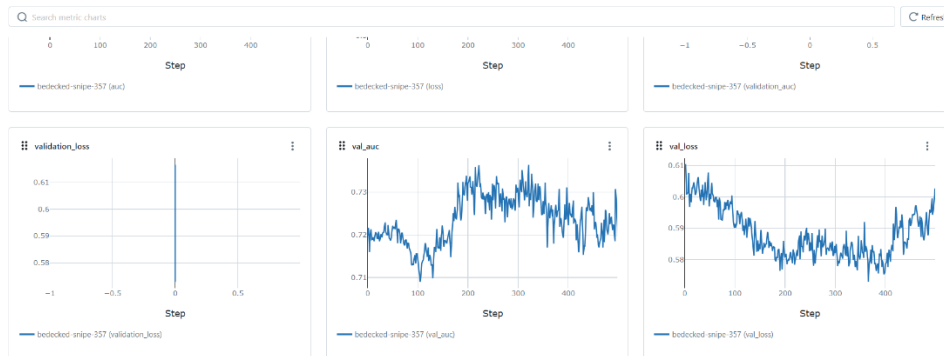


Alliance Food Fraud Detection Experiment >

bedecked-snipe-357

Register model

Overview **Model metrics** System metrics Artifacts



Alliance Food Fraud Detection Experiment >

bedecked-snipe-357

Register model

Overview **Model metrics** System metrics Artifacts

Search parameters

Parameter	Value
batch_size	30
class_weight	None
epochs	500
initial_epoch	0
max_queue_size	10
opt_emgrnd	False
opt_beta_1	0.9
opt_beta_2	0.999
opt_clipnorm	None
opt_clipvalue	None
opt_ema_momentum	0.99
opt_ema_overwrite_frequency	None
opt_epsilon	1e-07

Search metrics

Metric	Value
auc	0.809728608551025
loss	0.5115990042686462
validation_auc	0.719457030293237
validation_loss	0.602698564529419
val_auc	0.7184570302963257
val_loss	0.602698564529419

Figure 28 Managing the data-driven solution lifecycle through MLflow.



5.4 Food Fraud Detection Modelling

Food fraud poses a significant threat to the integrity of our food supply chains, compromising consumer safety and trust.⁵³ Predictive analytics, especially leveraging deep learning algorithms, emerges as a promising solution for both identifying and mitigating these risks.⁵⁴ In this subsection, we delve into the process of deploying predictive analytics for the detection of vulnerability risks related to food fraud. Our approach employs a champion-challenger framework, wherein multiple deep learning algorithms compete during the training phase to select the most effective model for the inference phase. By considering a range of performance metrics such as accuracy, AUC, F1 score, precision, and recall, we aim to optimize the detection process. This comprehensive approach extends from acquiring and cleaning data to training models, deploying them, and ensuring explainability, all while harnessing state-of-the-art technologies to achieve optimal effectiveness.

5.4.1 Data Acquisition and Cleansing

High-quality, clean data is pivotal for effective predictive analytics. Our process begins with the acquisition of diverse datasets encompassing various aspects of the food supply chain, including supplier information, transaction records, product specifications, and historical fraud incidents. Next, thorough data cleansing is essential to uphold the integrity of our analyses. This involves detecting and rectifying anomalies, handling missing values, standardizing formats, and addressing data inconsistencies. Advanced data cleansing tools and techniques, including outlier detection algorithms and automated validation processes, streamline this crucial step, enhancing the reliability and accuracy of our predictive models.

5.4.2 Model Training in the Champion-Challenger Framework

Our approach adopts a champion-challenger framework, wherein multiple deep learning algorithms⁵⁵ compete for superiority during the training phase. Leveraging libraries such as TensorFlow and Keras, we deploy an array of neural network architectures, including convolutional neural networks (CNNs), recurrent neural networks (RNNs), and transformer models, among others. During training, each model undergoes iterative optimization processes, competing against one another to maximize performance across designated metrics. We employ techniques such as hyperparameter tuning, ensemble learning, and cross-validation to enhance model robustness and generalization capabilities.

It is worth noting that the training data undergoes preprocessing steps such as normalization, feature scaling, and encoding categorical variables.⁵⁶ Techniques like feature engineering may be employed to extract relevant information and enhance model performance. As far as model parameters are concerned, these are initialized, and optimization algorithms such as stochastic gradient descent (SGD) and Adam are employed to iteratively update these parameters. Therefore, hyperparameter tuning techniques, including grid search or random search, are utilized to explore the hyperparameter space and identify optimal configurations. Furthermore,

⁵³ Manning, L. (2021). Food supply chain fraud: The economic, environmental, and sociopolitical consequences. *Advances In Food Security and Sustainability* 3, 253-276.

⁵⁴ Vinothkanna, A., Dar, O. I., Liu, Z., & Jia, A.-Q. (2024). Advanced detection tools in food fraud: A systematic review for holistic and rational detection method based on research and patents. *Food Chemistry* 446, 138893.

⁵⁵ Tayyab, M., Marjani, M., Jhanjhi, J., Hashem, I. A. T., Usmani, R. S. A., & Qamar, F. (2023). A comprehensive review on deep learning algorithms: security and privacy Issues. *Computers and Security* 131, 103297.

⁵⁶ Dinov, I. D. (2023). *Data science and predictive analytics*. Springer Nature Switzerland AG.





model architectures are customized based on the nature of the input data and the complexity of the fraud detection task.

Regarding the training and validation parts of the process, it should be mentioned that the training dataset is partitioned into training and validation sets, with the latter used for monitoring model performance during training (see Figure 29). As a result, models are trained iteratively on the training data, with performance evaluated on the validation set at regular intervals. Technique such as early stopping may be employed to prevent overfitting and ensure optimal model generalization.

To mitigate overfitting and improves the robustness of the final predictive model, ensemble learning⁵⁷ techniques, such as bagging or boosting, may be utilized to combine predictions from multiple base models, whereas K-fold cross-validation or stratified cross-validation techniques are employed to assess model performance across different subsets of the data. This provides a more reliable estimate of model performance and generalization capabilities. Performance metrics including accuracy, AUC, F1 score, precision, and recall are computed on both the training and validation sets. More specifically, model performance is analysed across different metrics to identify the most effective model for deployment.

5.4.3 Deployment and Explainability

Upon identifying the champion model through rigorous evaluation, we proceed to deploy it within the operational environment for real-time fraud detection. Furthermore, achieving model explainability is paramount for fostering trust and comprehension among stakeholders. We employ techniques such as SHAP (SHapley Additive exPlanations) values, LIME (Local Interpretable Model-agnostic Explanations), and feature importance analysis to elucidate the underlying factors driving model predictions.⁵⁸ This facilitates transparent decision-making and enables stakeholders to interpret and act upon the insights gleaned from the predictive analytics framework.

It is worth mentioning that containerization technologies like Docker are employed to package the predictive analytics application and its dependencies into lightweight, portable containers. The predictive analytics model is exposed as a RESTful API, allowing seamless integration with existing systems and applications. API endpoints facilitate real-time inference, enabling stakeholders to query the model and receive predictions promptly.

⁵⁷ Zhou, X., He, J., & Yang, C. (2022). An ensemble learning method based on deep neural network and group decision making. *Knowledge-based Systems* 239, 107801.

⁵⁸ Nagahisarchoghaei, M., Nur, N., Cummins, L., Nur, N., Karimi, M. M., Nandanwar, S., Bhattacharyya, S., & Rahimi, S. (2023). An empirical survey on explainable AI technologies: Recent trends, use-cases, and categories from technical and application perspectives. *Electronics*12(5), 1092.



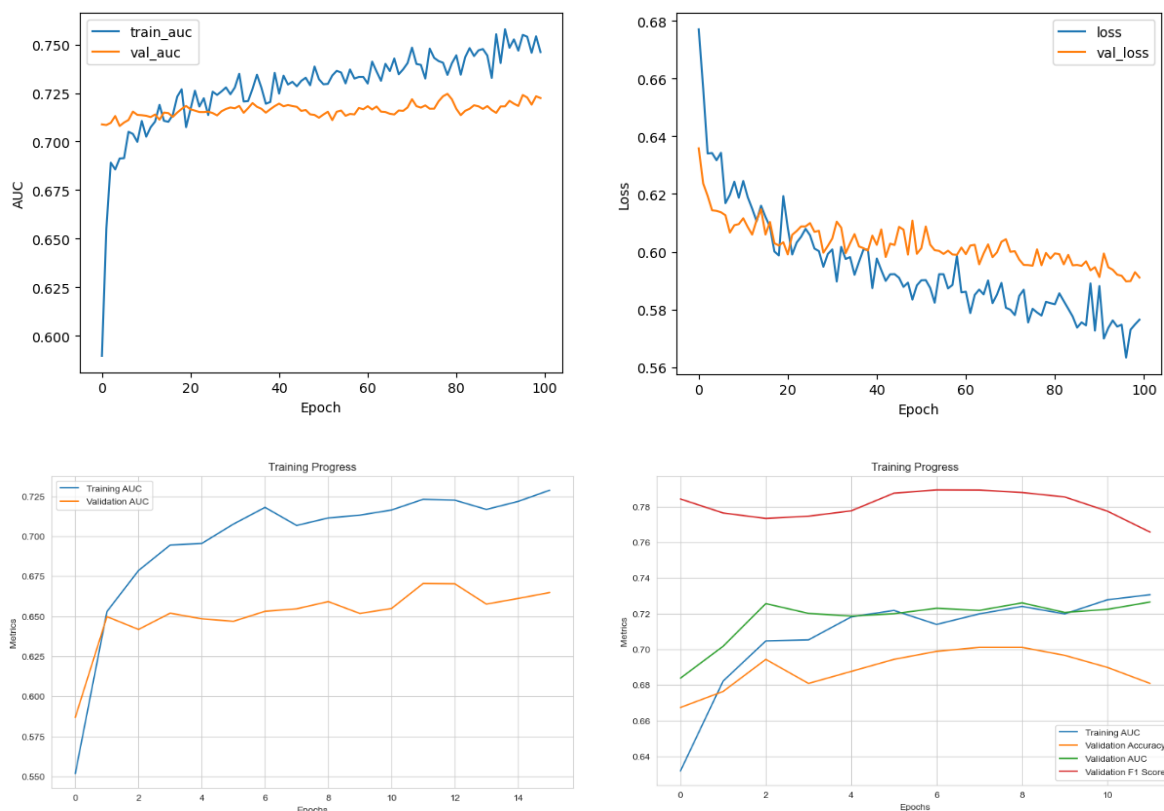


Figure 29 Training pipeline with respect to multiple performance measures.

In conclusion, predictive analytics powered by deep learning holds significant potential for detecting vulnerability risks associated with food fraud. By employing a champion-challenger framework and considering multiple performance metrics, we can identify optimal models for real-world deployment. Through thorough data acquisition, cleansing, and advanced model training techniques, we enhance the efficacy and reliability of our predictive analytics solution. Moreover, by prioritizing model explainability and leveraging cutting-edge technologies, we ensure transparency and facilitate informed decision-making within the food supply chain ecosystem. As we continue to innovate and refine our methodologies, predictive analytics will remain a cornerstone in safeguarding consumer safety and trust in the face of evolving food fraud challenges.

5.5 Visualisations and predictive Analytics

The insights obtained from the statistical analysis of each use case's data, along with additional information inferred from explainable AI models, are presented comprehensively to end users and stakeholders. This presentation is facilitated through interactive dashboards developed in Apache Superset, an open-source data visualization platform.

Apache Superset empowers end users to access valuable insights into the unique characteristics of each use case by interacting with concise dashboards tailored specifically for this purpose. The process of developing such dashboards involves several indicative steps:





- **Define objectives:** Through continuous communication with the involved stakeholders, the main objectives and KPIs are agreed upon, which create the most value in each business case.
- **Select visualizations:** The best visualization types, that most illustratively represent the data and satisfy the defined objectives are selected.
- **Explore the data:** Using Superset's integrated features, as well as custom scripts, the available data are explored and the subsets that can best support the chosen dashboards and visuals are extracted.
- **Create database views:** To reduce computing load on the tool's side, custom database views are created, which already contain pre-calculated values that will be used in the visuals.
- **Develop dashboards:** Superset is connected to the backend database using built-in connectors, the data of interest are retrieved, and the dashboards are created and optimized.
- **Publish online:** Once the dashboards are ready, they can be published through the tool, so that all interested parties can easily access them through their web browser.

In the context of ALLIANCE goals and objectives, some options of relevant dashboards and visualizations that can be developed are:

- Enable time series analysis for (i) plotting historical trends of fraud incidents over time, and (ii) overlaying predictive model forecasts to illustrate future trends and potential risks.
- Create risk heatmaps to visualize the distribution of food fraud risks and represent the severity or likelihood of fraud occurrence.
- Enable feature importance analysis to identify the most influential factors contributing to food fraud prediction.
- Show a predictive model performance dashboard to monitor the performance metrics of the predictive model in real-time.
- Run a cluster analysis to group similar food products or suppliers based on their characteristics or risk profiles.
- Run geospatial visualization to map the distribution of food fraud incidents geographically.

It is noteworthy that predictive analytics, visualizations, and explainable AI (see subsection 5.6) come together to form dashboards that go beyond conventional data representation, with the goal of equipping users with multifaceted capabilities. These dashboards are designed to facilitate the following:

1. The ability to perceive rich and complex information: By presenting data in visually engaging formats, users can grasp hidden patterns and relationships within datasets in a more intuitive manner.
2. The ability to discover causal information from observational data: Through interactive exploration, users can uncover causal relationships and identify factors that influence outcomes, enabling deeper insights into underlying mechanisms.
3. The ability to learn in a particular context: Dashboards contextualize data within specific domains or business contexts, facilitating continuous learning and adaptation to evolving circumstances.
4. The ability to abstract: Visual representations abstract complex datasets into understandable forms, enabling users to extract high-level insights without getting lost in the details.



5. The ability to create new meanings/concepts: By combining various data sources and illustrating where they overlap, dashboards inspire creativity and innovation, encouraging the emergence of new ideas and concepts.
6. The ability to reason for decision-making: Dashboards support informed decision-making by providing evidence-based reasoning and enabling users to evaluate alternative courses of action based on data-driven insights.
7. The ability to explain prediction/decision outcomes: Explainable AI techniques embedded within dashboards clarify the reasoning behind predictions and decisions, improving transparency and fostering trust in the analytics process.

Essentially, these dashboards serve as dynamic platforms that enable users to perceive, explore, learn, abstract, create, reason, and explain, thereby unlocking the full potential of predictive analytics and visualization in decision-making.

5.6 Concepts towards Explainable AI and rational decision-making

By analysing the training process in distinct phases, we can systematically optimize model performance and identify the most robust and reliable predictive model for detecting vulnerability risks related to food fraud. Each phase plays a crucial role in enhancing model efficacy and generalization capabilities, ultimately contributing to the success of the predictive analytics framework. However, a recurrent concern about AI that is based on deep learning algorithms is that they operate as “black boxes” without explaining their computational results in a language close to a human expert.⁵⁹

In the domain of AI-driven decision-making, the primary focus is typically on achieving accuracy. However, in the pursuit of precision, the significance of explainability is frequently overlooked. Although AI has become a key enabling technology for the sciences and industry, many of these AI systems are not able to explain their autonomous decisions and actions to human users.⁶⁰ The role of explanations in data quality in the context of data-driven machine learning models is very crucial especially when it is needed to address the issues of transparency as linked to AI.⁶¹⁻⁶² Recently, some ideas and insights have been provided by the research community as far as the Explainable AI is concerned, focusing not only on the AI models, but also on the data, that support the model development.⁶³ The focus is placed on the definition, identification, and explanation of errors in data and the appropriate repair actions. Moreover, a set of processes have been introduced to mitigate and manage general classes of bias in AI algorithms.⁶⁴ These biases are generally divided into those associated with the mapping of the business intent into the AI algorithm, those that arise due to the distribution of training samples, and those related to individual input samples. Therefore, understanding the

⁵⁹ Loyola-Conzalez, O. (2019). Black-Box vs. White-Box: Understanding their advantages and weaknesses from a practical point of view. *IEEE Access*, 7, 154096-154113.

⁶⁰ Arrieta, A.-B., Diaz-Rodriguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82-115.

⁶¹ Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267, 1-38.

⁶² Verheij, B. (2020). Artificial intelligence as law. *Artificial Intelligence and Law*, 28, 181-206.

⁶³ Bertossi, L., & Geerts, F. (2020). Data quality and explainable AI. *ACM Journal of Data Management and Information Quality*, 12(2), 1-9.

⁶⁴ Roselli, D., Matthews, J., & Talagala, N. (2019). Managing bias in AI. In L. Liu, & R. White (Eds.), *WWW '19: Companion Proceedings of the 2019 World Wide Web Conference* (pp.539-544). San Francisco, USA.



reason why a decision has been made by a machine is crucial to grant trust to a human decision-maker.⁶⁵

The goal is to explain the algorithmic decisions of AI solutions with non-technical terms to make these decisions trusted and easily understandable by humans. Finally, it is widely recognized that Explainable AI should also deal with the question of “How to evaluate the quality of explanations?” given by an explainable AI system. The System Causability Scale (SCS), as well as the Interpretable Confidence Measures (ICMs) have been introduced in the literature to measure the quality of explanations.⁶⁶⁻⁶⁷

Considering issues described above, SHAP (SHapley Additive exPlanations) values are employed to quantify the contribution of each feature to the model's predictions. By analysing SHAP values, stakeholders can understand the relative importance of different features and their impact on the model's decisions (see Figure 30). Moreover, LIME (Local Interpretable Model-agnostic Explanations) approach is suggested. Typically, LIME generates locally faithful explanations for individual predictions, providing insights into how the model arrived at a particular decision (see Figure 31). It highlights the most influential features for a specific prediction, enhancing interpretability and trust in the model.

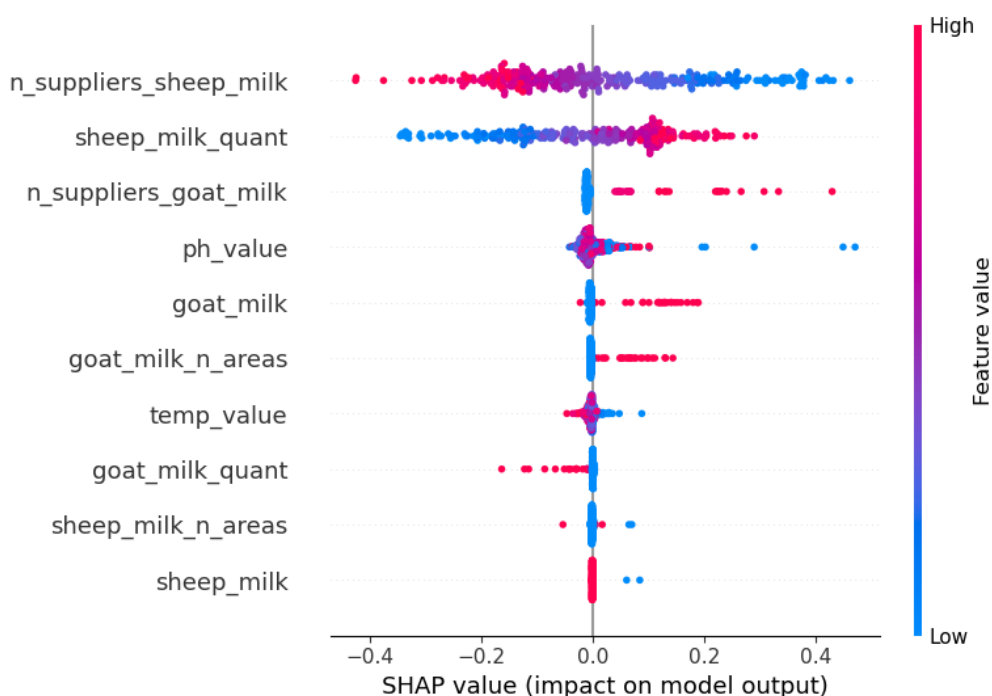


Figure 30 SHAP values for an illustrative example on feta cheese use case.

⁶⁵ Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G.-Z. (2019). XAI – Explainable artificial intelligence. *Science Robotics*, 4(37), 1-2.

⁶⁶ Holzinger, A., Carrington, A., & Muller, H. (2020). Measuring the quality of explanations: The system causability scale (SCS). *KI*, 34, 193-198.

⁶⁷ Van der Waa, J., Schoonderwoerd, T., van Diggelen, J., & Neerincx, M. (2020). Interpretable confidence measures for decision support systems. *International Journal of Human-Computer Studies*, <https://doi.org/10.1016/j.ijhcs.2020.102493>.



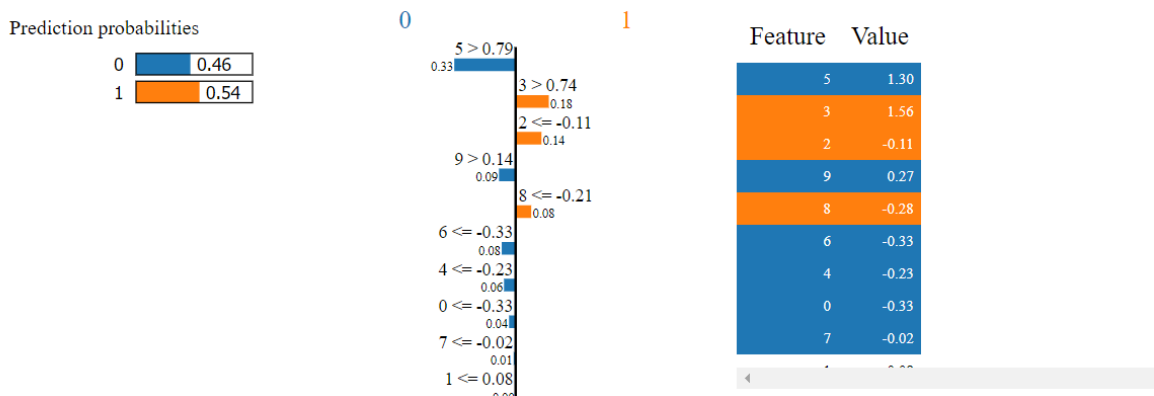


Figure 31 LIME approach for an illustrative example on feta cheese use case.

Additionally, feature importance techniques such as permutation importance or feature contribution plots are utilized to identify the most influential features across the entire dataset. Stakeholders can use these insights to gain a holistic understanding of the factors driving the model's predictions and make informed decisions accordingly.

By focusing on robust deployment practices and ensuring model explainability, stakeholders can confidently integrate predictive analytics solutions into their operations, leveraging actionable insights to mitigate vulnerability risks related to food fraud effectively. In this direction, interactive visualization tools and dashboards are developed to present model explanations in a user-friendly manner. These interfaces will allow stakeholders to explore model predictions, understand the underlying rationale, and gain actionable insights without requiring expertise in machine learning (see Figure 32).

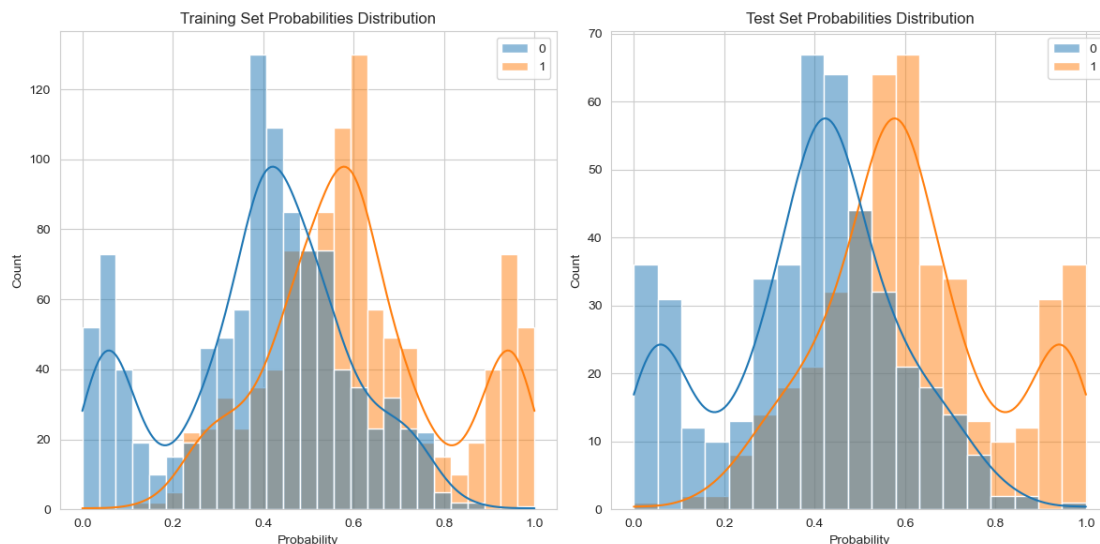


Figure 32 Probabilities (predictions) distributions for an illustrative example on feta cheese use case.

5.7 Next Steps

These steps encompass a strategic sequence, beginning with the acquisition and ingestion of diverse datasets from various facets of the food supply chain. Over the next months, the focus will shift towards processing and mining this data, leveraging advanced techniques to extract actionable insights indicative of potential fraud incidents. At the same time, we'll focus on





creating machine learning models customized for detecting food fraud, aiming to enhance efficacy and accuracy in identifying fraudulent activities.

As we move forward, the deployment of predictive analytics frameworks will play a pivotal role in operationalizing these models for real-time fraud detection. This phase will be characterized by continuous refinement and adaptation, ensuring that our predictive analytics strategies remain responsive to evolving fraud patterns. Finally, the integration of the developed predictive analytics system into existing food supply chain infrastructure (i.e., current ALLIANCE components) will solidify our efforts, facilitating seamless collaboration and data exchange among stakeholders (e.g., integration with assets developed by UTH and FINS in WP2 and WP3 respectively). Through these tasks, we aim to realize the envisioned outcomes of the ALLIANCE project, safeguarding the integrity and resilience of the food supply chain against fraudulent activities.



6 CONSUMER DEMAND ASSESSMENT AND STRENGTHENING

6.1 Introduction

The increasing concern about food fraud and the growing consumer demand for high-quality and safe food has become one of the most important issues in the current food framework. In an increasingly complex food system, it is necessary to strengthen the traceability system in order to increase and ensure the transparency of the food chain. Traceability often encounters problems because the information along the chain is not secure and can be manipulated or incorrect from one step to the next. The result is an insecure system where fraud can creep in, leading to health risks for consumers and mistrust. The consequences of food fraud and adulteration are manifold and affect different levels of society, e.g. consumer confidence in the supply chain, consumer health, economic and social impact⁶⁸. In this context, blockchain technology is proving to be a promising solution for improving the traceability of the food supply chain. By integrating blockchain technology, it is possible to create an immutable and transparent register of all the transfers and steps that a product must go through to reach the consumer. Blockchain helps to ensure that all parties involved in the production process can track and trace the product, ensuring detailed and secure information. Task 3.6 of the ALLIANCE project (which is the subject of this section) aims to monitor and evaluate consumer opinion on products produced under the innovative traceability system. It also aims to identify the main socio-cultural aspects that determine consumer choice in favour of these products. To this end, a public acceptance survey will be conducted with a representative sample of 500 citizens per participating country.

Consumers and stakeholders can significantly benefit significantly from the solutions proposed in this part of the project. By using blockchain technology to ensure transparency and traceability in the food supply chain, end consumers can be assured of the safety and quality of their products. This increased transparency allows them to make more informed decisions about the products they buy, contributing to greater consumer confidence. Transparent supply chains promote consumer confidence as they have access to detailed information about the origin, production processes and journey of the products they buy. This transparency promotes a sense of responsibility among those involved and reassures consumers that the food they are consuming has integrity and authenticity. In addition, blockchain-enabled traceability systems allow the end consumer to access comprehensive product information, including details on origin, production methods, certifications and quality ratings. Overall, the adoption of blockchain technology and transparent supply chain practises will benefit end consumers by promoting safety, quality, transparency, trust and sustainability in the agri-food sector.

The results of this study are important for both policy makers and producers in the agri-food sector. For policy makers, the results suggest that blockchain technology should be seen as an important tool to combat food fraud and ensure food safety and quality. Policies and regulations may be needed to promote the adoption and implementation of blockchain technology in the food supply chain. A transparent supply chain promotes better collaboration and communication. By sharing clear and well-documented information, all actors in the supply

⁶⁸ Bannor, R. K., Arthur, K. E., Oppong, D., & Oppong-Kyeremeh, H. (2023). A comprehensive systematic review and bibliometric analysis of food fraud from a global perspective. *Journal of Agriculture and Food Research*, 14, 100686. <https://doi.org/10.1016/j.jafr.2023.100686>



chain can gain a better overview of the intermediate processes in the supply chain. This information sharing can lead to greater operational efficiency and reduce lead times and operating costs. In addition, a transparent supply chain can facilitate product traceability, allowing companies to respond quickly to food safety or quality issues. Overall, supply chain transparency can create a more trusting and collaborative environment within the agri-food sector, promoting long-term sustainability, accountability and competitiveness. In addition, producers can utilise insights into consumer perceptions and trust in blockchain technology to develop marketing and communication strategies aimed at promoting their products tracked with this technology.

6.2 Literature review

Several studies have investigated the intention of consumers to buy products that are tracked on the blockchain. The 2021 study by Lin et al.⁶⁹ on consumers' intention to use blockchain technology for food traceability in organic food explores the factors that influence consumers' intentions regarding the adoption of Blockchain Food Traceability System (BFTS) technology, a blockchain-based food traceability system, to ensure food safety. The research proposed an integrated conceptual framework, that combines two consolidated theoretical models: the Theory of Planned Behaviour (TPB) and the Information Success (ISS) models. The authors argue that it is crucial that the BFTS is reliable and of high quality, as the decisions based on it are associated with risks for consumers. Therefore, this study suggests the importance of utilising initial trust to strengthen customer purchase intent. The analysis was conducted using a questionnaire with 300 Chinese consumers, with a particular focus on analysing organic food. The results show that Chinese consumers who have a positive attitude towards BFTS technology and believe that they can effectively control their behaviour when using this technology are more likely to adopt it. Subjective norms, i.e. the opinions and expectations of others, were positively but not significantly related to the intention to use BFTS. This suggests that social norms do not have a significant influence on the acceptance of this technology by Chinese consumers in the specific context of organic food.

The study by Dang & Tran (2020)⁷⁰, aims to investigate consumer intentions towards traceable food in the context of the animal disease epidemic and current food safety issues. The study utilises the TPB model, which was developed to predict attitudes and purchase intentions towards traceable pork. Other constructs include perceived risk, concern for food safety, perceived importance of a healthy diet and environmental impact. The extended TPB model was used to predict purchase intention/attitude towards traceable pork among 230 students in Vietnam. It was found that conscious consumers have a positive attitude towards traceable pork as they are concerned about their own health and the environment and believe it contributes to their commitment to a healthy and environmentally friendly lifestyle.

⁶⁹ Lin, X., Chang, S. -, Chou, T. -, Chen, S. -, & Ruangkanjanases, A. (2021). Consumers' intention to adopt blockchain food traceability technology towards organic food products. *International Journal of Environmental Research and Public Health*, 18(3), 1-19. doi:10.3390/ijerph18030912

⁷⁰ Dang, T. K., & Anh, T. D. (2020). A pragmatic blockchain based solution for managing provenance and characteristics in the open data context. In *Future Data and Security Engineering: 7th International Conference, FDSE 2020, Quy Nhon, Vietnam, November 25–27, 2020, Proceedings 7* (pp. 221-242). Springer International Publishing.



In the work of Dionysis et al. (2022)⁷¹, the factors influencing consumer purchase intentions for coffee with blockchain traceability were analysed using the TPB model. The original TPB model was extended to include additional constructs such as trust, past habits and environmental protection. The study was conducted using an online questionnaire to 123 participants, comparing two organic coffee traceability systems: one based on blockchain and the other based on traditional certification. It was found that participants who trusted that they could find and understand information about coffee without help showed a greater commitment to coffee with blockchain traceability. In addition, those who believed that coffee can be traced back to its true origin and that coffee with blockchain traceability is more likely to be authentic also showed a positive association with this type of coffee.

Prisco et al. (2022)⁷² investigate the factors influencing the adoption of blockchain in Italian companies: the moderating role of company size. The article presents an integrated approach that combining the TAM (Technology Acceptance Model) and the TPB and adding as benefits the additional factors “efficiency and security”, “reduced costs” and “quality of customer service” perceived by companies adopting blockchain technology. The results show that attitude and perceived behavioural control are the most significant predictors of intention to adopt blockchain, while perception of benefits is the most significant predictor of attitude. In addition, subjective norms were found to positively influence behavioural intention, while the effect of perceived ease of use on attitude was insignificant.

6.3 Theoretical Framework

Based on the analysis of previous literature, the TPB,⁷³ one of the most commonly used models in the literature for predicting intentions and behaviours, was chosen as the conceptual model in this study. The TPB is based on the idea that a person's behaviour depends on the intention to perform it. Behavioural intention is the result of the interaction of three factors:

1. Attitude (ATT): represents a person's inclination to perform a certain action. It is the opinion or judgement a person has about adopting or performing a particular behaviour, based on their values, beliefs and previous experience of that behaviour. A positive attitude leads to a greater likelihood of behaving in a way that is consistent with one's intention. Attitude towards a particular technology should be measured in terms of trust, i.e. the actor's tendency to trust the target behaviour.
2. Subjective norms (SN): refer to the influence of other people's thoughts and attitudes towards a particular behaviour. In other words, it is the social pressure to perform or avoid a certain action, which may result from the expectations, encouragement or opinions of others. Subjective norms therefore reflect the social weight that individuals perceive towards a particular behaviour and can influence their behavioural intentions and decisions.

⁷¹ Dionysis, S., Chesney, T., & McAuley, D. (2022). Examining the influential factors of consumer purchase intentions for blockchain traceable coffee using the theory of planned behaviour. *British Food Journal*, 124(12), 4304-4322. doi:10.1108/BFJ-05-2021-0541

⁷² Prisco, A., Abdallah, Y., Morandé, S., & Gheith, M. H. (2022). Factors affecting blockchain adoption in Italian companies: the moderating role of firm size. *Technology Analysis & Strategic Management*, 1–14. <https://doi.org/10.1080/09537325.2022.2155511>

⁷³ Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, 50(2), 179–211. [https://doi.org/10.1016/0749-5978\(91\)90020](https://doi.org/10.1016/0749-5978(91)90020)



3. Perceived behavioural control (CCP): refers to the perception of a person's ability to perform an action or the perception of the difficulty or ease of a particular behaviour depending on certain factors.

The combination of the three components results in behavioural intention, which is a predictor of actual behaviour.

However, this study aims to improve the predictive power of the TPB. In addition to the original elements of the TPB, such as attitude, subjective norms and perceived behavioural control, additional constructs will be introduced.

UNIBO plans to include also other constructs, such as consumer's buying habits, consumer's trust towards in organic/PDO-PGI labels, trust in technology and sociodemographic variables. This process led to different specifications of the same model and the original TPB.

6.4 Method

6.4.1 Sample Design and Data Collection

A public acceptance survey will be conducted among a representative sample of 500 citizens per participating country (Croatia, Greece, France, Italy, Serbia, Spain). Respondents will be recruited by a recruitment agency. The study in each country will look at the case study product of that country.

A pilot study was carried out in Italy to test the questionnaire. The dissemination period ran from 30 October 2023 to 28 February 2024. A total of 251 responses were collected in total, of which 201 were successfully completed, indicating a high participation rate.

6.4.2 Questionnaire

The questionnaire was developed on the Qualtrics platform to explore consumers' views on products produced through an innovative traceability system. This questionnaire was distributed online via popular social networking platforms (such as Whatsapp, Instagram and Facebook).

The draft questionnaire, which consists of a total of 50 questions, was divided into 7 different blocks to simplify the survey process and allow a targeted investigation of the different aspects of consumer opinions and preferences. This methodological approach allows valuable information to be gathered on consumer attitudes towards traceable products and the underlying socio-cultural influences that guide their choices.

The draft questionnaire is divided into following sections:

1. General overview of the study and guarantee of confidentiality of the information provided by the participants.
2. Knowledge of the topic under study.
3. TPB constructs: Intentions, attitude, subjective norms, perceived behavioural control. Each question was created with consideration of research from authors such as Dang & Tran (2020), Dionysis et al (2022), and Menozzi et al (2015) to ensure full coverage of key variables in the context of the study. Participants' responses will be analysed using a carefully structured rating scale that allowed participants to clearly express their agreement or disagreement with the statements contained in the questionnaire. This scale, like the

Likert scale, offers a range of options from "strongly agree" to "strongly disagree". This approach allows participants to provide detailed and nuanced responses, ensuring an accurate assessment of their opinions and perceptions.

4. Consumption habits. In this part of the questionnaire, participants' purchasing habits are collected in terms of place of purchase, frequency, preferences for product attributes, taking into account previous research such as Bandinelli et al (2023),⁷⁴ Dionysis et al (2022), and Menozzi et al. (2015). A flexible response method was used to collect the participants' answers.
5. Behaviour towards surveyed product. This question concerns respondents' attitudes towards the product under investigation by asking participants to rate the percentage of organic/PDO product they buy compared to their total conventional product purchases. This approach allows to assess the relative importance of this attribute in their purchasing decisions.
6. Consumer trust in the organic/PDO system and technology. This block includes the evaluation of trust in producers and sellers of organic/PDO products, taking into account previous studies conducted by Li et al. (2023),⁷⁵ as well as trust in technology.
7. Socio-demographic questions. Finally, demographic information about the participants is collected, including age, gender, education level, income and other variables relevant to the study.

6.4.3 Data Analysis

After developing the model and collecting data, we will analyse the correlation and assess the reliability of the data. To measure the reliability of the questionnaire, which is intended as a measurement tool, the Cronbach's alpha test will be used. To calculate the Cronbach's alpha coefficient, we will analyse the correlations between the different questions of the questionnaire within each block. The Cronbach's alpha coefficient ranges from 0 to 1, with higher values indicating greater internal consistency. In general, an alpha value above 0.70 is considered ideal, while values above 0.80 guarantee greater reliability.

According to the pilot study, the results obtained show a good correlation between the items and a satisfactory overall coherence, which confirms the reliability of the questionnaire.

To develop our econometric model, we will group all questions within each construct and use the average value obtained. Therefore, for each individual respondent, we will be a single value for each block that corresponds to the average value for the answers to the questions asked in each block. The only exception will be for the block 4, "buying habits", as some questions follow a different rating scale.

A multiple linear regression model will be used for the analysis, which attempts to model the dependent variable using several independent variables.

⁷⁴ Bandinelli, R., Scozzafava, G., Bindi, B., & Fani, V. (2023). Blockchain and consumer behaviour: Results of a Technology Acceptance Model in the ancient wheat sector. *Cleaner Logistics and Supply Chain*, 8, 100117. <https://doi.org/10.1016/j.clscn.2023.100117>

⁷⁵ Li, Y., Liao, A., Li, L., Zhang, M., Zhao, X., & Ye, F. (2023). Reinforcing or weakening? The role of blockchain technology in the link between consumer trust and organic food adoption. *Journal of Business Research*, 164, 113999.



6.5 Assessment of consumer perception and behaviour

UNIBO shared the questionnaire with all partners in order to collect their comments. The questionnaire will then be translated into the language in which the survey will be conducted and adapted to the individual case study with the help of the pilot partners.

We have already proceeded with the request for quotes from the recruitment agencies.

With reference to the pilot study, we have preliminary results which constitute for us expected results for the study we will conduct.

The results of the pilot study confirm the importance of trust in the technology as a significant determinant of consumers' willingness to buy traceable pasta via blockchain technology, which is consistent with the conclusions of Lin et al. (2021). However, the results on attitudes show no significant influence on purchase intention, which is consistent with the research by Dang & Tran (2020) and Prisco et al (2022). In contrast, perceived behavioural control proves to be a relevant predictor, which is consistent with the studies by Lin et al, (2021); Dang & Tran, (2020), and Dionysis et al, (2022) and Prisco et al, (2022).

Although subjective norms are not robust predictors in some studies, they appear to be significant in this context in all three models analysed. This suggests that the opinions and thoughts of others, along with social pressure and individual perceptions of ability to perform a certain behaviour, play a fundamental role in consumers' intention to purchase noodles tracked via blockchain technology.

Regarding habits, contrary to the finding of Dionysis et al., habits related to seeking information about the origin and production processes were found to significantly and positively influence consumers' purchase intentions. This result is in line with the theory of Verplanken & Aarts (1999)⁷⁶, which states that past behaviour can influence future intentions. This conclusion differs from the analysis of Dionysis et al. (2022), in which habits of seeking information about origin and production processes were not unpredictable and did not influence consumers' purchasing decisions.

6.6 Conclusion and policy implication

Today, the agricultural and food sector is more complex and interconnected than ever. However, this complexity also creates new problems for authorities seeking to protect human health and for consumers demanding safe and high-quality food. To ensure the safety and quality of food, it is essential to address these challenges through targeted regulations, policies, and controls. In this context, blockchain technology offers itself as a promising solution to improve traceability in the food supply chain. Blockchain provides an immutable and transparent record of all the transactions and steps a product has to go through to reach the end consumer. This can significantly help to prevent food fraud and ensure food quality and safety.

The pilot confirms how important individual perception and trust in blockchain technology is for consumer purchase intent. It is evident that subjective norms, perceived behavioural control and trust in blockchain technology play an important role in shaping consumer purchasing

⁷⁶ Verplanken, B., & Aarts, H. H. (1999). Habit, attitude, and planned behaviour: Is habit an empty construct or an interesting case of goal-directed automaticity? *European Review of Social Psychology*, 10(1), 101–134. <https://doi.org/10.1080/14792779943000035>





behaviour. While some purchasing habits may positively influence the intention to purchase products tracked with blockchain technology, other variables may not have the same influence.

The results of this study are relevant for both policy makers and producers in the agri-food sector. For policy makers, the results suggest that blockchain technology should be seen as an important tool to combat food fraud and ensure food safety and quality. Policies and regulations may be needed to promote the adoption and implementation of blockchain technology in the food supply chain.

A transparent supply chain promotes better collaboration and communication. By sharing clear and well-documented information, all actors in the supply chain can gain a better overview of the intermediate processes in the supply chain. This information sharing can lead to greater operational efficiency and reduce lead times and operating costs. In addition, a transparent supply chain can facilitate product traceability, allowing companies to respond quickly to food safety or quality issues. Overall, supply chain transparency can create a more trusting and collaborative environment within the agri-food sector, promoting long-term sustainability, accountability and competitiveness.

In addition, producers can utilise insights into consumer perceptions and trust in blockchain technology to develop marketing and communication strategies aimed at promoting their products tracked with this technology.

It is important to keep in mind that the study is based on a specific empirical analysis on the purchase of pasta tracked with blockchain technology. Therefore, although the results are transferable to similar contexts in the agri-food sector, they may not be directly transferable to other sectors or product types. Further research is required to confirm and extend the findings and explore other potential applications of blockchain technology in the agri-food sector and beyond. Another limitation that this analysis may have concerns the sample used in the study, which may not be representative of the general population. The present study may raise new research questions or highlight areas of enquiry that have not yet been fully explored. Despite the growing interest in the use of blockchain technology for product traceability, there may be practical or technical limitations to the effective implementation of this technology in the food sector. These limitations may include high costs, limited interoperability or privacy and security concerns.



7 CONCLUSIONS AND FUTURE DIRECTIONS

7.1 Recap of Achievements

Although the main achievements of T3.2-T3.6 reported up to M18 have been presented in detail in sections 2-7, a brief summary is also provided in this section. The experimental design of portable NIR and HSI technologies for food fraud detection, revealed four novel genetic molecular markers (two of which are the most prominent for further analysis) which were tested across the six olive varieties. The first results from the usage of the portable DNA kit for the EVOO authentication process are promising as they could potentially provide valuable insights into the application of DNA fingerprinting technology. The development of a fast, non-destructive and robust methodology using portable NIR and HSI technologies and ML to authenticate the geographical origin of beans, as geographical origin significantly influences quality and price gave promising results. The fundamental elements of the ALLIANCE Digital Knowledge Base have formed the basis of its design and development. Establishing the framework for building a scalable and robust predictive analytics system for detecting food fraud and developing deep learning models has been successfully conducted, along with a first consumer demand assessment analysis and detailed selection of the methodologies that will be followed in the future.

7.2 Addressing Project Objectives

The ALLIANCE success is ensured by the achievement of its key objectives as described in section 1.1 of the Grant Agreement. The current deliverable (and its final version) with the information it contains is considered as the means of verification for objectives 1, 2, 4, 5, and 6.

7.3 Future Directions

The work that has been presented in this deliverable is ongoing and all participants in T3.2-T3.6 has been putting efforts on the realisation of the tasks objectives as described in the last subsection of each section. The final solution and results of the ALLIANCE novel tools that use advanced portable qPCR DNA sequencing, portable NIR and HSI (offline) Spectroscopy, leverage AI and predictive analytics to detect adulteration and consumer study will be reported in D3.3 "Final AI-enabled tools & Digital Knowledge Database for Detecting Food Fraud using novel portable rapid testing for on-site inspection", which is to be submitted in M30.



8 ANNEX

The morphological assessment consisted of scrutinising the physical structure and shape of the beans by measuring the length, width and thickness of the beans using a caliper. The colour analysis aimed to quantify the visual appearance of the whole beans, using a Konica Minolta CM-26dG colourimeter, which reports colour in L*, a*, and b* values, where L* represents lightness from black to white (0-100), a* indicates green (-127) to red (+127), and b* signifies blue (-127) to yellow (+127). In addition, the measurement of the weight of the whole beans provided information on their density and mass distribution (Fig. 6). Finally, the evaluation of the water absorption capacity was crucial to understand the efficiency with which each bean sample can absorb moisture, which is an important factor in various processing and quality control applications, by weighting approximately 16 grams of beans per sample and immersing them in 100 ml of distilled water, collecting bean weight data every hour until 6 hours into the experiment and measuring the weight again after 24 hours.

The moisture of the bean flour was assessed using the oven-drying method, in which samples are heated under controlled conditions to calculate the moisture content as a function of weight loss. The colour of the bean flour was determined using the colourimeter by taking five measurements of each well homogenised sample and then calculating an average value.

In the other hand, the specifications of the HSI cameras used are as follows:

Specim FX-10:

- Spectral range (VNIR): 400-1000 nm.
- Spectral resolution: 5.5 nm.
- Signal-to-noise ratio (maximum): 600:1.
- Number of spectral bands: 224.
- Spatial resolution: 1024 pixels.
- Frame rate: 330 FPS (full spectral range).

Specim FX-17:

- Spectral range (NIR): 900-1700 nm.
- Spectral resolution: 8 nm.
- Signal-to-noise ratio (maximum): 1000:1.
- Number of spectral bands: 224.
- Spatial resolution: 640 pixels.
- Frame rate: 670 FPS (full spectral range).

The HSI image acquisition configuration consists of the following items:

- Hyperspectral camera (FX-10 or FX-17).
- Scanner with moving tray, element that provides movement to the sample. In addition, the scanner includes two illumination units with three halogen lamps in line, that are placed on both sides of the hyperspectral camera.
- Reference for hyperspectral images (Specim White Calibration Tile).
- Computer.
- Frame Grabber: CL Xtium MX4. This board is used to collect data from the hyperspectral cameras.
- LUMO Software Scanner: Software that manages the acquisition of hyperspectral images.