ALLIANCE

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

DELIVERABLE 3.3 - FINAL AI-ENABLED TOOLS & DIGITAL KNOWLEDGE DATABASE FOR DETECTING FOOD FRAUD USING NOVEL PORTABLE RAPID TESTING FOR ON-SITE INSPECTION

GRANT AGREEMENT NUMBER: 101084188

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: 13.05.2025 (Month 30)

Version: 1.0

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	Ready for internal review	INTRA
0.4	Comments received by IRs	ASINCAR, and UNIZG
0.5	Comments from IRs addressed, and deliverable sent to the PC	BIOC, ASINCAR, FINS, INTRA, and UNIBO
1.0	Final QA'ed version and submission to the EC Portal	UTH

Authors

Author	Partner
Amalia Ntemou	Netcompany-Intrasoft (INTRA)
Athanasia-Maria Dourou, Sofia Tzagkaraki, Stylianos Arhondakis	BioCoS (BIOC)
Noemí Quintanal, 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)
Alessandra Castellini, Giulia Maesano, Seyyedehsara Sadrmousavigargari	Alma Mater Studiorum -University of Bologna (UNIBO)

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

List of figures

List of tables

List of Abbreviations

Abbreviation	Description
EVOO	Extra Virgin Olive Oil
qPCR	Quantitative Polymerase Chain Reaction
PCR	Polymerase Chain Reaction
HRM	High Resolution Melting Analysis
ML	Machine Learning
PCA	Principal Component Analysis
PDO	Protected Designation of Origin
PGI	Protected Geographical Indication
AI	Artificial Intelligence
WP	Work Package
NIR	Near-infrared
HSI	Hyperspectral imaging
SAM	Spectral Angle Mapping (SAM)
SHAP	SHapley Additive exPlanations

Executive Summary

The submission of Deliverable 3.3 marks a significant milestone in ALLIANCE, which is related to the final development, integration and validation of the ALLIANCE AI-enabled tools and the digital knowledge base. These solutions have been designed in order to enable the detection, monitoring, and prevention of food fraud across various food supply chains.

More specifically, the ALIANCE systemic innovations include advanced portable technologies, qPCR DNA sequencing, Near-Infrared (NIR) and Hyperspectral Imaging (HSI) spectroscopy, as well as AI-driven predictive analytics and a digital knowledge base.

A comprehensive overview of the advancements achieved, from the experimental design until the validation and final implementation, for each of the ALLIANCE tools, are described in this report. Key results from the analysis on the use of portable qPCR devices to identify the provenance and verify the authenticity of organic PDO/PGI EVOO, with a focus on Italian varieties and the validation of additional biomarkers are also presented. The validation of NIR and HSI models showcased their performance in detecting fraudulent mixtures in PGI Asturian faba beans. In addition, preliminary results of a complementary model based on the use of finer zone labels related to the spatial distribution of faba beans samples from different areas within the region is presented. The current deliverable also presents the ALLIANCE digital knowledge base, a centralized platform that consolidates structured information on food fraud cases, detection methods, and relevant regulations, supporting both risk assessment and knowledge sharing among stakeholders. Additionally, the predictive analytics system offers robust fraud detection capabilities and supplier risk profiling, as demonstrated in the Feta Cheese food supply chain (FSC), while also illustrates the systems and adaptability to other FSCs. Last but not least, this deliverable includes the consumer demand assessment methodology, along with its results and key lessons learned.

The ALLIANCE solutions contribute to significantly advancing food fraud detection and prevention, by delivering innovative tools that effectively address the growing challenges and needs of industrial stakeholders, regulatory authorities as well as consumers.

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 up to Month 30 towards the afore-mentioned objectives along with the outcomes from each activity are comprehensively documented in the current deliverable, named D3.3 "Final AI-enabled tools & Digital Knowledge Database for Detecting Food Fraud using novel portable rapid testing for on-site inspection". This document serves as the second and final version of D3.2 " Interim AI-enabled tools & Digital Knowledge Database for Detecting Food Fraud using novel portable rapid testing for on-site inspection ", which was successfully submitted in M18.

1.2 Relationship to project work

The final versions and results of the ALLIANCE novel tools, which include advanced portable qPCR DNA sequencing, Near-Infrared (NIR) and Hyperspectral Imaging (HSI) Spectroscopy, and leverage AI and predictive analytics to detect food in Food Supply Chains, are comprehensively presented in the current deliverable. These solutions have been developed and implemented within Work Package 3 (WP3), and more specifically in Tasks T3.2 to T3.6, and their delivery is achieved through this document.

This deliverable provides a detailed description of the experimental design, implementation, use case applications, data collection methods, processing workflows, analytical frameworks, key results, and validation processes for each of the ALLIANCE tools. The DNA-based authentication and traceability tool, which demonstrated an accurate method for EVOO supply chain traceability, are the outcome of T3.2, led by BIOC. In Task T3.3, ASCINCAR contributed to the advancement of NIR and HSI spectroscopy tools, validating their robustness in detecting fraud in PGI Asturian faba beans. D3.3 also presents the ALLIANCE digital knowledge base (implemented in T3.4), which is a strategic tool developed within the project to enhance detection, monitoring, and prevention of food fraud across complex supply chains. Furthermore, the food fraud prevention system with predictive analytics is part of this deliverable. More specifically, it includes a thorough analysis of the development of a complete milk quality assessment and fraud detection platform for detecting hidden associations among the FSC performance parameters and potential vulnerability risks.

Last but not least, this deliverable also presents the outcomes from the monitoring and assessment of consumer perceptions of products and the innovative traceability systems, as carried out in Task 3.6.

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

The herein deliverable represents the continuation of the work carried out and reported in detail until M18 in the D3.2. Considerable work has been done since the last reporting period (that will be explained through the following subsections). Since the technical methods used and the importance of DNA-based authentication in preventing food fraud have already been reported in D3.2, it is important to emphasize that no deviations from the original plan of Task 3.2 have occurred. The DNA-based authentication and traceability tool has been proven to be robust and highly accurate, and therefore we are confident that it will serve as a useful tool both during the pilot phase, as well as beyond the project lifespan.

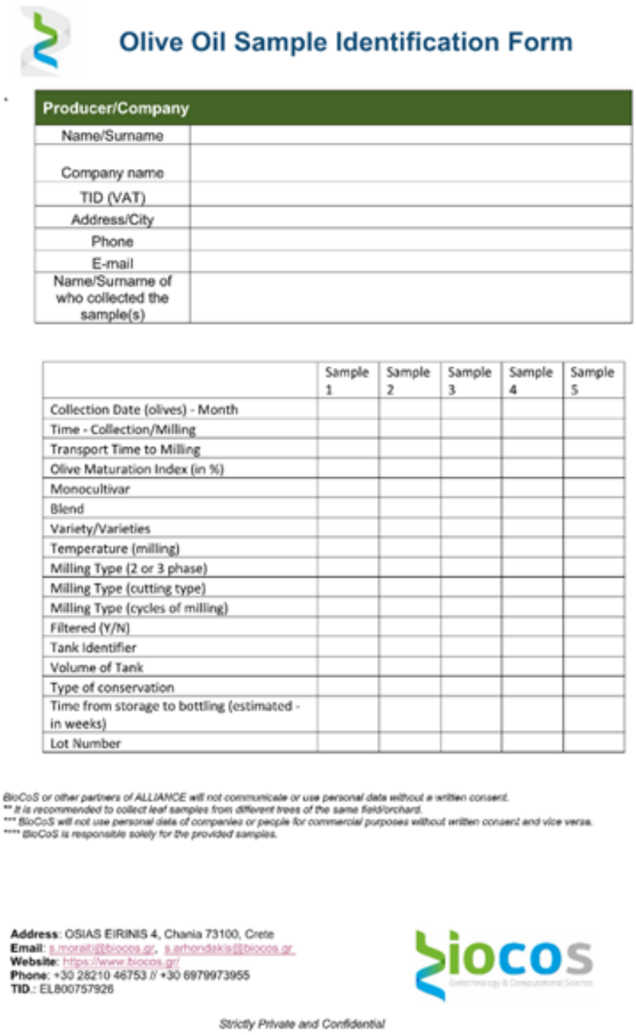
2.2 Experimental design and implementation

2.2.1 Use case: Extra Virgin Olive Oil (EVOO)

Task 3.2 is Interconnected with the ALLIANCE pilot demonstration focusing on PDI/PGO EVOO (Task 4.2). Throughout the pilot demonstration, the efforts undertaken in this task will be utilized to establish an end-to-end closed system for the CIA partners of Umbria, as well as the retail partner of MASOU. DNA data generated in Task 3.2 are utilized to train a machine learning/artificial intelligence (ML/AI) algorithm, automating the classification of Umbria PDO olive varieties (leaves), and their represented EVOO. 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 blockchain system developed by UTH, ensuring complete DNA-based traceability from field to store.

2.2.2 Data collection and (pre-)processing

As reported in D3.2, BioCoS focused on discovering and validating novel biomarkers in order to identify 6 Italian olive cultivars, namely Moraiolo, Frantoio, Leccino, Dolce Agogia, Rajo, and San Felice. The samples collection has been carried out through collection forms that were presented In D3.2, while additional forms (Figure 2-1) were made available to CIAUM in order to provide us with EVOO samples. On the one hand, the collection forms (one entry per sample) and on the other hand the DNA results of entry facilitate the import of data to the blockchain. Moreover, these anonymised data, also represent to a great extend the link between the different stakeholders of the olive oil supply chain, as this information is currently utilized by the blockchain platform to connect each node (i.e., field, mill, storage, bottling, retailer) of the value chain.



Olive Oil Sample Identification Form

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

	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5
Collection Date (olives) - Month					
Time - Collection/Milling					
Transport Time to Milling					
Olive Maturation Index (in %)					
Monocultivar					
Blend					
Variety/Varieties					
Temperature (milling)					
Milling Type (2 or 3 phase)					
Milling Type (cutting type)					
Milling Type (cycles of milling)					
Filtered (Y/N)					
Tank Identifier					
Volume of Tank					
Type of conservation					
Time from storage to bottling (estimated - in weeks)					
Lot Number					

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 solely for the provided samples.*

Address: OSIAS EIRINIS 4, Chania 73100, Crete
Email: s.moraiti@biocos.gr, s.arhondakis@biocos.gr
Website: https://www.biocos.gr/
Phone: +30 28210 46753 // +30 6979973955
TID.: EL800757926

biocos
Biotechnology & Computational Science

Strictly Private and Confidential

Figure 2-1 Olive oil sample collection form

While in the previous reporting period the focus has been given to leaf samples; the second half of Task 3.2 has been focusing on EVOO samples. During discussions with CIAUM, considering both EVOO samples availability (2023 and 2024 harvesting years have been quite devastating for many Italian regions) and KPIs related to Task 3.2 and Task 4.2, we agreed on giving emphasis on the following PDO EVOO labels: a) Moraiolo, b) San Felice and c) Rajo. All three labels are the sole representatives of 3 sub-regions of PDO EVOO production of Umbria; Assisi Spoleto, Martani and Amerini, respectively. In collaboration with CIAUM, we collected EVOO samples from the storage tank (right after EVOO extraction). This process allowed us to utilize the biomarkers developed and validated for the leaf samples, directly to EVOO samples.

2.2.3 Data analysis and Machine Learning for DNA-data classification

A key highlight from this period was identifying and validating one additional proprietary biomarker, namely OL9, that can distinguish two varieties (namely Moraiolo from San Felice), that was previously reported as one cluster (D3.2). Having said that, this represents a significant step forward as currently we are able to correctly classify 5 out of 6 varieties.

In order to assess and validate the biomarkers efficiency, our partners from LGL carried out experiments regarding the efficiency of our method, as well as the limit of detection (LOD). The total number of biomarkers assessed was 5 (namely OL3, OL4, OL6, OL7, OL8 and OL9), with only one of them (OL6) not performing satisfactorily, and therefore was neglected for further assessment. Specifically, Table 2.1 represents the biomarkers utilized for the ML (that were the top-performing), as well as the newly discovered one, their efficiency and R^2, and their DNA amount, as reported by LGL.

Table 2.1 Validation of Novel Biomarkers in terms of qPCR efficiency, R2 and their DNA amount, expressed as copy numbers per microliter. The R2 Is determined from the equation of the slope from the efficiency experiments.

Biomarker name	qPCR Efficiency	R^2	DNA Amount (cp/uL)
OL4	95.5%	0.999	100
OL8	100.1%	0.999	50
OL9	95.3%	0.995	20

The EVOO samples received were subjected to DNA Isolation, and then a real-time PCR coupled with a High-Resolution Melting Analysis (HRM) was carried out, for all the above-mentioned biomarkers. In total, 100 experiments per variety were carried out to validate the ML pipeline (that was initially trained with olive leaf samples) and investigate whether the latter needed further fine tuning. The possibility that the results from the DNA-data obtained from leaf samples compared to DNA-data obtained from EVOO samples to be slightly different is expected. This can be attributed to factors such as the isolated DNA quality from the EVOO, but also the presence of a pollinator variety in the field where the olive fruit were collected, milled, and its DNA presence in the produced EVOO. The reason that this is mentioned herein, is because during the collection of the field data, the producers were asked to provide this information (presence/absence of a pollinator and its variety). This facilitated the interpretation of the results coming from the field compared to the ones coming for the tank. For the sake of clarity, the presence of a pollinator or pollinators in a field would translate also in the presence of other varieties in a monovarietal EVOO, which is not considered fraud, and it is in reality a quite beneficial act from the producers to maximize their yield. For the sake of simplicity, the herein presented EVOO results derived from fields that no pollinator was utilized.

In respect to the ML pipeline, additional training was performed with the latest marker that we discovered (OL9), in order to inspect the overall performance of the model. We observed that utilizing solely the latest biomarker for the varieties Moraiolo (MO), San Felice (SF) and Rajo (RA), resulted in improved metrics performance to the overall model classification. More specifically, the following table (Table 2.2) provides the relevant metrics of the system.

Table 2.2 Performance metrics. MO denotes Moraiolo, RA denotes Rajo and SF denotes San Felice.

Statistics by Class	Class: MO	Class: RA	Class: SF
Sensitivity	0.8667	0.9643	0.9333
Specificity	0.9483	1.0000	0.9310
Pos Pred Value	0.8966	1.0000	0.8750
Neg Pred Value	0.9322	0.9836	0.9643
Prevalence	0.3409	0.3182	0.3409
Detection Rate	0.2955	0.3068	0.3182
Detection Prevalence	0.3295	0.3068	0.3636
Balanced Accuracy	0.9075	0.9821	0.9322

Our ML model demonstrates strong and well-balanced performance across all three classes (MO, RA, and SF). It achieves high sensitivity and specificity for each class (variety), with particularly outstanding results for class RA, which is predicted with perfect specificity and precision. Class MO, while slightly lower in sensitivity, still maintains solid performance metrics, and class SF shows a good balance between recall and precision. The detection rates closely align with the true prevalence of each class, indicating that the model is not biased toward any specific category. Overall, the model exhibits high balanced accuracy (ranging from 0.91 to 0.98 across classes) and robust predictive capabilities, making it a reliable tool for multi-class classification in this domain. As a next step, we continued with testing the model with new HRM-data that was not part of the original training-testing dataset. Taking the above into consideration, we started testing the ML pipeline with DNA-data deriving from the EVOO samples.

2.3 Key Results

During the second part of Task 3.2 we were able to discover and validate one additional biomarker that allowed us to distinguish 5 out of 6 olive varieties. After several Improvement done in the ML model, the pipeline is now able to classify MO, SF, and RA, whereas in the previous version of the model (reported In D3.2) the varieties MO and SF were clustering together. For the classification of the rest of the varieties (Frantoio, Leccino and Dolce d'Agogia), the previous model (reported in D3.2) is currently utilized, since both models tested performed the same for the aforementioned varieties. The performance of the model in terms of sensitivity increased 38.5% (Relative improvement; correct classification of the actual positives, Absolute gain; 25%). Moreover, balanced accuracy that accounts for both sensitivity and specificity (correctly identifying both positives and negatives) increased nearly 19% better overall. Our ML model represents a very promising tool to analyse and classify not only DNA from olive leaf samples, but also DNA outcomes from EVOO. Additionally, it is worth mentioning that all the biomarkers discovered herein by BIOC were validated by our partners at the LGL, and their overall performance - except from OL6 - was satisfactory and shows potential for further scalability, including more varieties to deliver a panel of markers for EVOO DNA-based authentication.

2.4 Validation and final implementation

As mentioned in an earlier section, the validation of the markers carried out at the LGL, while the ML model was validated using EVOO samples. The tool presented herein, as well as Task 3.2, is interconnected with Task 4.2, which Is the EVOO pilot. Therefore, our next steps are linked to the implementation of the tool to the pilot (Task 4.2) and its demonstration. The first step is to perform DNA testing in all the nodes of the olive oil supply chain and import the outcomes to the ML model. The model will classify the EVOO labels in their respective varieties, and these results will be uploaded to the blockchain system. As described above, the different nodes have been carefully selected in order to include all the relevant stakeholders, so that the latter will have access to information that is of value to them. The DNA testing and the ML classification will act as the linkage between the documents (e.g. hard copies) and/or the physical Items (e.g. bottled olive oil), but instead of tracing solely these, they will also trace the content (the olive oil), and therefore an additional layer of assurance will be provided enhancing the traceability and authenticity of the product.

3 Enhanced Food Fraud Detection with Advanced Spectroscopy

3.1 Introduction

In this deliverable, we continue with the approach outlined in D3.2 regarding the application of NIR and HSI technologies in the food sector. The fundamental approaches and methodologies that underpin the use of portable devices to ensure product authenticity and quality remain unchanged, demonstrating the robustness and relevance of the strategy adopted from the project's inception.

It is important to note that while considerable progress has been made in data collection and analysis, details of which will be discussed in subsequent sections, in this section the update is limited solely to reaffirming the previously established theoretical and methodological framework without any substantive modifications. In this study we are testing two distinct strategies based on the use of: i) handheld NIR sensor device; and ii) portable HSI camera, each with the same goal of rapid, on-site authentication but relying on different devices to maximize our chances of robust fraud detection.

3.2 Experimental design and implementation

3.2.1 Use case: PGI Asturian Faba Beans

As mentioned in D3.2, PGI Asturian faba beans are dried and dehulled from the Phaseolus vulgaris L. species, grown in Asturias and classified within the Extra and First categories. They are characterized by a creamy white colour, a kidney-shaped form, and an average of 100 -110 beans per 100 g, offering a balanced nutritional profile. Moreover, they are the main ingredient in the traditional Fabada Asturiana.

The main objective of this pilot is to develop and validate, in an operational control environment, a digital tool based on low-cost and portable advanced optical sensor devices, specifically utilizing NIR and HSI technologies, for the detection of fraudulent practices in the PGI Asturian Faba bean products. The approach can be summarized in one primary use case and one proof of concept: i) primary use case focuses on detecting the intentional blending of PGI-certified beans with lower-priced beans from South America (specifically from Bolivia); and ii) the proof of concept aims to establish distinct fingerprints based on the plot location within Asturias, identifying and differentiating between bean batches from various certified plots areas.

The primary end-users of these tools are the two main quality control actors for the faba beans products with PGI certification: i) namely the PGI control body (partner "IGP Faba de Asturias" - "IGPFA"); and ii) the competent public authority responsible for food quality and authenticity (associated partner "CMAS"). During the validation rounds, and more specifically during the validation campaigns, both end-users will integrate the digital tools within their existing control measures and protocols to assess performance and benchmark results against established control procedures.

3.2.2 Data collection and (pre-) processing

To date, samples from the 2022, 2023, and 2024 harvests have been measured following the experimental design detailed in D3.2, section 3.3.1, with one slight modification. Specifically, ground beans measurements have been eliminated to reduce the measurement time, in line with the project reviewer's comment during the project review meeting. It was determined that measuring ground beans is unnecessary for the purpose of using a precise, fast, and portable measurement method, since including ground bean measurements would require additional sample preparation that greatly increases the data acquisition time. In Figure 3-1 and Figure 3-2, the modified experimental design is illustrated.

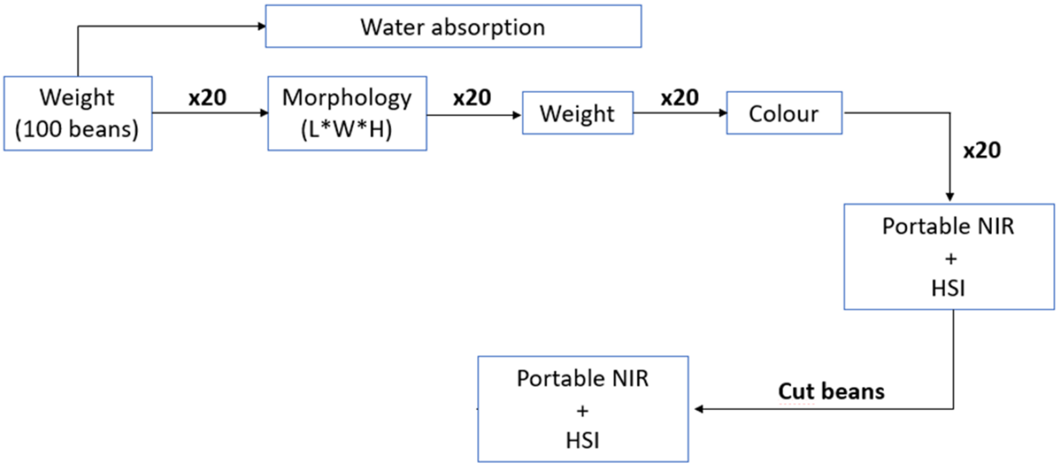


Figure 3-1 Schema of the modified protocol for physico-chemical and spectral measurements.

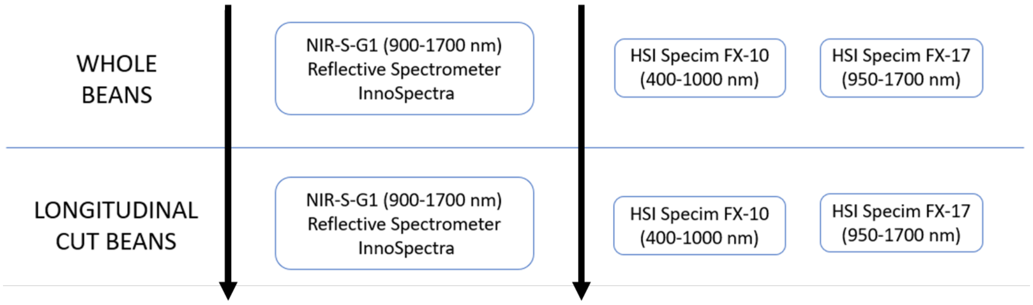

WHOLE BEANS
NIR-S-G1 (900-1700 nm) Reflective Spectrometer InnoSpectra
HSI Specim FX-10 (400-1000 nm)
HSI Specim FX-17 (950-1700 nm)
LONGITUDINAL CUT BEANS
NIR-S-G1 (900-1700 nm) Reflective Spectrometer InnoSpectra
HSI Specim FX-10 (400-1000 nm)
HSI Specim FX-17 (950-1700 nm)

Figure 3-2 Schema of the modified protocol measurement NIR and HSI.

The current dataset comprises 36 samples from the 2022 harvest (28 from Asturias, 2 from Galicia, and 6 from Bolivia), 55 samples from the 2023 harvest (34 from Asturias, 1 from Galicia, and 20 from Bolivia), and 51 samples from the 2024 harvest (33 from Asturias, 4 from Galicia, and 14 from Bolivia).

For each sample, physico-chemical data were collected using classical laboratory techniques. However, it is important to note that this analysis was performed only on the samples from the 2022 harvest and just on the Bolivia samples from the 2023 harvest (Table 3.1). This selective approach was adopted to achieve an approximately balanced number of samples from Asturias and Bolivia and to evaluate potential differences between PGI Asturian faba beans and foreign faba beans.

Table 3.1 Physico-chemical parameters collected to date by classical laboratory techniques.

	WHOLE bean	LONGITUDINALY CUT bean	MILLED bean
Morphology (length, width and thickness)	3360	-	-
Colour (L*, a*, b*)	1120	-	168
Weight/100 beans	56		
Weight/ each bean	1120	-	-
Humidity	-	-	56
Absorption	56	-	-

On the other hand, spectral data were obtained via NIR and hyperspectral imaging (HSI) methods. In the case of NIR, each spectra generates a ".csv" file. By contrast, in the case of HSI, each hyperspectral image is loaded into memory along with its corresponding white (reference) and black (dark current) images. Segmentation techniques based on thresholding are then applied to isolate the bean region, and an average spectrum is computed from all pixels within this region, as more extensively detailed in D3.2. These average spectra are saved in ".csv" format and later imported into Python for advanced data processing. A summary table, which consolidates the acquired spectra and hyperspectral images, is also included (Table 3.2)

Table 3.2 Information gathered to date by spectroscopy techniques.

	WHOLE bean	LONGITUDINALY CUT bean	MILLED bean
Portable NIR Inno	5680 spectra (two spectra for each bean)	5680 spectra (two spectra for each bean)	0
HSI-FX10 camera	142 images (20 beans for each image)	142 images (20 beans for each image)	36 images (one foe each sample)
HSI-FX17 camera	142 images (20 beans for each image)	142 images (20 beans for each image)	36 images (one foe each sample)
NIR ASD	0	0	108 spectra (three spectra for each sample)

This comprehensive dataset forms the basis for further data analysis and model development, ultimately leveraging NIR and HSI technologies for food authentication purposes.

3.2.3 Data analysis

Physico-chemical analysis

As described in Deliverable D3.2, an initial statistical analysis of the physico-chemical properties was carried out using box-and-whisker plots and violin plots. For this analysis, data from Bolivian faba beans collected in 2023 were also incorporated, resulting in a total of 28 samples from Asturian beans and 26 from Bolivian beans, thus ensuring a more balanced dataset.

The results show that significant differences in colour continue to be observed (Figure 3-3), albeit to a lesser extent than in previous analyses. These differences primarily appear in the colour parameter *a*nd the *Hue angle*. The parameter *a* is part of the CIELAB colour system and represents the red-green axis, while the *Hue angle* indicates the perceived tone, typically calculated from the a and b values (e.g., using the function arctan(b/a)). These parameters allow for the identification of subtle tonal variations between samples of different origins.

A key observation emerges from the violin plot representing the *ac* parameter (Figure 3-3). Asturian samples form a relatively symmetrical distribution around zero, with a modest negative tail extending slightly below, 0.5 and a median hovering just above zero. By contrast, the

Bolivian distribution extends further in the positive direction, reaching values close to 1.0, and shows a median slightly higher than that of Asturias. This indicates that Bolivian beans exhibit greater variability in *ac*, particularly toward higher values, which could be linked to differences in cultivation practices or inherent varietal characteristics.

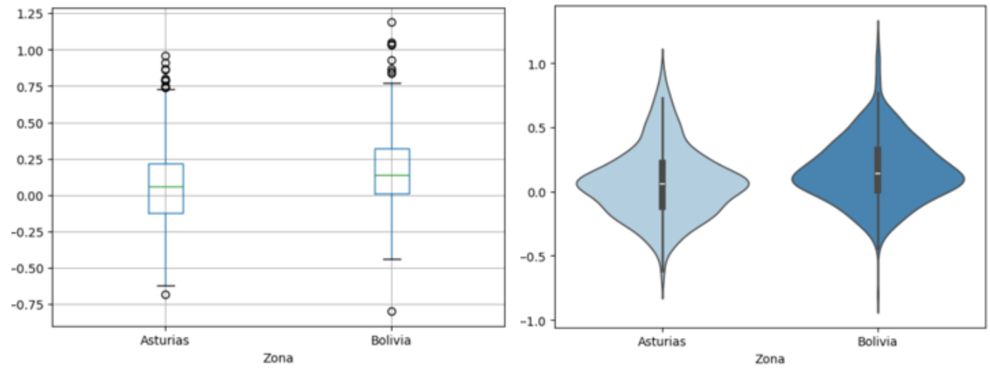


Figure 3-3 Box and whisker diagrams (left) and violin diagrams (right) for colour ac parameter.

The violin plot comparing Hue angle for Asturian (left) and Bolivian (right) (Figure 3-4) faba beans reveals further insights. The Asturian samples display a broader distribution of Hue angles, suggesting higher variability or the existence of distinct subgroups; the “double-lobed” shape indicates that while some beans cluster around slightly negative Hue values, others lean towards positive values. Meanwhile, the Bolivian samples exhibit a narrower, more centered distribution, with the central boxplot area positioned around moderately positive Hue values, indicative of a more uniform hue.

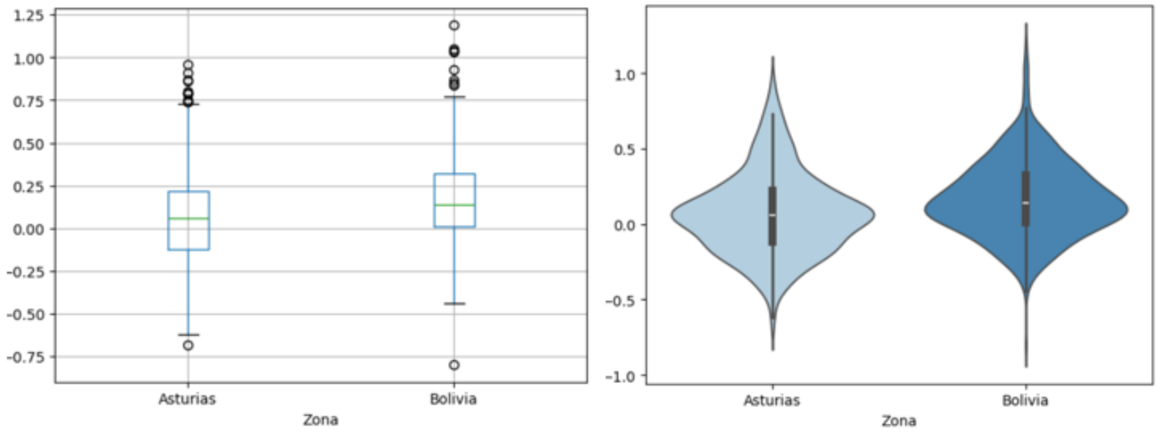


Figure 3-4 Box and whisker diagrams (left) and violin diagrams (right) for Hue angle.

Taken together, the differences in distribution shape and range indicate that colour tonality, as measured by the Hue angle, continues to be a useful parameter for differentiating Asturian and Bolivian beans, reinforcing other physico-chemical evidence presented in this study.

In addition, differences in water absorption were noted (Figure 3-5), with the Asturian faba beans exhibiting a more homogeneous absorption pattern with defined limits compared to the Bolivian beans. These findings reinforce the evidence that, despite reduced differences in magnitude, the physico-chemical properties continue to differentiate the faba beans based on their origin, which is key for product authentication.

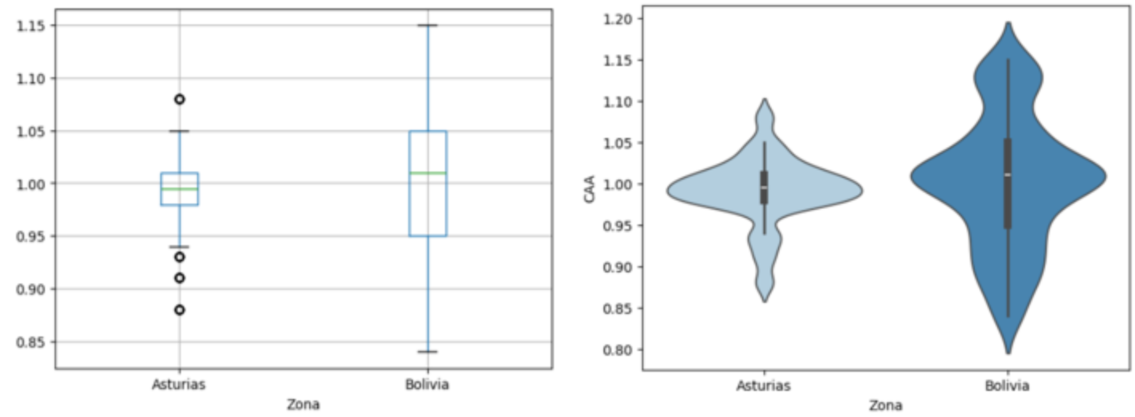


Figure 3-5 Box and whisker diagrams (left) and violin diagrams (right) for water absorption

Overall, the balanced dataset and the refined analysis approach provide a strong foundation for future work aimed at further validating and enhancing the authentication methodology.

Spectra analysis

The spectroscopic data obtained using NIR and HSI were analyzed in Python, with the spectra stored in a data frame alongside their class labels. Only the wavelengths between 930 and 1670 nm (for the portable device NIR-S-G1 and the FX17 camera) and between 430 and 970 nm (for the FX10 camera) were used to avoid noise at the measurement limits. The analysis followed three key steps: pre-treatment (applying methods such as smoothing, normalization, and baseline correction), variable (wavelength) selection (using techniques like stepwise selection, recursive feature elimination, mutual information, and PCA loadings), and the development of classification models. For the first use case, models such as PLS-DA (often combined with other algorithms like Random Forest, XGBoost, or non-linear SVM) were applied, focusing on samples from Asturias and Bolivia only, as Galicia's representation was too limited (n=7).

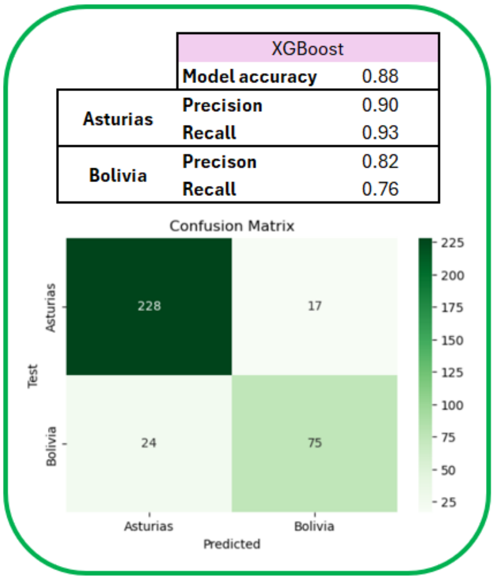
Much of the detailed methodology and experimental procedures, along with the extensive discussion of each pre-treatment and modeling step, is more comprehensively described in D3.2. This summary serves as an overview of the spectral analysis approach while emphasizing that the approach was refined through rigorous testing to achieve robust classification performance based on the available spectral data

Use case: Analysis of spectral data using portable NIR

Spectral data analysis from the portable NIR device (NIR-SG1 Innospectra) was conducted on both whole and longitudinally cut faba beans. In this analysis, the raw dataset, now extended to include samples from the three available campaigns, was split into training and test sets at an 80/20 ratio. Various pre-treatments, including SNV, Savitzky-Golay smoothing (SG), MSC, and DeTrend, were applied to the training data in different combinations to enhance data quality.

Following pre-treatment, PCA was used for wavelength selection in order to reduce the dimensionality of the dataset. While alternative methods such as Stepwise selection are also being evaluated, the primary focus was on reducing dimensionality via PCA before applying the subsequent classification models. Subsequently, a PLS-DA algorithm was employed to further reduce dimensionality, and the latent variables obtained were input to multiclass classification models. A 10-fold cross-validation was performed on the training set, and the predicted values were calculated using the test set to derive model parameters, including accuracy, precision, and recall.

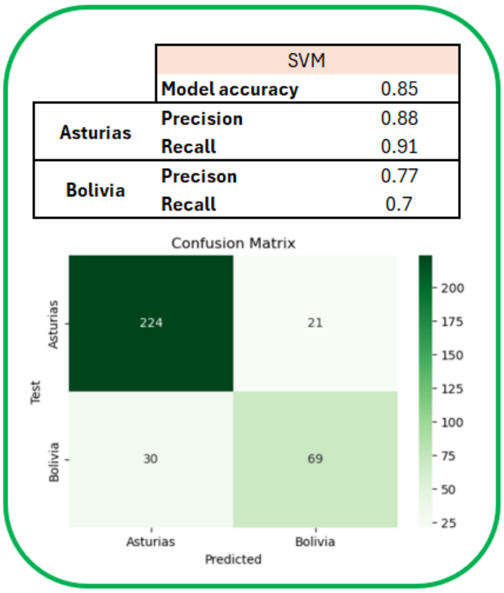
The best results for both whole and cut beans were obtained using a combination of PLS-DA and XGBoost. For this method, hyperparameter tuning was carried out, and the optimal parameters selected were: (n_components=25, max_depth=5, n_estimators=200). Using SNV as the sole pre-treatment (since DeTrend did not provide significant improvements), this approach achieved a best cross-validation score of 0.9235 and a model accuracy of 0.8808. Specifically, for Asturias samples, the model achieved a precision of 0.9048 and a recall of 0.9306, while for Bolivia samples, the precision and recall were 0.8152 and 0.7576, respectively (Figure 3-6).



	XGBoost	
	Model accuracy	0.88
Asturias	Precision	0.90
	Recall	0.93
Bolivia	Precison	0.82
	Recall	0.76

Figure 3-6 Example accuracy, precision, recall and confusion matrix for whole beans obtained using XGBoost model.

Additionally, when employing a nonlinear SVM approach—with SNV as the primary pre-treatment and optimized hyperparameters (C=1, gamma=0.0001, kernel='linear') using GridSearchCV, the results for both whole and cut bean spectra were found to be similar. Specifically, the model achieved an overall accuracy of 0.85. For Asturian samples, the precision was 0.8819 and the recall 0.9143, while for Bolivian samples the precision was 0.7667 and the recall 0.6970 (Figure 3-7).



	SVM	
	Model accuracy	0.85
Asturias	Precision	0.88
	Recall	0.91
Bolivia	Precison	0.77
	Recall	0.7

Confusion Matrix

Test \ Predicted	Asturias	Bolivia
Asturias	224	21
Bolivia	30	69

Figure 3-7 Example accuracy, precision, recall and confusion matrix for whole beans obtained using SVM model.

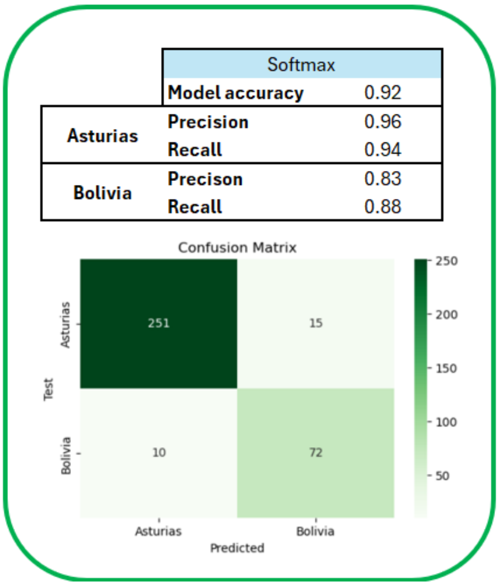
Both the PLS-DA plus XGBoost and the nonlinear SVM approaches demonstrated effective classification performance, but some differences in the metrics suggest that one may have a slight edge over the other. Notably, both techniques produced similar results for whole and cut bean spectra. Overall, while both models are robust, the XGBoost approach appears to offer a modest improvement in predictive performance across the evaluated metrics.

This analysis integrates the spectral data obtained from the three campaigns, providing a comprehensive and balanced dataset for model development. A more detailed discussion of these results, along with graphical outputs of the models, will be included in the final report.

Use case: Analysis of spectral data using HSI

The analysis of spectra obtained via HSI, with both the FX10 and FX17 cameras, follows a process similar to that employed for NIR data, with the main difference being the extraction of a spectrum for each individual bean pixel. After applying suitable pre-treatments (such as SNV, Savitzky-Golay smoothing, MSC, or DeTrend) to the individual pixel spectra, an average spectrum is computed for each bean. The data, now extended to include the spectral information collected across all three available campaigns, is split into training and test sets using an 80/20 ratio. Subsequently, the PLS-DA method is applied in combination with various classification algorithms, and a 10-fold cross-validation is conducted to evaluate performance. For HSI data, the SNV pre-treatment consistently proves optimal across different bean formats and camera types, as previously noted in Deliverable D3.2.

The best results for in-line HSI data analysis were achieved using a combination of PLS-DA coupled with SoftMax under the optimal parameter setting (Best params: (n_components=20). Under the same model conditions, the FX10 camera produced an accuracy of 0.9248, which is notably higher than the 0.8649 accuracy obtained with the FX17 camera. These findings indicate that, for the current dataset and pre-treatment settings, the FX10 camera offers superior predictive performance. Such results further support the role of in-line HSI technology in enhancing traceability and food authenticity for PGI Asturian faba beans. A more detailed discussion of these findings, including graphical presentations of the model results, is provided in Deliverable D3.2 and will be elaborated upon in subsequent sections.



		Softmax
	Model accuracy	0.92
Asturias	**Precision**	0.96
	Recall	0.94
Bolivia	**Precison**	0.83
	Recall	0.88

Figure 3-8 Example accuracy, precision, recall and confusion matrix for whole beans obtained using SVM model.

Proof of concept: Analysis of regional variations using NIR and HSI

In addition to the primary use case, detecting fraudulent mixtures between PGI Asturian and foreign beans, a proof-of-concept was conducted to investigate whether significant spectral differences exist among faba beans from different regions within Asturias. For this analysis, both NIR and HSI spectral data were acquired from samples originating from multiple Asturian zones, following the same standardized data collection and pre-treatment protocols as described previously.

The spectral data were pre-treated using methods such as SNV and DeTrend, and dimensionality reduction was performed using PCA and PLS-DA. Advanced classification algorithms were then applied to assess regional variations in the spectral signatures. However, the analysis did not reveal statistically significant differences between beans from various Asturian regions. In fact, the PLS-DA results by zones demonstrated only a moderate overall accuracy of approximately 0.75, with subtle spectral variations that are insufficient for reliable discrimination of geographic origin within the region.

This indicates that, under the current experimental conditions and with the available sample sizes, NIR and HSI techniques are more effective for distinguishing between beans of fundamentally different origins (e.g., Asturian versus foreign) than for capturing the finer regional distinctions among Asturian zones. The confusion matrix presented below further illustrates this moderate classification performance and the variability in predictions across councils (Figure 3-9).

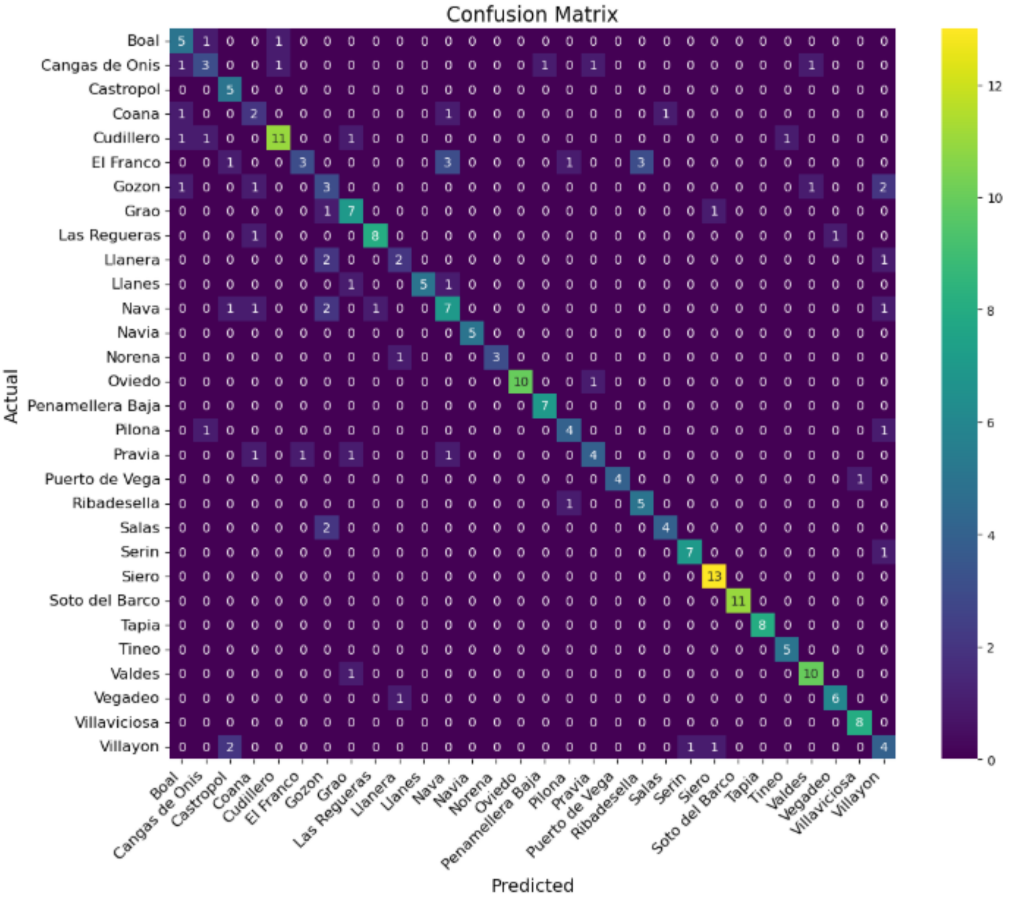


Figure 3-9 Confusion matrix for the PLS-DA classification by council using spectral data from Asturian faba beans.

Further investigations were performed to refine the regional analysis. The data from the three campaigns were normalized using Standard Scaler, and histograms along with kernel density plots of the LDA scores were generated to compare distributions, such as “Coast” versus “Inland” (Figure 3-10). The histogram for the first LDA component (LDA1) shows that Costa samples tend to exhibit more negative values, while Interior samples lean toward positive values. Although there is significant overlap near LDA1 ≈ 0, indicating that some samples from both regions share similar spectral characteristics, the distinct peaks suggest that LDA1 captures meaningful differences between the two groups. Additionally, overall metrics, including an average intra-zone spectral distance of 0.4395 versus an inter-zone distance of 0.5056, indicate that the overall differentiation between regions is modest.

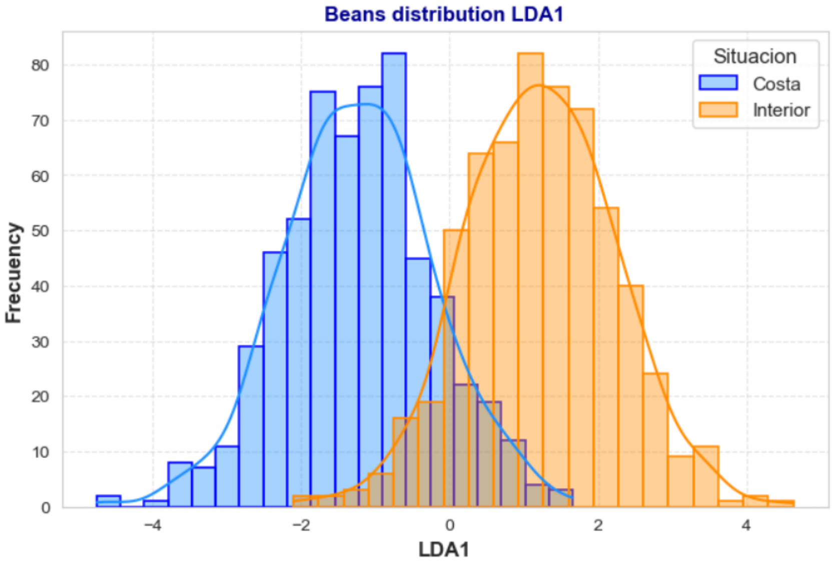
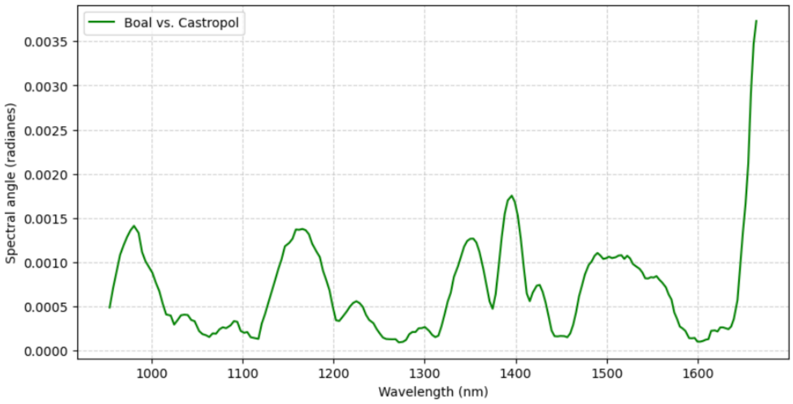


Figure 3-10 Distribution of LDA1 scores for faba bean samples from “Coast” and “Inland” groups[1].

In a complementary analysis, finer zone labels based on individual councils were employed. Advanced visualization techniques, including Kernel PCA, t-SNE, and UMAP, along with statistical tests such as the Kolmogorov-Smirnov and Anderson-Darling tests, were applied to further interrogate the data. A moving-average approach to Spectral Angle Mapping (SAM) revealed that certain spectral bands are particularly relevant for regional discrimination, with differences appearing more pronounced among some councils than others (Figure 3-11). Although these findings hint that targeted analysis of key spectral regions may enhance geographic differentiation, the overall regional variations remain subtle.



[1] Note: Although the graph uses the original Spanish labels “Costa” and “Interior”, these correspond to “Coast” and “Inland” in English.

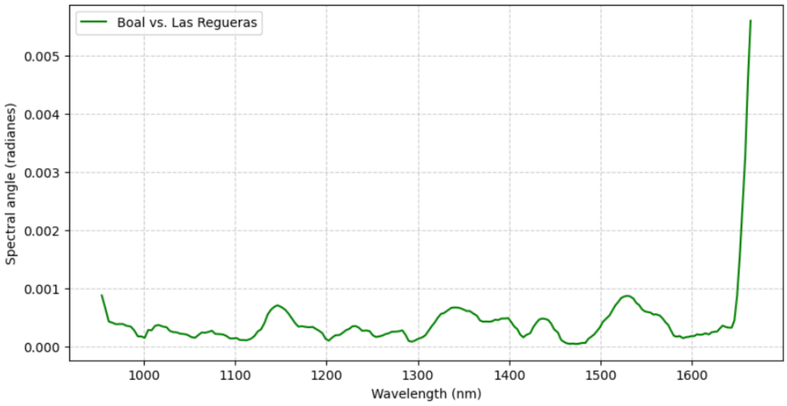
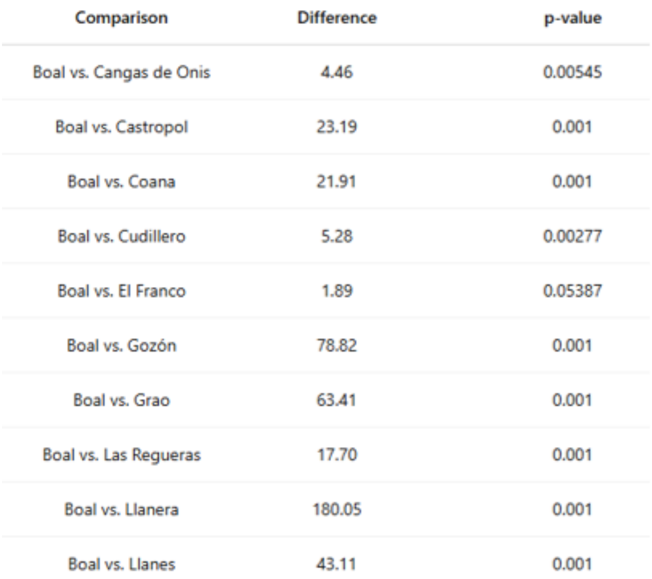


Figure 3-11 SAM comparisons highlight significant spectral differences between Boal and Cangas de Onis (upper image) and minimal differences between Boal and Las Regueras (lower image).

These insights are further supported by quantitative metrics. For example, the Table 3.3 summarizes the differences and corresponding p-values obtained when comparing Boal samples with various councils, highlighting that while most comparisons yielded statistically significant differences ($p < 0.05$), the differences vary notably in magnitude across councils:

Table 3.3 Pairwise Comparisons: Boal vs. Other Councils (Differences and p-values).



Comparison	Difference	p-value
Boal vs. Cangas de Onis	4.46	0.00545
Boal vs. Castropol	23.19	0.001
Boal vs. Coana	21.91	0.001
Boal vs. Cudillero	5.28	0.00277
Boal vs. El Franco	1.89	0.05387
Boal vs. Gozón	78.82	0.001
Boal vs. Grao	63.41	0.001
Boal vs. Las Regueras	17.70	0.001
Boal vs. Llanera	180.05	0.001
Boal vs. Llanes	43.11	0.001

These metrics, along with results from the Spectral Angle Mapping (SAM) analysis with moving averages, indicate that certain spectral bands show greater relevance for differentiating between some councils than others. Although overall differences between zones remain subtle, this quantitative evidence reinforces our hypothesis that targeted analysis of key spectral regions could improve geographic discrimination. For a detailed view of the results, including comprehensive graphical presentations and confusion matrices, please refer to Figure 10.

One notable limitation is that spectral data have been collected from only three harvests, which means that, for each council, there are only three samples (one per harvest). Although each sample consists of measurements from 20 individual grains (totalling 60 grains per council), the overall number of samples per council may not be sufficient to fully capture the natural variability within each region. This limited sample size could affect the statistical power of the regional differentiation analysis and may require additional data collection in future campaigns to achieve more representative and robust conclusions.

For enhanced visualization, the samples were mapped onto a representation of the Asturias region using Python code, with different colours corresponding to the PCA results. This approach facilitates the observation of differences among faba beans from various zones, as shown in Figure 3-12.

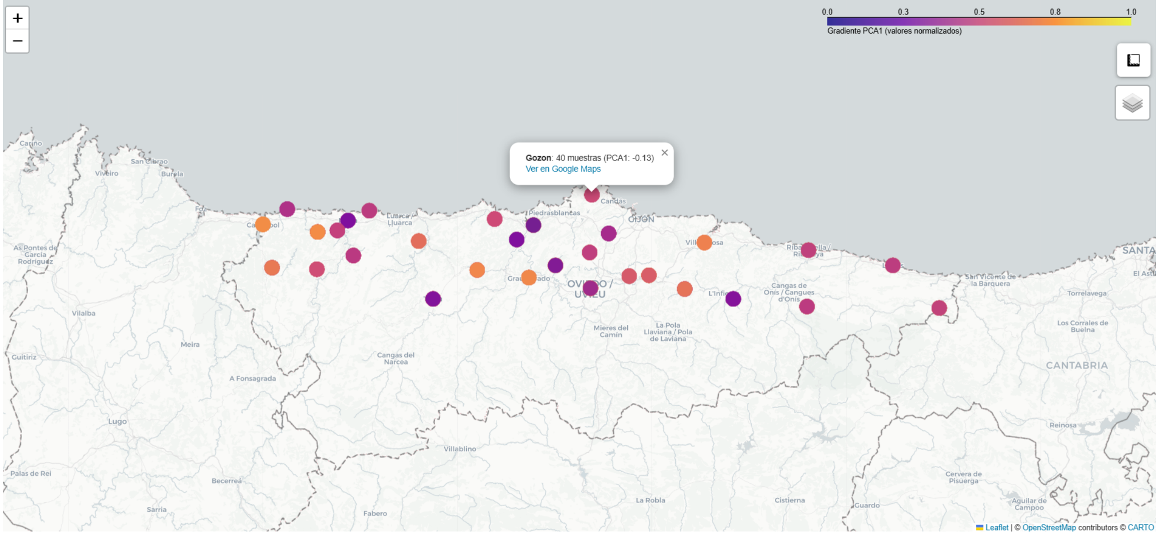


Figure 3-12 Map of the Asturias region displaying the spatial distribution of faba bean samples. Colors represent the clustering obtained from PCA.

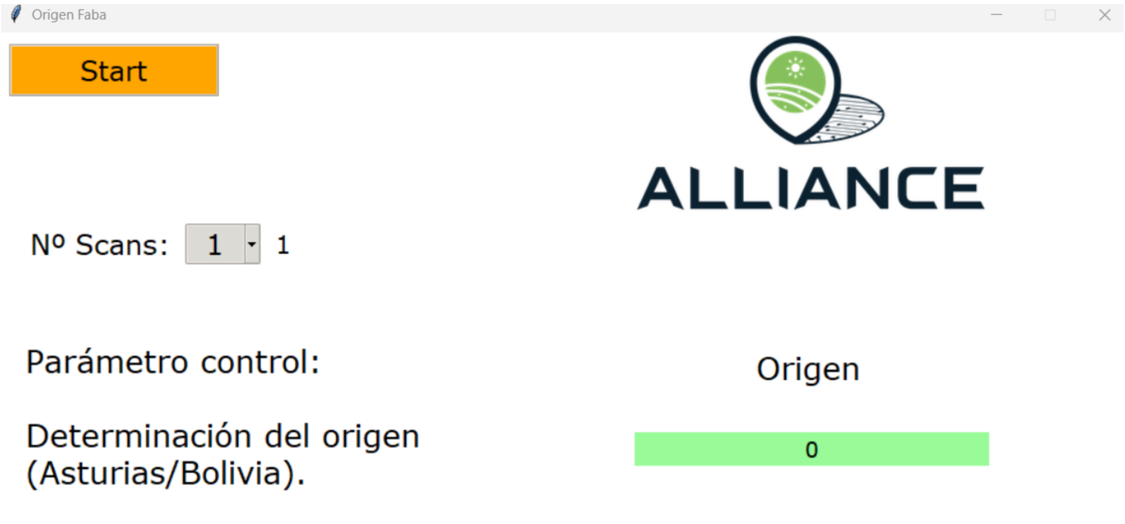
Future work will aim to increase the number of samples per council and investigate additional spectral features that could further enhance the differentiation between regions. These steps will help to validate and improve the authentication methodology for PGI Asturian faba beans.

Overall, while the proof-of-concept did not establish clear, statistically significant regional differences in the spectral signatures of Asturian faba beans, it provides valuable insights. The advanced analyses highlight both the potential and the limitations of the current methodology, serving as a basis for future refinements aimed at enhancing the differentiation capability and, ultimately, the authentication process.

3.2.4 Data visualization

Finally, two interfaces were developed to display the analysis results: one for NIR data and another for HSI data. The main idea behind both interfaces is to ensure they are user-friendly for non-experts, minimizing the learning curve while providing accurate, real-time predictions regarding bean origin. Python was chosen as the optimal language for this development because of its versatility and the ease with which packages for graphical user interfaces (such as TkInter) can be used.

For the NIR interface, the program receives two spectra, one from each side of the faba bean. These spectra are averaged to generate a representative spectrum, which then undergoes pre-treatment, wavelength selection, and is input into the classification model. The prediction result is automatically displayed on the screen for that sample (Figure 3-13). Similarly, the HSI interface functions in a comparable manner; however, the input data differ, as it receives processed hyperspectral data instead of the two discrete NIR spectra. In addition, while the underlying development approach remains the same, the design of the HSI screen has been slightly adjusted to accommodate its specific data presentation requirements (Figure 3-14).



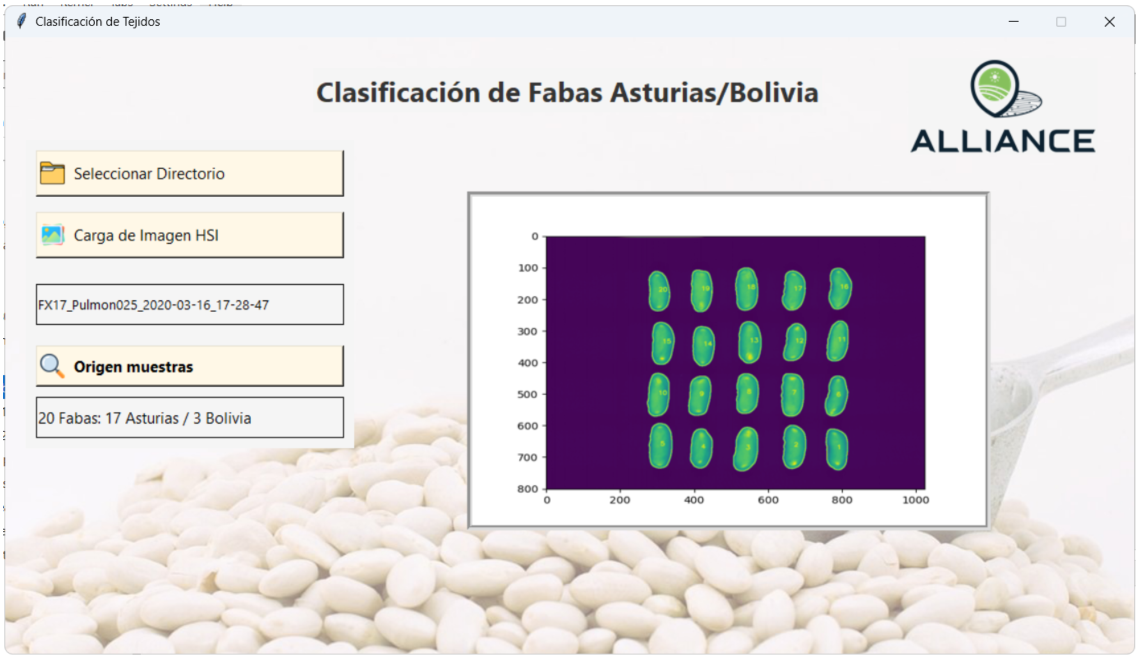
Origen Faba

Nº Scans: 1 1

Parámetro control: Origen

Determinación del origen (Asturias/Bolivia). 0

Figure 3-13 Results display window for end-consumers for NIR technology.



Clasificación de Tejidos

Clasificación de Fabas Asturias/Bolivia

FX17_Pulmon025_2020-03-16_17-28-47

20 Fabas: 17 Asturias / 3 Bolivia

Figure 3-14 Results display window for end-consumers for HSI technology.

The internal functioning of the application is as follows: a Python class stores the pre-treatment parameters, wavelengths, and algorithm configurations used to develop the chemometric model. When the spectrometer generates the necessary spectra, an object of this class is instantiated, modularly performing all pre-treatment and predictive steps. The final prediction is then extracted from the object and displayed on screen. The workflow for both interfaces includes running the prediction program, automatically collecting the required spectra, and displaying the result, with simple controls to restart the measurement or change the directory where data are stored. Further details on the development and functioning of these visualization interfaces are provided in D3.2, section 3.3.4.

3.3 Key results

For the primary use case, detecting fraudulent mixtures between PGI Asturian faba beans and cheaper foreign varieties, significant differences were observed between samples originating from Asturias and those from Bolivia. Classification models based on portable NIR data yielded satisfactory performance, with accuracy, precision, and recall values in the range of 0.8 to 0.9. In contrast, the in-line HSI models delivered even higher performance, consistently achieving metrics above 0.9. These results underscore the superior robustness and precision of HSI technology, suggesting it can offer more reliable outcomes for food authentication and fraud detection in this context.

The proof-of-concept analysis, designed to identify potential spectral differences among faba beans from various regions within Asturias, revealed no significant regional variations. This finding indicates that, under the current experimental conditions and with the available sample set, the spectral characteristics of Asturian faba beans are too similar to allow for reliable discrimination based on their geographic origin.

Additional key results highlight that the study's comprehensive dataset—including thousands of physico-chemical determinations and over 8,640 spectral measurements—provides a strong foundation for future model refinement. Preliminary comparisons of various chemometric approaches (including PLS-DA paired with XGBoost and non-linear SVM models) demonstrated that both techniques yield high predictive performance, with the HSI-based models slightly outperforming their NIR counterparts.

Overall, these promising results not only validate the use of portable NIR and in-line HSI as effective tools for detecting major fraudulent practices in PGI Asturian faba beans, but they also suggest the potential for broader applications to other PGI/PDO foods. The integration of feedback from the PGI control body and food safety authorities has been instrumental in refining the experimental design and achieving these outcomes.

3.4 Validation and final Implementation

The developed digital tool has undergone an extensive validation process under operational control conditions. This phase involved integrating the measurement system, with both portable NIR and in-line HSI technologies, into the routine control protocols of the PGI and the competent public authority responsible for food quality and authenticity. The validation was aimed at assessing both the analytical performance of the chemometric models and the practicality of the tool in real-world settings.

Field tests confirmed that under controlled settings, the NIR-based models achieved acceptable performance parameters with accuracy, precision, and recall values ranging from 0.8 to 0.9, while the HSI-based models consistently exceeded these metrics, with values above 0.9. These promising results validate the tool's ability to reliably detect fraudulent practices involving the mixture of PGI Asturian faba beans with lower-priced foreign beans, specifically under the primary use case of adulteration detection.

A first validation campaign was carried out as a training phase, during which end users tested the portable NIR equipment both at their installations and in the field. This allowed them to become familiar with the tool and the data acquisition process. During this campaign, several challenges were detected that resulted in a decrease in the overall performance of the NIR-based model. Specifically, issues such as variations in field conditions, sample handling differences, and calibration deviations contributed to lower accuracy, precision, and recall values than expected. In response, corrective actions have been identified and will be implemented to address these issues. A second validation campaign is scheduled for October 2025 to verify whether these improvements lead to enhanced performance.

The final implementation phase focused on refining the user interface and ensuring the system's modularity and ease of use. An intuitive, user-friendly interface was developed for both NIR and HSI data visualization, allowing non-expert users to operate the tool with minimal training. This interface displays real-time predictions of bean origin, enabling rapid and informed decision-making during inspections.

Moreover, user feedback collected during the pilot trials with the PGI control body, and the regional food quality and safety authority was instrumental in fine-tuning the tool. This collaborative approach ensured that the final implementation not only meets the technical performance criteria but is also fully aligned with the operational needs and practices of the end users.

In summary, the validation and final implementation of the digital tool demonstrate its potential to serve as a revolutionary method for food authentication in PGI products, combining precision, speed, and portability. With corrective actions being taken based on the initial training campaign results and a follow-up validation planned for October 2025, the initiative is set to further enhance its accuracy and reliability, paving the way for broader applications to other PGI/PDO foods in the future.

More detailed information about the validation results of the NIR/HSI digital tools during the two validation campaigns planned on the PGI Asturian Faba Bean pilot will be included in D4.2.

4 Digital Knowledge Base for Food Fraud Mitigation

4.1 Introduction

The Digital Knowledge Base for Food Fraud is a strategic tool developed within the project to enhance detection, monitoring, and prevention of food fraud across complex supply chains. Addressing growing concerns over food integrity, consumer safety, and market transparency, the platform serves as a centralized hub for structured knowledge, actionable insights, and risk assessment tools. Designed for a wide range of stakeholders—including regulatory authorities, producers, and researchers—it integrates internal and external data sources through a flexible, modular architecture and user-friendly interface. This digital solution strengthens early detection capabilities, supports evidence-based decision-making, and reinforces trust within Quality Labelled and general Food Supply Chains.

4.1.1 Digital Knowledge Base Overview

Food fraud, encompassing deliberate adulteration, misrepresentation, and substitution of food products, continues to challenge global food systems. The increasingly complex and globalized nature of food supply chains—coupled with growing pressure on pricing, labelling, and consumer trust—has exposed vulnerabilities that bad actors can exploit. In response, this project recognized the urgent need for a robust, dynamic, and data-integrated digital infrastructure to consolidate dispersed knowledge, improve visibility into fraudulent practices, and empower stakeholders with actionable intelligence. The Digital Knowledge Base is designed to fulfil this role, bringing together both existing and project-generated data to enable risk identification, evidence-based action, and improved coordination among stakeholders working to ensure food authenticity and safety.

4.2 Experimental Design and Implementation

4.2.1 Knowledge Base Architecture and Features

System Overview

The architecture of the Digital Knowledge Base is grounded in a modular, scalable design that facilitates the seamless integration of diverse data sources, processing layers, and user-facing functionalities. This architecture consists of distinct but interconnected layers—data ingestion, processing and analysis, knowledge representation, and visualization—each optimized to manage and leverage complex datasets. External data sources such as certification documents, product specifications, scientific literature, and fraud alerts are combined with structured results produced by the project (e.g., test outcomes, classifications, vulnerability assessments). This layered design ensures robust data handling, meaningful insights, and user-friendly access to critical information.

Key Functionalities

- A search bar and filterable menu system for querying product types, fraud categories, and detection tools.
- Categorization by fraud types (e.g., dilution, origin misrepresentation).
- Integration of external links and metadata (scientific papers, certifications).
- Connection to a broader Early Warning and Decision Support System.
- Result cards with links, summaries, source documents, and contact points.

User Experience

The Digital Knowledge Base was developed with a user-centric design philosophy. The interface is intentionally intuitive, minimizing the learning curve for first-time users, including those with limited technical expertise. From the homepage, users can navigate through fraud categories, product-specific risk profiles, and tools for fraud detection and prevention. Each result page contains a summary, source references, and links to relevant documents or databases. This modular structure enables users to perform focused searches, explore related fraud cases, and build a contextual understanding of risks and interventions. User interaction is enhanced through smart filters and visual elements that support decision-making without overwhelming the user with complexity.

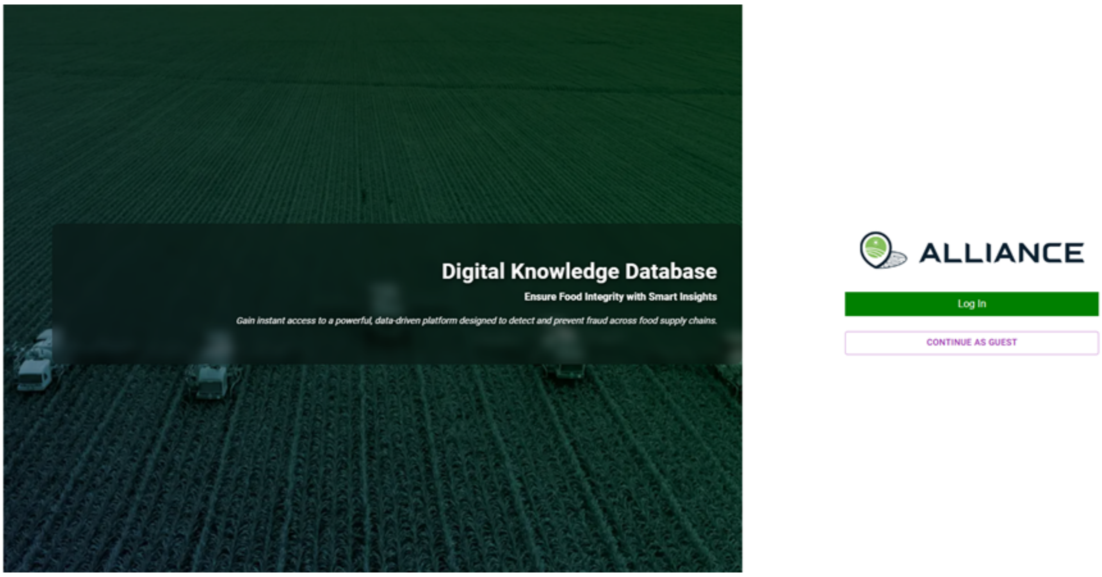
4.3 Development and Implementation

4.3.1 Technology Stack

The backend is developed using FastAPI, while the frontend is implemented in React. Hosting and deployment utilize AWS services, supporting scalability, cloud storage, and secure data access. The system employs a database for indexing and storing structured knowledge entries, as well as raw documents and associated metadata.

4.3.2 Current Status

The current implementation is a fully functioning prototype deployed online. It supports core functionalities including data upload, fraud tagging, search by category, and dynamic generation of result cards. The backend infrastructure manages structured storage of partner-submitted data, while the frontend offers an engaging, responsive experience. Ongoing development focuses on refining the search logic, enhancing performance, and expanding categorization coverage. Initial partner access has validated the basic flow of data navigation and retrieval, with additional refinements to be introduced as more data and use cases are incorporated. The platform remains under active iteration, allowing for the continuous integration of project outcomes and partner feedback.

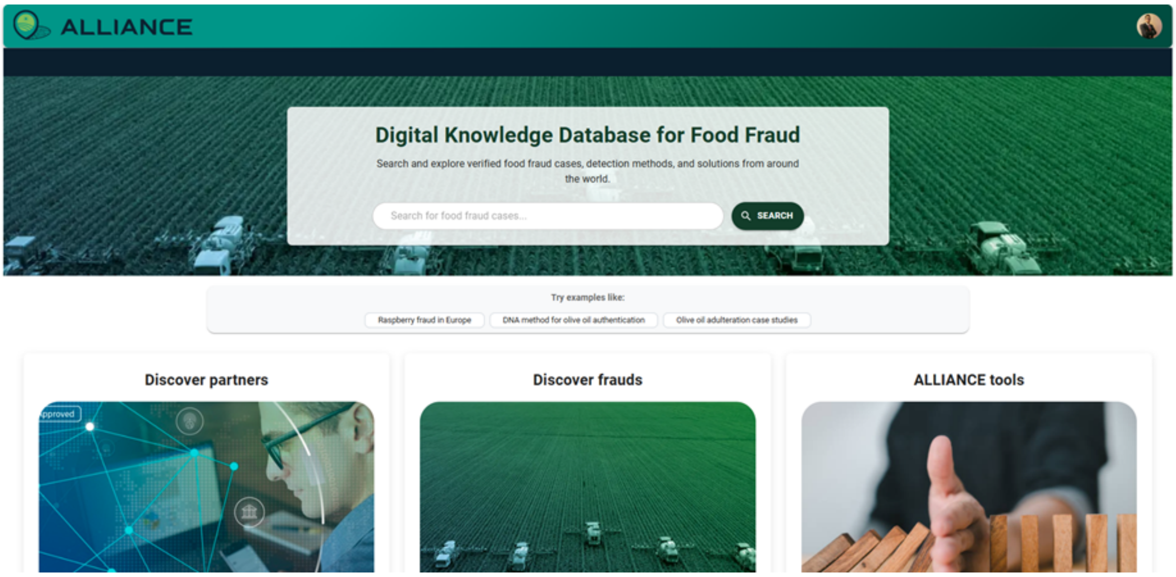


Digital Knowledge Database

Ensure Food Integrity with Smart Insights

Gain instant access to a powerful, data-driven platform designed to detect and prevent fraud across food supply chains.

Figure 4-1 Results display window for end-consumers for HSI technology.

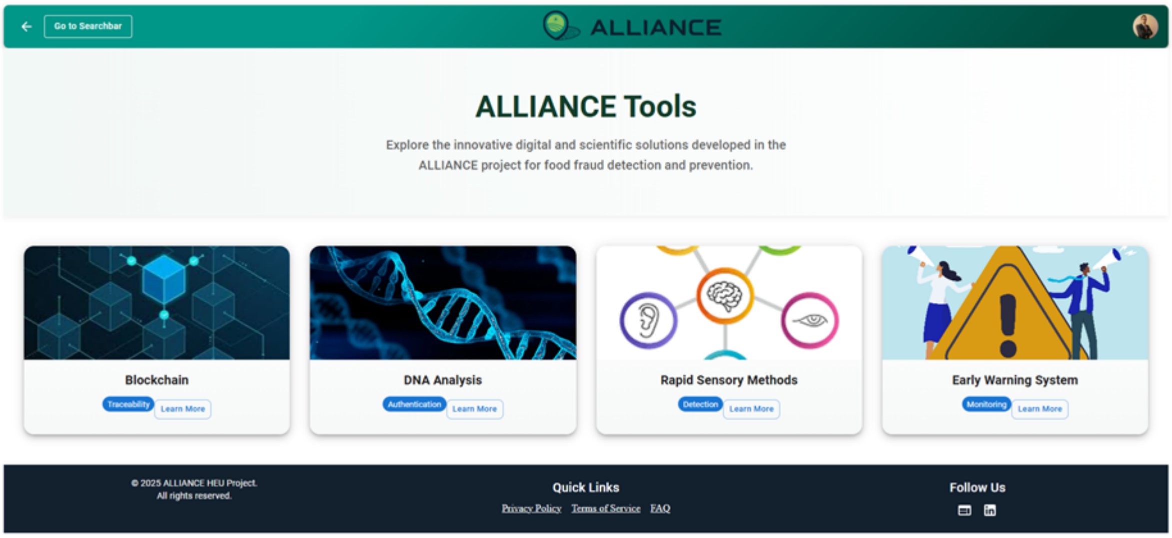


Digital Knowledge Database for Food Fraud

Search and explore verified food fraud cases, detection methods, and solutions from around the world.

ALLIANCE
Search for food fraud cases...
SEARCH
Try examples like:
Raspberry fraud in Europe
DNA method for olive oil authentication
Olive oil adulteration case studies
Discover partners
Discover frauds
ALLIANCE tools
pproved

Figure 4-2 Main dashboard showing categories and search bar functionality.



ALLIANCE Tools

Explore the innovative digital and scientific solutions developed in the ALLIANCE project for food fraud detection and prevention.

© 2025 ALLIANCE HEU Project.
All rights reserved.

Figure 4-3 Example of a fraud detection tool card with metadata and references.

4.4 Validation and final Implementation

Initial validation efforts have been carried out in collaboration with project partners who are contributing data to the system. Feedback was gathered through structured user testing sessions and direct interviews. These early adopters have tested core functionalities and offered insight into usability, navigation clarity, and content relevance, which have informed iterative design improvements. Use cases explored to date include detection of fraudulent substitution in specific product lines, browsing of solution cards related to analytical methods, and identification of regulatory documentation for protected food labels. These tests have informed UI improvements and refinement of content tagging. Future validation will include a broader user base and stress-testing of the search system under realistic data loads.

4.4.1 Key Results and Achievements

The development of the Digital Knowledge Base has yielded several important outcomes. A live, modular prototype has been delivered, capable of ingesting and displaying structured knowledge about food fraud cases and prevention tools. Data from internal project work packages and partner contributions are now harmonized and searchable. Integration with the early warning system framework has been initiated, establishing the foundation for risk scoring and predictive analytics. User experience design has successfully bridged technical complexity and intuitive exploration, making the tool accessible to various stakeholders.

4.4.2 Future Work and Sustainability

Looking ahead, several enhancements are planned to strengthen the platform's analytical power – particularly in the areas of data collection and analysis, sematic search, discovery of relationships within the data, as well as operational sustainability. Future developments include advanced semantic search using natural language understanding, automated tagging of uploaded documents, and network-based visualization of fraud relationships using graph databases. Further integration with real-time monitoring tools and external fraud alerts will improve responsiveness. To ensure long-term sustainability, governance and maintenance plans will be established, along with potential expansion toward other Protected Designation of Origin (PDO) and Protected Geographical Indication (PGI) products. This will position the Digital Knowledge Base as a cornerstone resource for combating food fraud across Europe and beyond.

5 Food Fraud Prevention with Predictive Analytics

5.1 Introduction

In Deliverable D3.2, we introduced the conceptual architecture for food fraud prevention through predictive analytics, highlighting key components such as prediction algorithms for identifying fraud incidents in the food supply chain, with a particular emphasis on explainability and trustworthiness. We also presented the potential of clustering algorithms designed to uncover hidden patterns among suppliers. As mentioned in D3.2, the architecture is supported by a robust technology stack, including database storage, Apache NiFi for data acquisition and ingestion, RESTful APIs for exposing prediction and clustering functionalities as services, and Superset for visual analytics.

In the current deliverable, we move beyond the conceptual design to describe the actual development work, deployed services, and the operational system. To demonstrate the system's functionalities, we created synthetic datasets simulating a feta cheese supply chain. These datasets were specifically generated for demonstration purposes and are intended to illustrate system capabilities; however, they come with natural limitations compared to real-world data. Although, the system is designed to empower stakeholders to utilize their own datasets to achieve more realistic and actionable insights.

5.2 Synthetic Dataset Creation for Demonstration

To effectively demonstrate the functionalities and capabilities of the developed system, a synthetic dataset covering four years was created. The dataset is structured around multiple variables related to the supply chain and quality control of feta cheese, with a focus on fraud detection. The table below presents a summary of each field in the dataset (Table 5.1).

Table 5.1 Data dictionary

Column Name	Description
collection_date	Date when the milk was collected from the supplier.
supplier_id	Unique identifier for the milk supplier.
milk_type	Type of milk supplied (sheep, goat). Each supplier delivers a specific type or both.
quantity	Amount of milk collected (liters), adjusted for ice bowl distribution.
fat	Fat content of the milk sample (%).
protein	Protein content of the milk sample (%).
truck_plate	Truck identifier that collected the milk. Each truck has multiple compartments.
route	The collection route (regional origin of milk).
compartment_id	The specific compartment in the truck where the milk was stored.
icebowl_id	The ice bowl where the milk was originally placed by the supplier before being transferred.
sample_barcode_comp	Unique barcode for the milk sample taken from the compartment.
sample_barcode_ice	Unique barcode for the milk sample taken from the ice bowl.
pH_comp	pH value of the milk sample from the compartment.
pH_ice	pH value of the milk sample from the ice bowl.
temperature_comp	Temperature of the milk in the compartment (°C).

temperature_ice	Temperature of the milk in the ice bowl (°C).
cow_fraud_comp	Indicator (0/1) for cow milk fraud in the compartment (adulteration with cow milk).
cow_fraud_ice	Indicator (0/1) for cow milk fraud in the ice bowl (used to estimate compartment fraud).
water_fraud_comp	Indicator (0/1) for water dilution fraud in the compartment.
water_fraud_ice	Indicator (0/1) for water dilution fraud in the ice bowl.
goat_fraud_comp	Indicator (0/1) for excessive goat milk fraud in the compartment.
goat_fraud_ice	Indicator (0/1) for excessive goat milk fraud in the ice bowl.
goat_percentage_comp	Percentage of goat milk detected in the compartment.
goat_percentage_ice	Percentage of goat milk detected in the ice bowl.

It should be mentioned that the data generation process for this dataset has been carefully designed to simulate the real-world complexities of milk collection, transportation, and fraud risks.

5.2.1 Supplier Generation

Each supplier in the dataset is assigned unique characteristics that reflect typical milk suppliers in the industry. These characteristics include geographical regions (routes) from which milk is sourced, milk type specialization, fraud risk profile, and the number of trucks used for transportation. To ensure balanced representation, 40% of suppliers are specialized in sheep milk, 30% in goat milk, and 30% handle both types of milk. Fraud risk is distributed as follows: 70% of suppliers are low-risk, 20% medium-risk, and 10% high-risk, based on industry standards, thus mirroring real-world conditions where low-risk suppliers are far more common than high-risk ones. In addition to this, each supplier is randomly assigned one to three trucks, contributing to the diversity of transportation patterns in the dataset.

5.2.2 Truck and Compartment Assignment

In total, the dataset includes 500 trucks, each with 20 compartments. These trucks collect milk from various suppliers who store their milk in ice bowls. To prevent bias, suppliers are assigned multiple trucks, ensuring there are no one-to-one correlations between a truck and a supplier. Milk from different suppliers may be mixed in the compartments, which are randomly selected for each delivery.

5.2.3 Ice bowl Assignment

Suppliers store their milk in randomly assigned rice bowls, with each supplier using between one and four rice bowls per collection. There is no strict capacity limit for the ice bowls, allowing suppliers to store as much milk as needed for each collection.

5.2.4 Collection Process

The collection process is randomized to reflect the natural variability of milk deliveries. Each supplier delivers milk between 30 and 50 times per year, with the amount of milk and its properties depending on the milk type. Sheep milk typically has a higher fat and protein content than goat milk. In the dataset, sheep milk contains 5.5% to 8.5% fat and 4.5% to 7.5% protein, while goat milk ranges from 3.5% to 6.5% fat and 3.0% to 6.0% protein. To maintain realistic mixing ratios, sheep milk should have minimal contamination with goat milk, and goat milk should contain at least 50% goat milk.

5.2.5 Fraud Injection

Fraud risks are introduced at both the ice bowl and compartment levels, simulating common types of fraud that occur in milk production, such as the addition of cow milk or water, or excessive goat milk in sheep milk. Fraud probabilities are linked to supplier risk levels: low-risk suppliers have a 5% chance of fraud, medium-risk suppliers have a 20% chance, and high-risk suppliers have a 50% chance. To prevent deterministic behaviour and introduce a degree of randomness, fraud probabilities are adjusted with normal distribution noise, ensuring a more realistic and varied distribution of fraud cases. Additionally, fraud events at the ice bowl and compartment levels are not fully correlated, simulating the uncertainty that occurs in real-world fraud detection.

5.2.6 Sample Collection and Testing

For each milk batch, two sample barcodes are assigned: one for the compartment and one for the ice bowl. The pH, temperature, and goat percentage of the milk are measured and vary slightly between the ice bowl and compartment samples due to differences in handling and storage. Temperature decreases as milk is stored longer, and the goat percentage fluctuates within a ±10% range due to mixing errors.

5.3 Development of Prediction Services

5.3.1 Data Processing and Feature Engineering

To prepare the synthetic dataset for model development and evaluation, the pre-processed data was first filtered based on the presence of specific fraud indicators (cow milk fraud, goat milk fraud, and water addition fraud). For each fraud type, datasets were created by aggregating and engineering relevant features, including milk quantity, quality parameters (fat, protein, pH, temperature), milk type breakdown, and temporal features (month, weekday). Geographical coordinates were also mapped based on collection areas. The following tables (Table 5.2-Table 5.6) summarize the key features created during the data processing, with a distinction between compartment-related features (denoted by "_comp") and ice bowl-related features (denoted by "_ice"), along with a brief description of each feature's purpose and role in the analysis.

Table 5.2 Aggregated Features

Feature	Description
total_quantity_of_goat_milk	Total quantity of goat milk collected in the compartment, aggregated by collection date and compartment.
total_quantity_of_sheep_milk	Total quantity of sheep milk collected in the compartment, aggregated by collection date and compartment.
area	The route/area where the milk was collected in the compartment, used as a categorical feature for geographic information.
ph_comp	The pH level of the milk in the compartment, used for quality control in the compartment.
temperature_comp	The temperature of the milk in the compartment during collection.
goat_percentage_comp	The percentage of goat milk in the compartment.

Table 5.3 Milk Quality Features (Ice bowl)

Feature	Description
avg_goat_milk_ph_ice	Average pH level of goat milk in the ice bowl during storage.
avg_sheep_milk_ph_ice	Average pH level of sheep milk in the ice bowl during storage.
avg_goat_milk_temperature_ice	The average temperature of goat milk in the ice bowl during storage.
avg_sheep_milk_temperature_ice	The average temperature of sheep milk in the ice bowl during storage.
avg_goat_milk_fat_ice	Average fat content of goat milk in the ice bowl during storage.
avg_sheep_milk_fat_ice	The average fat content of sheep milk in the ice bowl during storage.
min_goat_milk_fat_ice	Minimum fat content of goat milk in the ice bowl during storage.
min_sheep_milk_fat_ice	Minimum fat content of sheep milk in the ice bowl during storage.
max_goat_milk_fat_ice	Maximum fat content of goat milk in the ice bowl during storage.
max_sheep_milk_fat_ice	Maximum fat content of sheep milk in the ice bowl during storage.
avg_goat_milk_protein_ice	The average protein content of goat milk is in the ice bowl.
avg_sheep_milk_protein_ice	The average protein content of sheep milk is in the ice bowl.
min_goat_milk_protein_ice	Minimum protein content of goat milk in the ice bowl, indicative of quality or potential fraud in the milk supply.
min_sheep_milk_protein_ice	Minimum protein content of sheep milk in the ice bowl.
max_goat_milk_protein_ice	Maximum protein content of goat milk in the ice bowl, useful for identifying milk composition irregularities.
max_sheep_milk_protein_ice	Maximum protein content of sheep milk in the ice bowl, useful for fraud detection based on milk composition.

Table 5.4 Temporal Features

Feature	Description
collection_month	The month in which the milk was collected, derived from the collection date, and was used for seasonal analysis.
collection_weekday	The weekday (0=Monday, 6=Sunday) on which the milk was collected, useful for analyzing collection patterns.

Table 5.5 Geographical Features

Feature	Description
area_longitude	The longitude of the area where the milk was collected in the compartment, mapped to specific locations in the dataset.
area_latitude	The latitude of the area where the milk was collected in the compartment, mapped to specific locations in the dataset.

Table 5.6 Target Variable

Feature	Description
fraud_comp	The fraud indicator for each record in the compartment, indicating whether fraud was detected in the milk supply chain.

To improve model accuracy, the datasets for both machine learning and deep learning models were cleaned by removing irrelevant columns, encoding categorical variables, and standardizing numerical features. Each dataset, corresponding to the cow milk fraud, goat milk fraud, and water addition fraud use cases, was divided into separate subsets for training, validation, and testing. This ensures that the models are evaluated on unseen data to mitigate overfitting. Special attention was given to handling multicollinearity by identifying and removing highly correlated features (above a threshold of 0.9), thus enhancing model performance and reducing redundancy.

Eventually, the following features were selected for the training:

- total_quantity_of_goat_milk,
- total_quantity_of_sheep_milk,
- area,
- ph_comp,
- temperature_comp,
- goat_percentage_comp,
- avg_goat_milk_temperature_ice,
- avg_sheep_milk_temperature_ice,
- collection_month, and

- collection_weekday.

5.3.2 Machine Learning Training Process

The training process for machine learning models follows a series of structured steps, beginning with dataset pre-processing, model selection, and evaluation. To address class imbalance in the datasets, various resampling techniques such as class weighting (i.e., assign higher importance to the underrepresented class during model training by adjusting the loss function), under-sampling (i.e., reduce the number of samples from the overrepresented class to balance the class distribution in the dataset), and SMOTE[2] (i.e., generate synthetic samples for the minority class by interpolating between existing instances to address class imbalance) were applied[3]. Initially, the datasets are split into training, validation, and testing subsets to ensure models are evaluated on unseen data and to mitigate overfitting. The models trained include Logistic Regression, Random Forest, LightGBM, and Balanced Random Forest, which are selected for their versatility and ability to handle imbalanced datasets effectively[45].

Hyperparameter tuning was performed using GridSearchCV and StratifiedKFold cross-validation6. The parameter grids for each model were customized to optimize performance while considering computational efficiency. For example, hyperparameters like the number of estimators for Random Forest, the learning rate for LightGBM, and regularization parameters for Logistic Regression were tuned. During the training process, the models were evaluated using performance metrics such as AUC (Area Under Curve) and accuracy.

The following table (Table 5.7) presents the performance of various machine learning models on a synthetic dataset related to goat milk fraud detection. Notably, the algorithms achieved strong results across all methods, with LightGBM demonstrating exceptional performance. The table highlights the AUC and accuracy metrics for each model and method combination, showcasing the effectiveness of these approaches in addressing the goat fraud use case.

Table 5.7 Performance of machine learning models with different Imbalanced dataset handling methods

Machine Learning Model	Selected Method for treating Imbalanced Dataset	AUC	Accuracy	Execution Time (seconds)
Logistic Regression	Class Weighting	0.85	0.74	25.27
	Under-sampling	0.85	0.74	0.63
	SMOTE	0.85	0.75	20.69
Random Forest	Class Weighting	0.86	0.73	262.35
	Under-sampling	0.86	0.73	9.41
	SMOTE	0.86	0.73	481.75
LightGBM	Class Weighting	0.86	0.96	13.19
	Under-sampling	0.86	0.73	2.71
	SMOTE	0.86	0.80	32.04
Balanced Random Forest	Class Weighting	0.86	0.73	43.40
	Under-sampling	0.86	0.73	14.15

[2] Synthetic Minority Over-sampling Technique

[3] Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., & Bing, G. (2017). Learning from class-Imbalanced data: Review of methods and applications. *Expert Systems with Applications 73*, 220-239.

[4] Afane, K., & Zhao, Y. (2024). Selecting classifiers and resampling techniques for imbalanced datasets: A new perspective. *Procedia Computer Science 246*, 1150-1159.

[5] Awe, O., & Vance, A. (2025). *Practical statistical learning and data science methods: Case studies from LISA 2020 global network, USA*. Springer.

[6] Ou, G., Zhu, Z., Dong, B., & E, W. (2024). *Introduction to Data Science.* World Scientific Publishing.

	SMOTE	0.86	0.73	740.54

In addition to the performance table, ROC curves are presented for each combination of machine learning model and method used to handle imbalanced datasets (Figure 5-1-Figure 53?Figure 5-25).

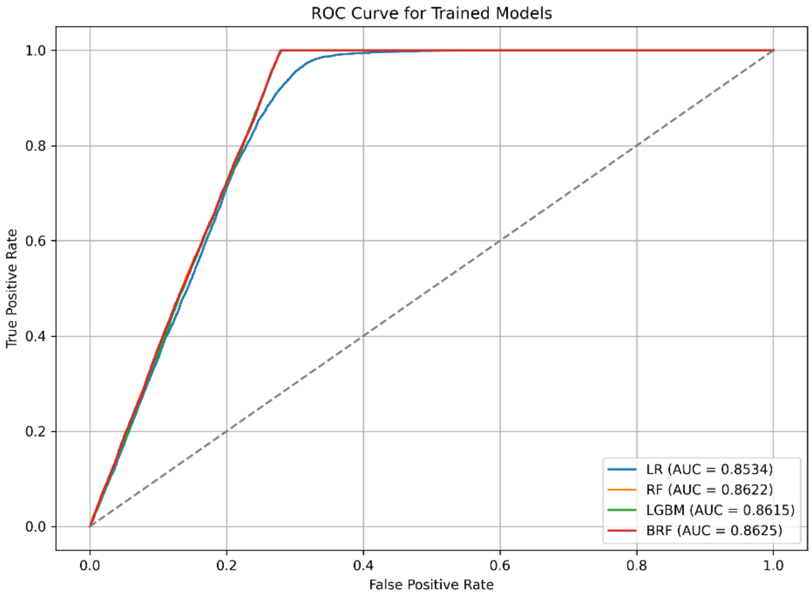


Figure 5-1 ROC curve for trained models (class-weighting)

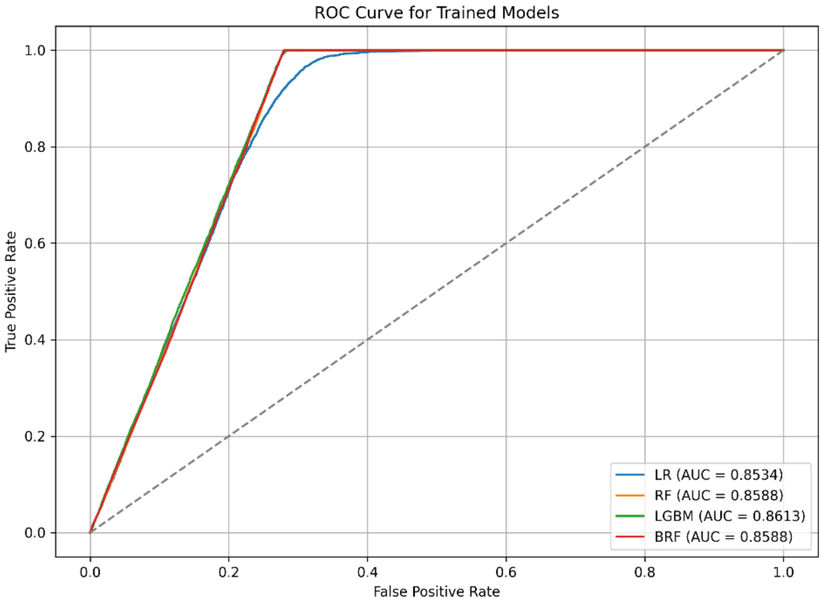


Figure 5-2 ROC curve for trained models (under-sampling)

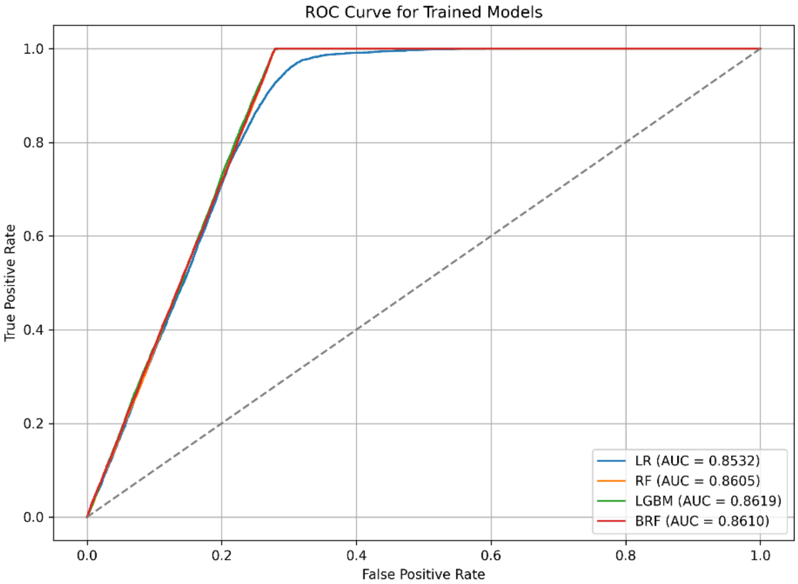


Figure 5-3 ROC curve for trained models (smote)

5.3.3 Deep Learning Model Training and Evaluation

The deep learning model is trained to detect fraudulent patterns in the dataset using neural network architecture. The process begins by preparing the dataset, which is split into three subsets: training, validation, and testing. In particular, the model is trained on one set of data, validated on a separate set to tune hyperparameters, and finally tested on unseen data to evaluate performance.

To address the class-imbalance present in the dataset, the model employs Focal Loss[7], a loss function specifically designed to down-weight the impact of easy-to-classify examples and focus more on harder, misclassified ones. The model is trained using the Adam optimizer[8], which adapts the learning rate for each parameter, thus improving training efficiency. During the training process, the model's performance is continually monitored through various metrics, such as accuracy and AUC. The model is then evaluated on the test dataset, and key metrics are extracted and recorded.

The results obtained from the model using the same dataset mentioned in Sub-section 5.3.3 were as follows:

- **Accuracy**: 0.73
- **AUC:** 0.86

These results indicate that the deep learning model achieved a reasonable level of performance in distinguishing fraudulent from non-fraudulent instances, with an accuracy of approximately 73%. Furthermore, the high AUC value (0.86) reflects the model's strong ability to correctly

[7] Mukhoti, J., Kulharia, V., Sanyal, A., Golodetz, S., Torr, P. H. S., & Dokania, P. K. (2020). Calibrating deep neural networks using focal loss. NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing System. Article No.: 1282, pp. 15288-15299.
[8] Hossain, E. (2024). *Machine learning crash course for engineers*. Springer.

classify the positive (fraudulent) and negative (non-fraudulent) classes, even in the presence of class imbalance.

5.3.4 Inference and Model Explainability to understand Predictions and Feature Contributions

In the inference phase, after training the machine learning or deep learning models, the next critical step is to predict the probability and classification indicators for new, unseen data. This phase also involves generating explainability graphs to better understand the model's decision-making process.

5.3.4.1 Prediction of Probabilities and Indicators

In the inference phase, for each dataset (e.g., cow fraud, goat fraud, and water fraud), the trained models are used to predict the probabilities and indicators of fraud. Specifically:

- **Probabilities**: The model outputs a probability value between 0 and 1, which indicates the likelihood that a given instance belongs to the positive class (fraud).
- **Indicators**: A binary classification decision is made by thresholding the predicted probabilities at 0.5. If the probability is greater than or equal to 0.5, the instance is classified as fraudulent (indicator = 1), and if it is lower, the instance is classified as non-fraudulent (indicator = 0).

5.3.4.2 SHAP (Shapley Additive Explanations) for Model Interpretability

To understand how different features contribute to the fraud detection decisions, SHAP values[9] are computed. Then, we create the following plots to highlight the importance of features based on computed SHAP values:

- **SHAP Summary Plot**: This plot provides a global view of feature importance across all instances in the dataset. It displays how each feature's SHAP value varies for different predictions (Figure 5-4).
- **SHAP Bar Plot**: Similar to the summary plot, the SHAP bar plot presents the average SHAP value for each feature. It visually emphasizes which features are most significant in influencing the predictions. The higher the SHAP value, the more influential the feature is in making the final prediction. The longer the bar for a feature, the more it contributes to the prediction of fraud (Figure 5-5-Figure 5-6).

[9] SHAP values represent the contribution of each feature to the model's prediction for each individual instance.

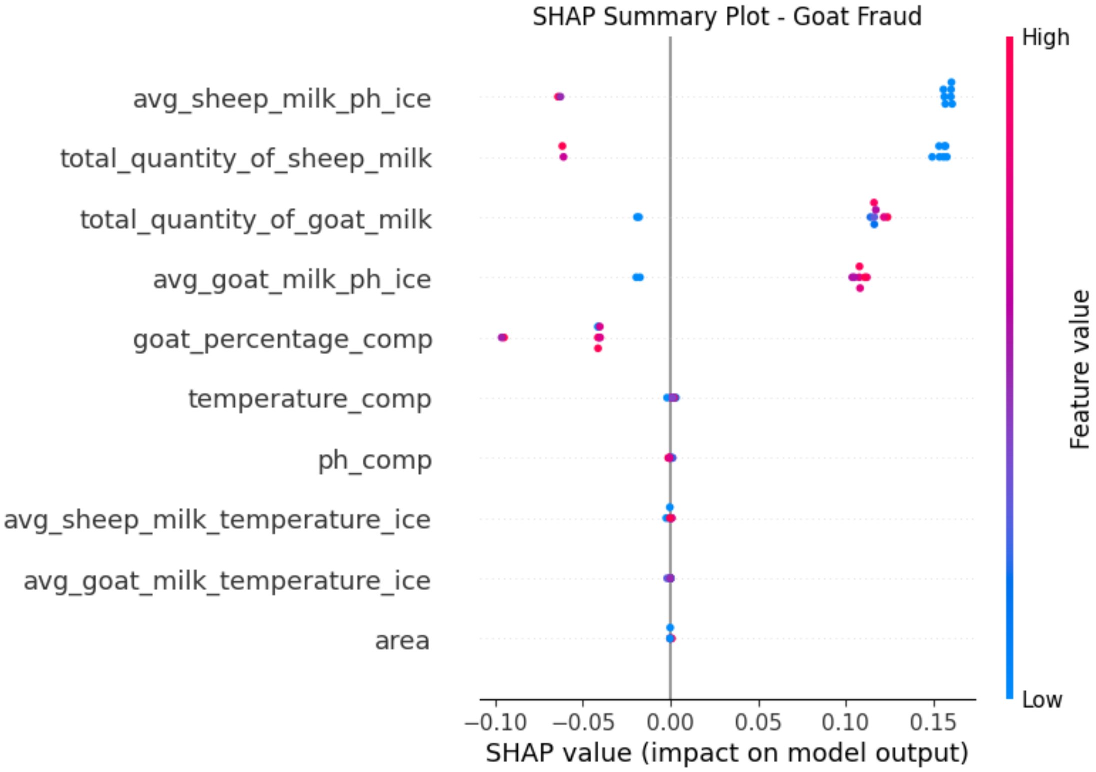


Figure 5-4 SHAP summary plot

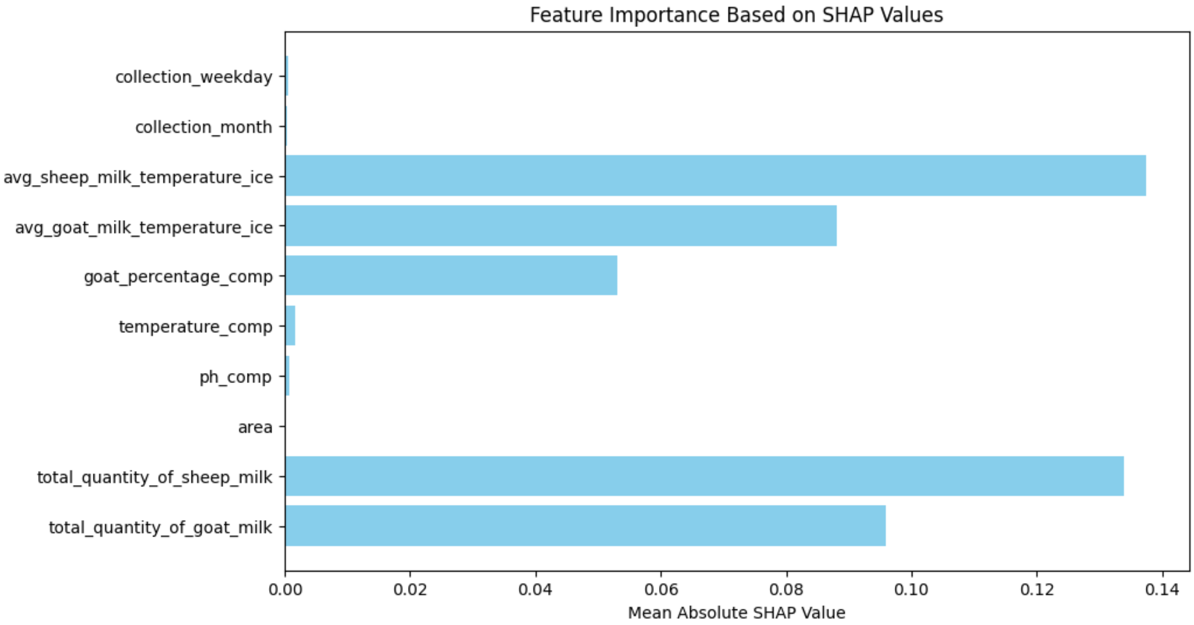


Figure 5-5 SHAP bar plot

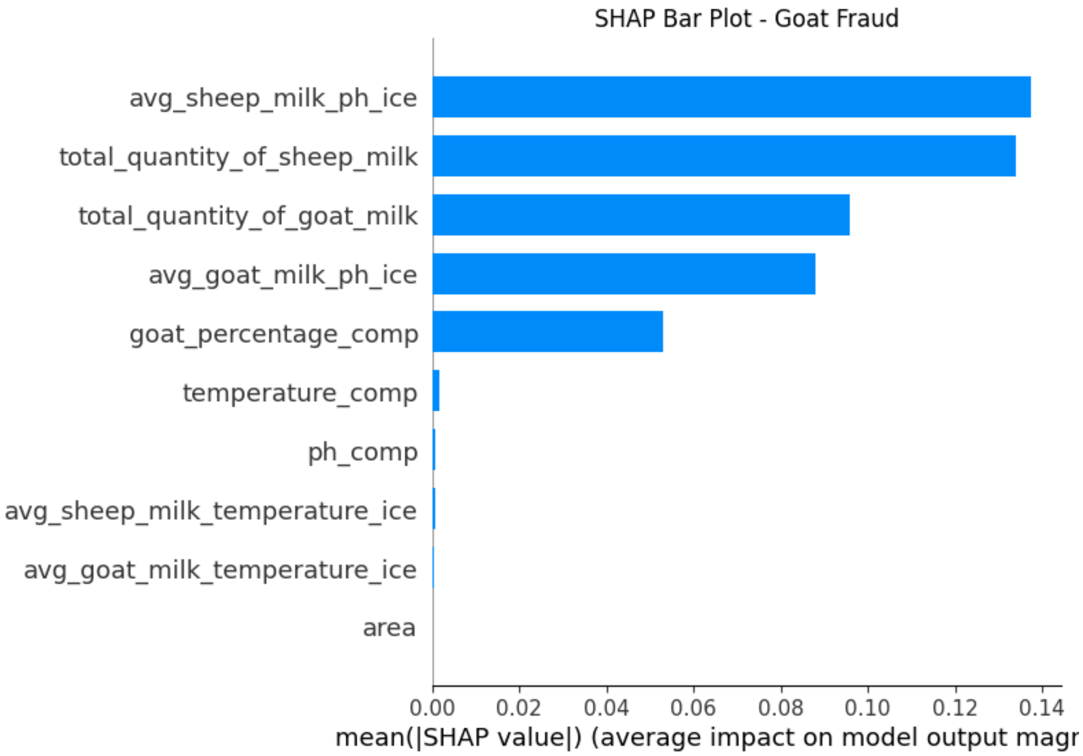


Figure 5-6 SHAP bar plot (sorting features in descending order according to their SHAP values)

5.3.4.3 Logistic Regression Feature Analysis

For the logistic regression model, feature coefficients are plotted to understand how each feature impacts the decision. Logistic regression models assign a weight (coefficient) to each feature, which determines its influence on the model's output. A positive coefficient increases the likelihood of fraud, while a negative coefficient decreases it.

- **Feature Coefficients Plot**: A bar plot of feature coefficients to visualize the strength and direction of each feature's influence. Larger absolute values indicate more influence, while the sign (positive or negative) shows whether the feature increases or decreases the probability of fraud (Figure 5-7).

- **Mean Contribution Bar Plot**: For each feature, the mean contribution across all instances is calculated, and a bar plot is generated to visualize this. The mean contribution provides insight into how much each feature, on average, contributes to the model's decision for each instance in the dataset (Figure 5-8).

- **Box Plot of Contributions**: This plot shows the distribution of feature contributions for each feature, providing insight into the spread and variability of contributions across different instances. It highlights outliers or instances where certain features contribute significantly to the prediction (Figure 5-9).

- **Feature Contribution Heatmap**: A heatmap is used to visualize the contribution of features across all instances. This plot helps identify patterns in feature contributions and can reveal interesting correlations or trends in how different features influence predictions (Figure 5-10).

- **Probability Evolution Plot**: This plot shows how predicted probabilities evolve based on the total contributions (log-odds) of all features. It provides a clearer picture of how specific contributions from each feature affect the predicted probability of fraud for individual instances (Figure 5-11).

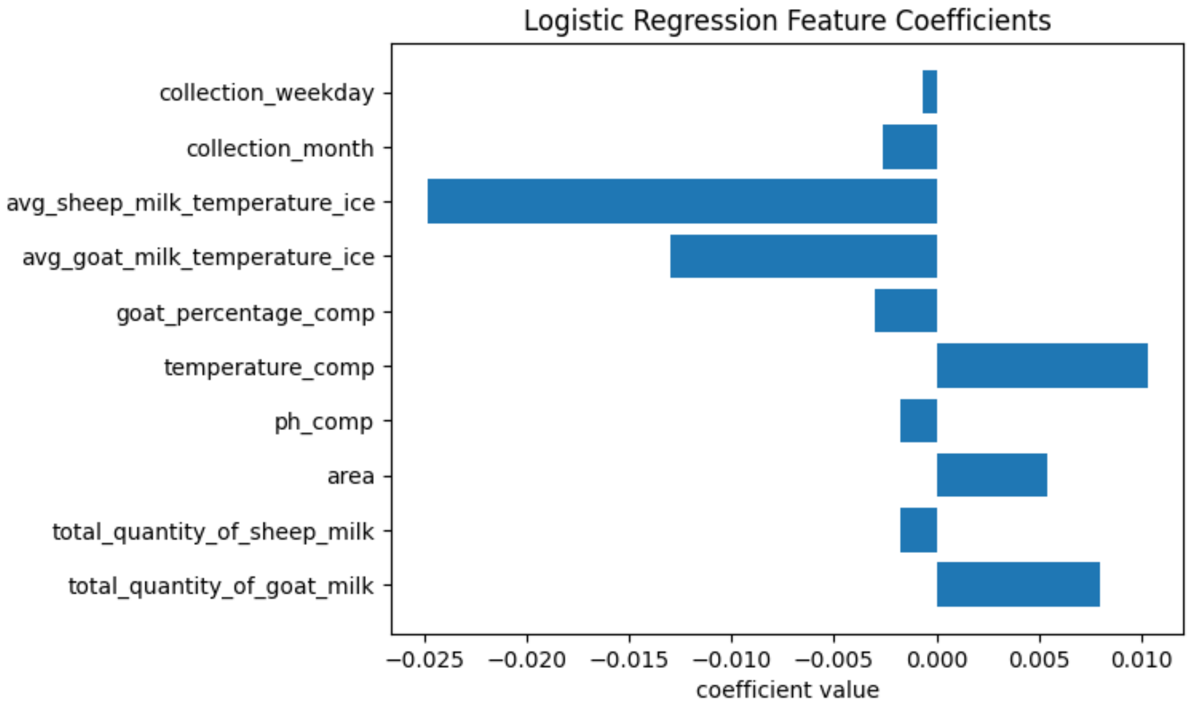


Figure 5-7 Feature coefficients plot

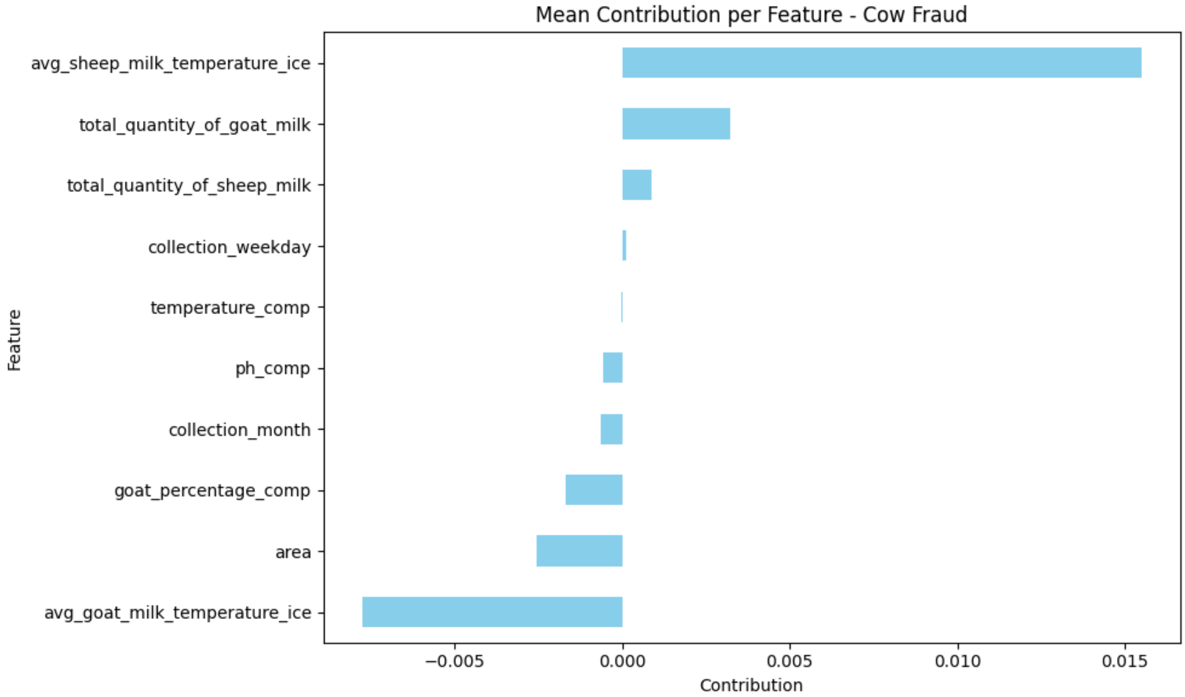


Figure 5-8 Mean contribution bar plot

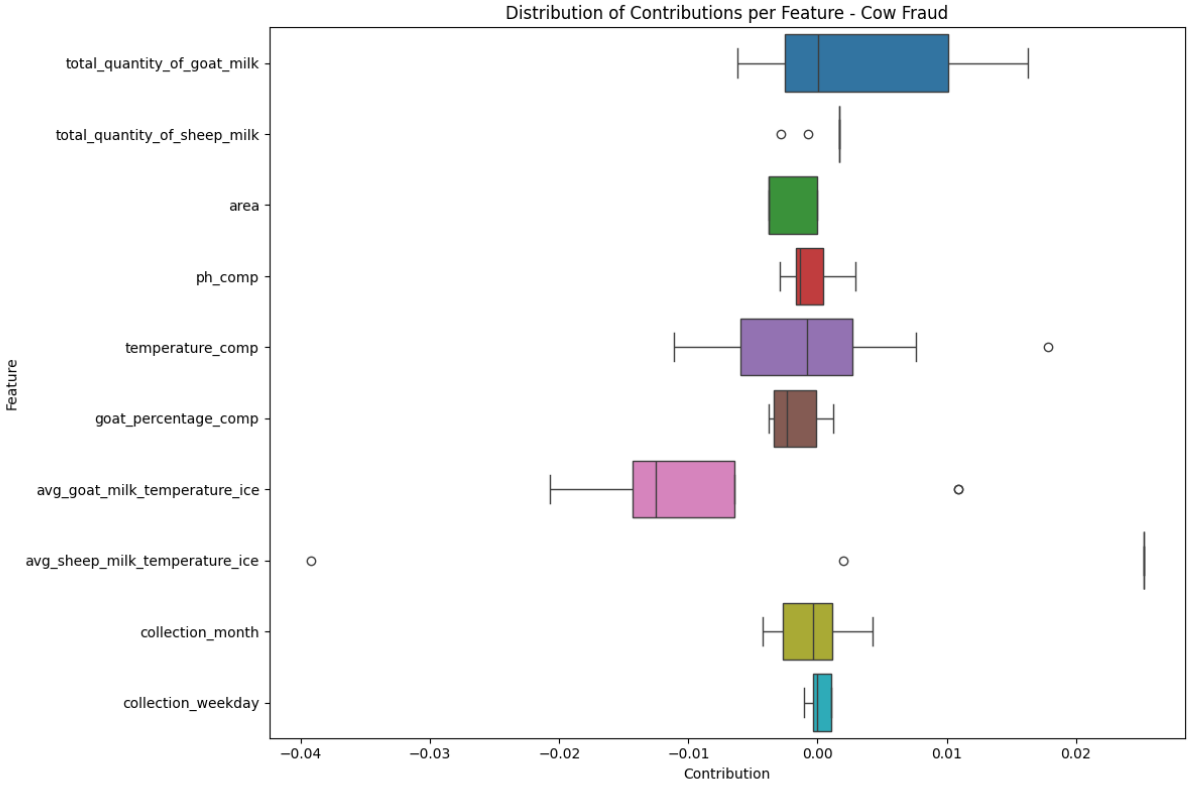


Figure 5-9 Boxplot of contributions

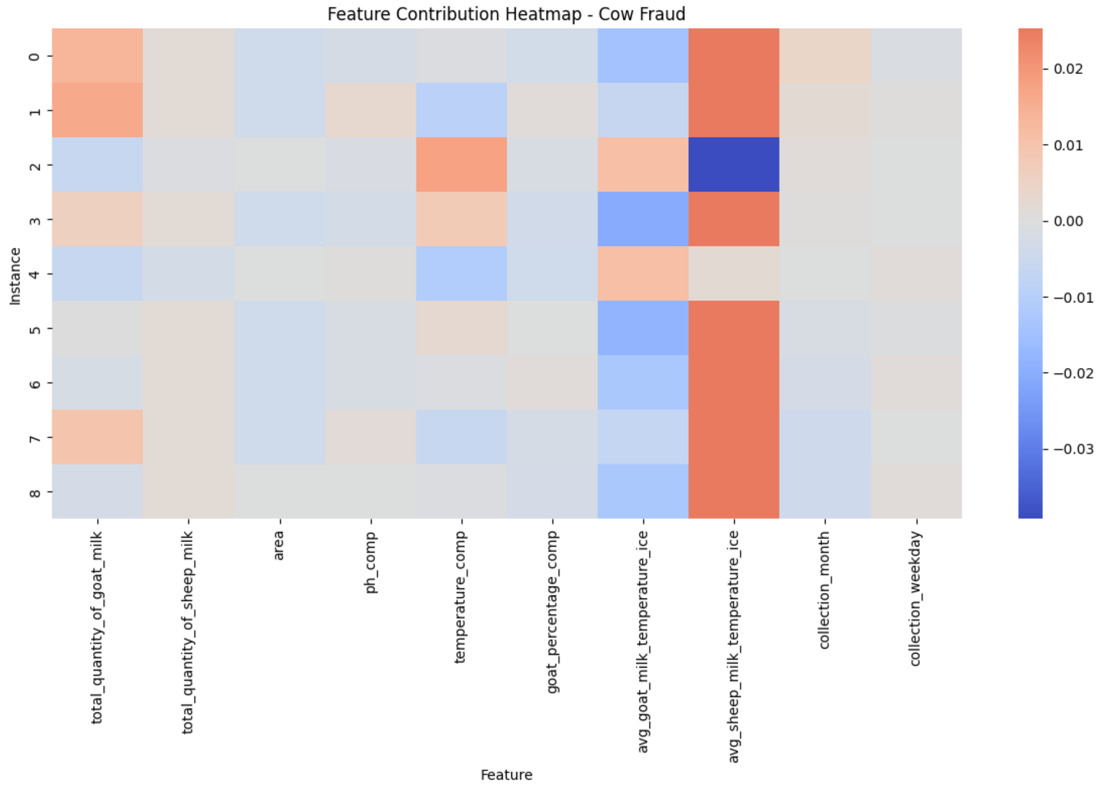


Figure 5-10 Feature contribution heatmap

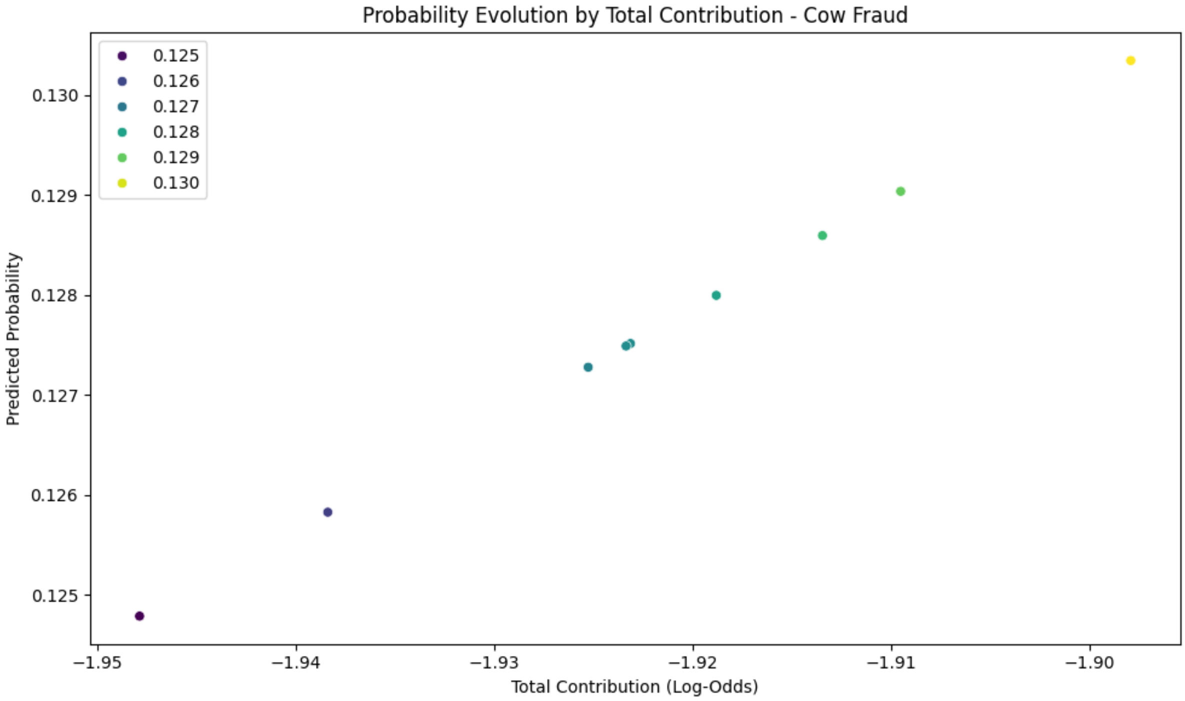


Figure 5-11 Probability evolution plot

5.4 Clustering Services for Supplier Risk Analysis

To uncover hidden patterns and better understand the behavior of milk suppliers, a clustering approach was developed using a set of carefully engineered features derived from raw milk collection data. These features included:

- **Milk Quantity Metrics**: Average, maximum, and minimum quantities supplied for both goat and sheep milk.
- **Milk Composition Metrics**: Average, maximum, and minimum fat and protein content values for both types of milk.
- **Milk pH and Temperature Metrics**: Average, maximum, and minimum pH levels and temperatures recorded.
- **Fraud Detection Indicators**: Counts of detected fraud incidents, such as the presence of cow or goat milk in sheep milk, or dilution with water, highlighting the authenticity risks for each supplier.

The goal of the clustering analysis was to group milk suppliers into homogeneous clusters based on their quantity patterns, milk composition metrics, and historical fraud incidents. The process involved the following steps:

1. **Data Preparation:** First, only numerical features were selected from the dataset, including quantities, fat content, protein content, pH, temperature measurements, and fraud incident counts. These features were standardized to ensure all variables contributed equally to the clustering, regardless of their original scale.
2. **Determining the Optimal Number of Clusters:** A K-Means clustering algorithm was used to group suppliers. To select the best number of clusters (K), K-Means models were fitted for different values of K (from 2 to 10). For each K, the Silhouette Score was computed to evaluate how well suppliers fit within their assigned clusters versus others.
3. **Final Clustering:** After identifying the best K, the final K-Means model was trained on the standardized data. The resulting cluster labels were assigned to each supplier and added to the dataset (see Figure 5-12).
4. **Visualization of Clusters:** To visualize the clustering results, the high-dimensional data was reduced to two principal components using Principal Component Analysis (PCA) (see Figure 5-13).
5. **Cluster Profiling and Insights:** Detailed analysis was performed to characterize each cluster:
   - The average goat and sheep milk quantities per cluster were computed and compared using bar charts (see Figure 5-14).
   - The average number of detected fraud incidents (cow fraud in sheep milk, water fraud in sheep milk, goat fraud in sheep milk, etc.) was analyzed across clusters using boxplots (see Figure 5-15), barplots with standard deviations (see Figure 5-16), heatmaps (see Figure 5-17), and radar charts (see Figure 5-18).

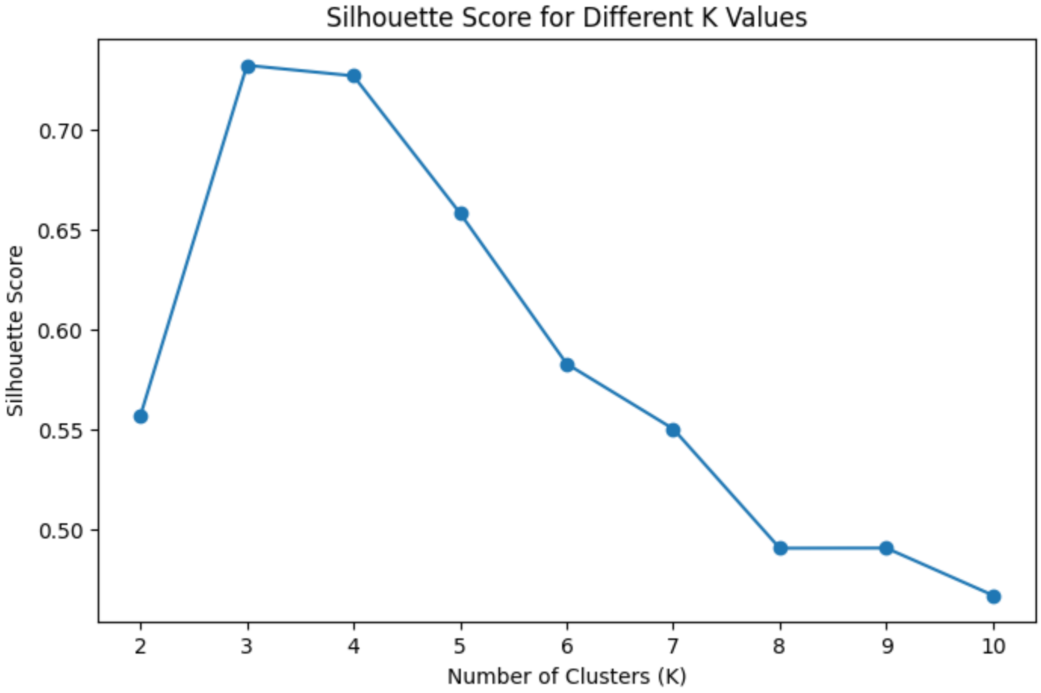


Figure 5-12 Silhouette score for different K values

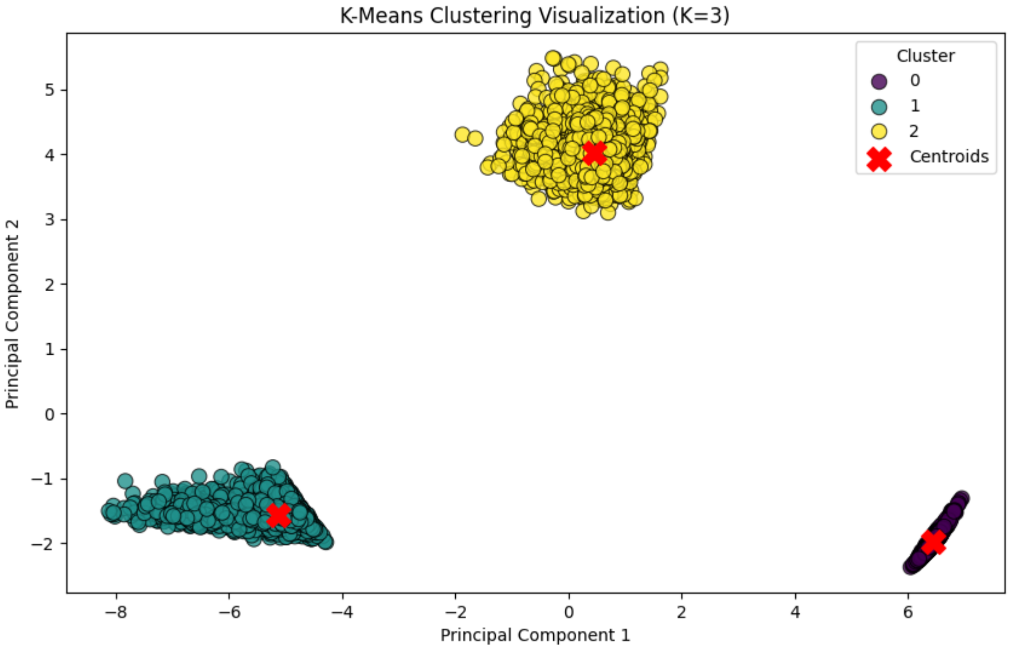


Figure 5-13 Clustering result

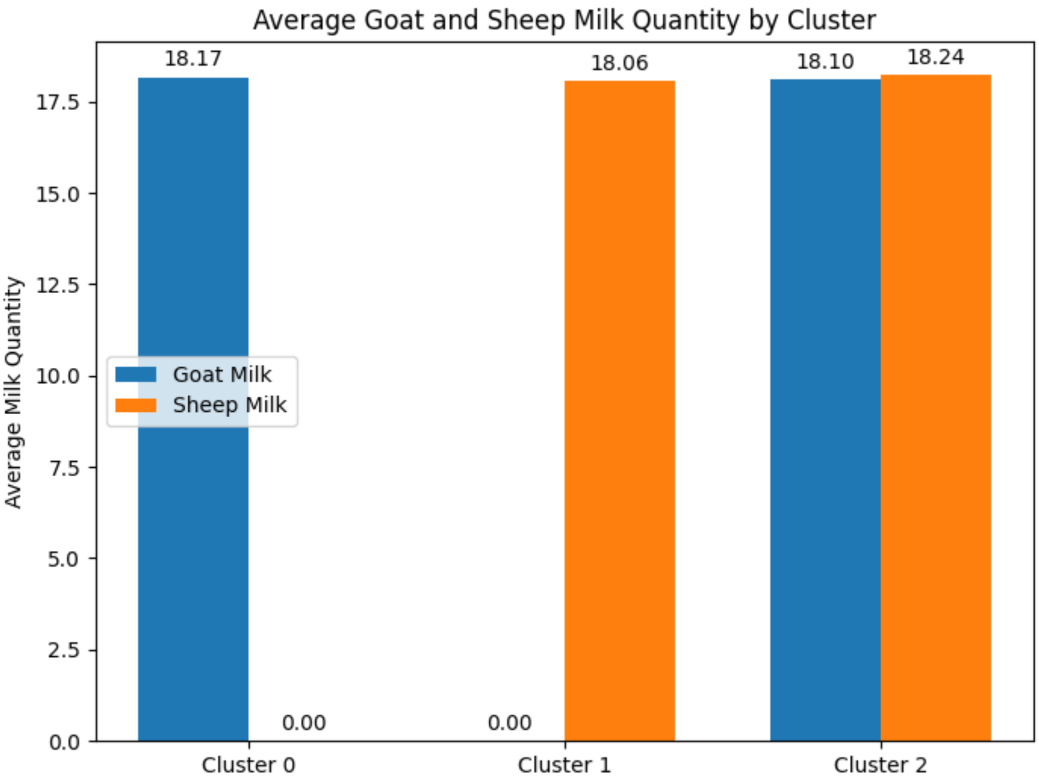


Figure 5-14 Average goat and sheep milk quantity by cluster

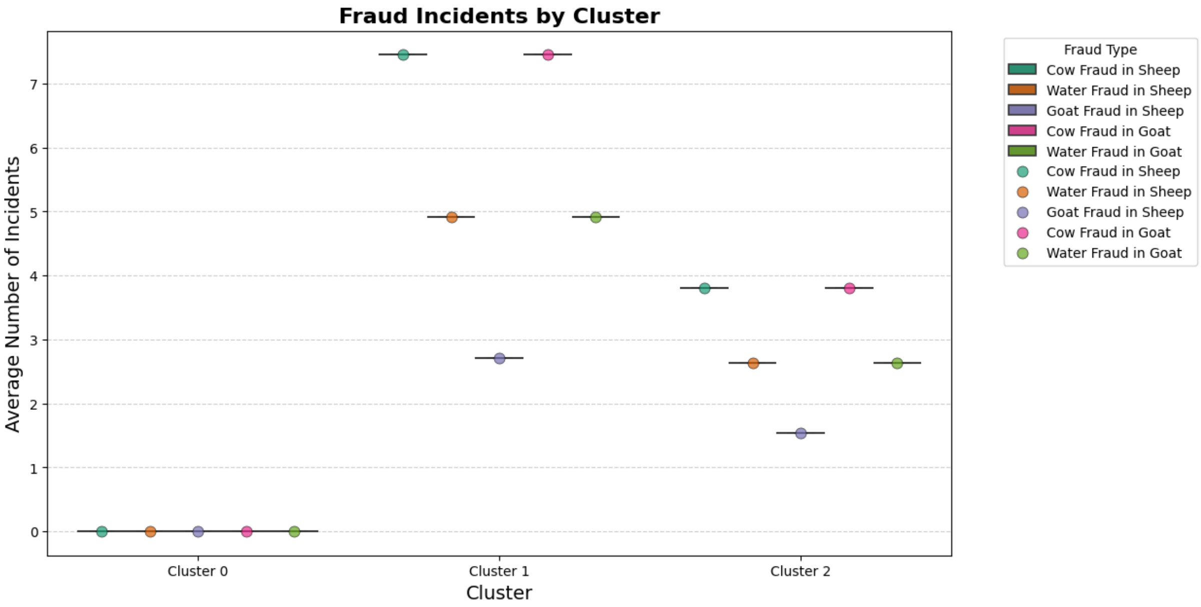


Figure 5-15 Boxplot depicting fraud incidents by cluster

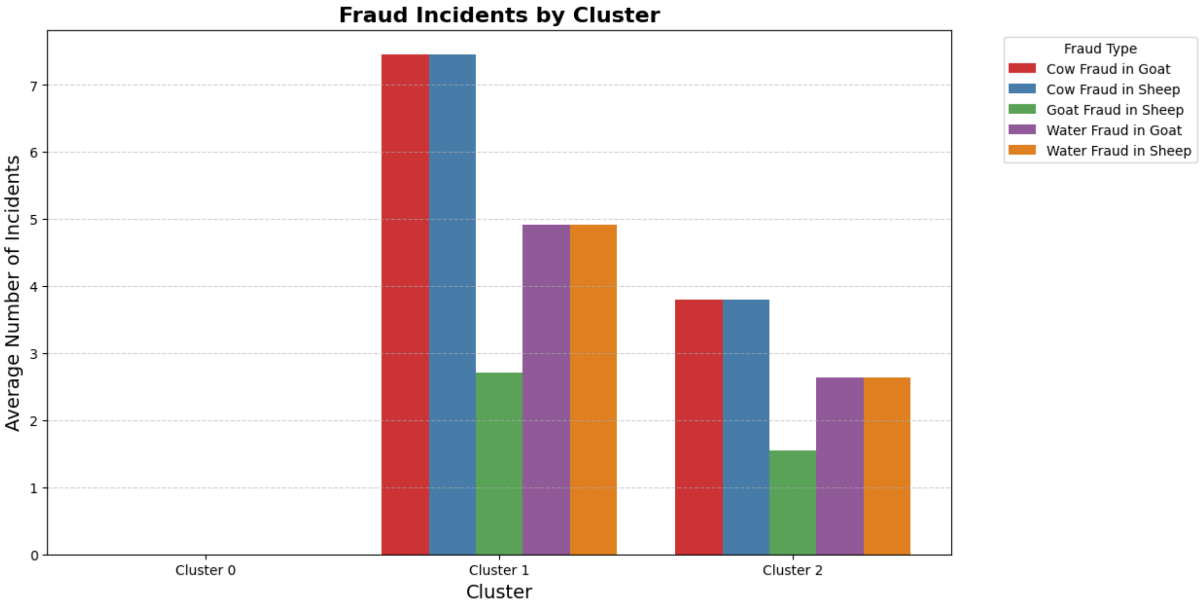


Figure 5-16 Bar plot depicting fraud incidents by cluster

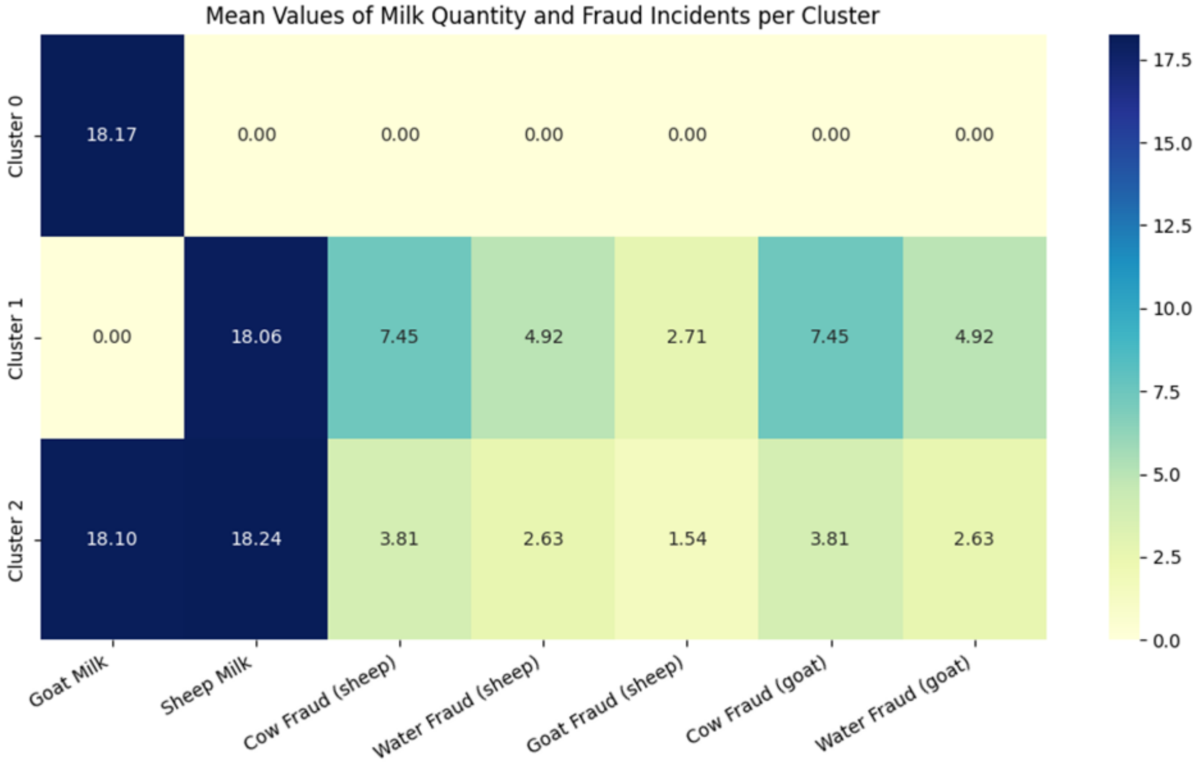


Figure 5-17 Mean values of milk quantity and fraud incidents per cluster

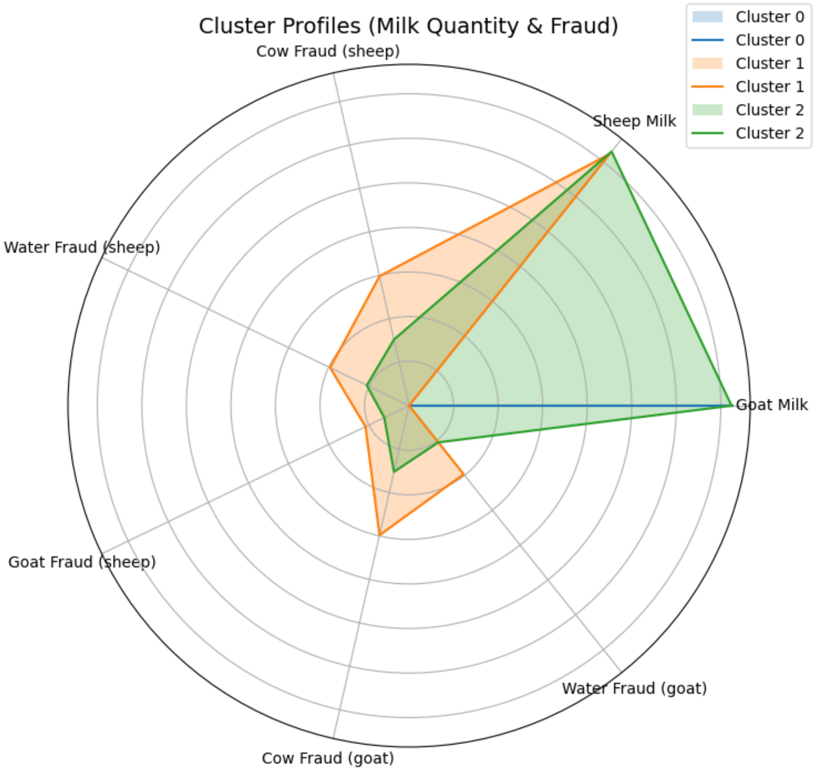


Figure 5-18 Bar plot depicting fraud incidents by cluster

The clustering analysis revealed three distinct groups of milk suppliers, each demonstrating unique profiles in terms of milk type, quantity, and incidence of fraud.

- **Cluster 0 (goat milk, no fraud)** consists of suppliers who exclusively provide goat milk, with an average quantity of 18.17 liters and very low variability (standard deviation of 1.88). No sheep milk was supplied in this cluster, and crucially, no fraud incidents of any kind were detected. This indicates a group of highly reliable and consistent goat milk suppliers.
- **Cluster 1 (sheep milk, high fraud)** is made up entirely of sheep milk suppliers, with an average quantity of 18.06 liters (standard deviation of 1.82). However, this cluster is strongly associated with high levels of fraud incidents, including cow milk contamination and water adulteration. Fraud levels are significant, with mean values such as 7.45 incidents for cow fraud and 4.92 incidents for water fraud in sheep milk, both accompanied by high variability. The cluster represents a high-risk group for producers focused on pure sheep milk products and highlights the need for careful supplier vetting and rigorous testing protocols.
- **Cluster 2 (mixed milk, moderate fraud)** represents suppliers who provide both goat and sheep milk in nearly equal quantities (around 18 liters each, with slightly higher standard deviations of about 2.6 to 2.7). Fraud incidents are present in this cluster as well, but at moderate levels compared to Cluster 1. The mean incidents of cow fraud and water fraud are lower (around 3.81 and 2.63, respectively) and are associated with more moderate variability. Suppliers in Cluster 2 could be acceptable partners for operations

that use mixed milk, provided there is regular quality control to detect and address any inconsistencies.

5.5 Technology Stack and Infrastructure

5.5.1 Data Collection and Integration

To serve the purpose of developing an end-to-end food fraud prevention system with predictive analytics, a multi-stage data pipeline has been formulated, originating in the acquisition of raw data directly from the various stages of the supply chain and leading all the way up to the creation of concise and meaningful insights based on highly curated information.

5.5.1.1 Storage layer - database

At the heart of the developed system lies the storage layer, which serves as a single point of reference, both for the incoming raw data, as well as for the results produced by the implemented pipeline. Data is stored into a relational database, in several tables insightfully structured in order to best serve the purpose of the system. MariaDB was the RDBMS of choice, as a reliable and well-known open-source option.

The various tables used in the implemented architecture, along with their schemas, are provided below.

milk_info

The Table 5.8 containing the raw data, as obtained by the blockchain network. Each row represents a milk batch received by a supplier on a given day. This table is also enhanced with the prediction results generated by the Deep Learning Models and serves as an aggregated point of reference for any reference.

Table 5.8 Milk info data

Field Name	Type	Description
id	bigint(20)	Primary key, auto incrementing
collection_date	datetime	Date that the batch was collected
supplier_id	varchar(16)	Id of the supplier
milk_type	varchar(8)	Type of the milk received (cow/sheep)
quantity	float	Quantity of the milk received
fat	float	Fat percentage of the milk batch
protein	float	Protein percentage of the milk batch
truck_plate	varchar(8)	Plate of the truck collecting the milk
route	varchar(16)	Route followed
compartment_id	varchar(16)	Compartment id on the given truck in which the milk was poured
icebowl_id	varchar(8)	Id of the icebowl in which the milk was received
sample_barcode_comp	varchar(8)	Barcode of the sample collected from the compartment
sample_barcode_ice	varchar(8)	Barcode of the sample collected from the icebowl
ph_comp	float	Ph of the compartment
ph_ice	float	Ph of the icebowl
temperature_comp	float	Temperature of the compartment

temperature_ice	float	Temperature of the icebowl
cow_fraud_comp	int(11)	Cow fraud indication in the compartment
cow_fraud_ice	int(11)	Cow fraud indication in the icebowl
water_fraud_comp	int(11)	Water fraud indication in the compartment
water_fraud_ice	int(11)	Waterfraud indication in the icebowl
goat_fraud_comp	int(11)	Goat fraud indication in the compartment
goat_fraud_ice	int(11)	Goat fraud indication in the icebowl
goat_percentage_comp	float	In case of goat fraud, percentage of the goat milk in the compartment
goat_percentage_ice	float	In case of goat fraud, percentage of the goat milk in the icebowl
collection_year	int(11)	Collection year
createdAt	datetime	Timestamp of the moment the record was inserted in the blockchain network
updatedAt	datetime	Timestamp of the moment the record was last updated in the blockchain network
total_quantity_of_goat_milk_comp	float	Total quantity of goat milk in the respective compartment
total_quantity_of_sheep_milk_comp	float	Total quantity of sheep milk in the respective compartment
avg_goat_milk_ph_ice	float	Average ph of the goat milk on the respective compartment
avg_sheep_milk_ph_ice	float	Average ph of the sheep milk in the compartment
avg_goat_milk_temperature_ice	float	Average temp of the goat milk in the respective compartment
avg_sheep_milk_temperature_ice	float	Average temp of the sheep milk in the respective compartment
avg_goat_milk_fat_ice	float	Average fat percentage of the goat milk in the icebowls that were poured into the respective compartment
avg_sheep_milk_fat_ice	float	Average fat percentage of the sheep milk in the icebowls that were poured into the respective compartment
min_goat_milk_fat_ice	float	Minimum fat percentage of the goat milk in the icebowls that were poured into the respective compartment
min_sheep_milk_fat_ice	float	Minimum fat percentage of the sheep milk in the icebowls that were poured into the respective compartment
max_goat_milk_fat_ice	float	Maximum fat percentage of the goat milk in the icebowls that were poured into the respective compartment

max_sheep_milk_fat_ice	float	Maximum fat percentage of the sheep milk in the icebowls that were poured into the respective compartment
avg_goat_milk_protein_ice	float	Average protein percentage of the goat milk in the icebowls that were poured into the respective compartment
avg_sheep_milk_protein_ice	float	Average protein percentage of the sheep milk in the icebowls that were poured into the respective compartment
min_goat_milk_protein_ince	float	Minimum protein percentage of the goat milk in the icebowls that were poured into the respective compartment
min_sheep_milk_protein_ince	float	Minimum protein percentage of the sheep milk in the icebowls that were poured into the respective compartment
max_goat_milk_protein_ice	float	Maximum protein percentage of the goat milk in the icebowls that were poured into the respective compartment
max_sheep_milk_protein_ice	float	Maximum protein percentage of the sheep milk in the icebowls that were poured into the respective compartment
collection_month	int(11)	Month of the collection
collection_weekday	int(11)	Weekday of the collection
area_longitude	float	Longitude value representing the area that was covered during this route
area_latitude	float	Latitude value representing the area that was covered during this route
cow_fraud_probability_comp	float	Probability of cow fraud in the respective compartment, as calculated by the deep learning model
cow_fraud_indicator_comp	int(11)	Cow fraud presence indicator in the respective compartment, as calculated by the deep learning model
goat_fraud_probability_comp	float	Probability of goat fraud in the respective compartment, as calculated by the deep learning model
goat_fraud_indicator_comp	int(11)	Goat fraud presence indicator in the respective compartment, as calculated by the deep learning model
water_fraud_probability_comp	float	Probability of water fraud in the respective compartment, as

		calculated by the deep learning model
water_fraud_indicator_comp	int(11)	Water fraud presence indicator in the respective compartment, as calculated by the deep learning model

cow_fraud_prediction_data

The Table 5.9 containing the data related to cow fraud prediction, as obtained by the deep learning model. Each row represents the contents of a single compartment of a company truck on a given day. The table is also enhanced with the actual fraud value obtained by the blockchain, as a way to validate the predictions inferenced.

Table 5.9 Cow fraud prediction data

Field Name	Type	Description
collection_date	datetime	Date that the compartment was filled
compartment_id	varchar(255)	Compartment id of interest
total_quantity_of_goat_milk	float	Total quantity of goat milk in the compartment
total_quantity_of_sheep_milk	float	Total quantity of sheep milk in the compartment
area	varchar(255)	Geographical area of the milk collection
ph_comp	float	Ph of the compartment
temperature_comp	float	Temperature of the compartment
goat_percentage_comp	float	Percentage of goat milk in the compartment
avg_goat_milk_ph_ice	float	Average ph of the goat milk on the compartment
avg_sheep_milk_ph_ice	float	Average ph of the sheep milk in the compartment
avg_goat_milk_temperature_ice	float	Average temp of the goat milk in the compartment
avg_sheep_milk_temperature_ice	float	Average temp of the sheep milk in the compartment
avg_goat_milk_fat_ice	float	Average fat percentage of the goat milk in the icebowls that were poured into the compartment
avg_sheep_milk_fat_ice	float	Average fat percentage of the sheep milk in the icebowls that were poured into the compartment
min_goat_milk_fat_ice	float	Minimum fat percentage of the goat milk in the icebowls that were poured into the compartment
min_sheep_milk_fat_ice	float	Minimum fat percentage of the sheep milk in the icebowls that were poured into the compartment
max_goat_milk_fat_ice	float	Maximum fat percentage of the goat milk in the icebowls that were poured into the compartment

max_sheep_milk_fat_ice	float	Maximum fat percentage of the sheep milk in the icebowls that were poured into the compartment
avg_goat_milk_protein_ice	float	Average protein percentage of the goat milk in the icebowls that were poured into the compartment
avg_sheep_milk_protein_ice	float	Average protein percentage of the sheep milk in the icebowls that were poured into the compartment
min_goat_milk_protein_ice	float	Minimum protein percentage of the goat milk in the icebowls that were poured into the compartment
min_sheep_milk_protein_ice	float	Minimum protein percentage of the sheep milk in the icebowls that were poured into the compartment
max_goat_milk_protein_ice	float	Maximum protein percentage of the goat milk in the icebowls that were poured into the compartment
max_sheep_milk_protein_ice	float	Maximum protein percentage of the sheep milk in the icebowls that were poured into the compartment
collection_month	int(11)	Month of the collection
collection_weekday	int(11)	Weekday of the collection
area_longitude	float	Longitude value representing the area that was covered during this route
area_latitude	float	Latitude value representing the area that was covered during this route
fraud_comp	int(11)	Fraud presence indicator, as obtained by the raw data
fraud_probability	float	Probability of fraud, as calculated by the deep learning model
fraud_indicator	int(11)	Fraud presence indicator, as calculated by the deep learning model

goat_fraud_prediction_data

The Table 5.10 containing the data related to goat fraud prediction, as obtained by the deep learning model. Each row represents the contents of a single compartment of a company truck on a given day. The table is also enhanced with the actual fraud value obtained by the blockchain, as a way to validate the predictions inferenced.

Table 5.10 Goat fraud prediction data

Field Name	Type	Description
collection_date	datetime	Date that the compartment was filled
compartment_id	varchar(255)	Compartment id of interest
total_quantity_of_goat_milk	float	Total quantity of goat milk in the compartment
total_quantity_of_sheep_milk	float	Total quantity of sheep milk in the compartment
area	varchar(255)	Geographical area of the milk collection

ph_comp	float	Ph of the compartment
temperature_comp	float	Temperature of the compartment
goat_percentage_comp	float	Percentage of goat milk in the compartment
avg_goat_milk_ph_ice	float	Average ph of the goat milk on the compartment
avg_sheep_milk_ph_ice	float	Average ph of the sheep milk in the compartment
avg_goat_milk_temperature_ice	float	Average temp of the goat milk in the compartment
avg_sheep_milk_temperature_ice	float	Average temp of the sheep milk in the compartment
avg_goat_milk_fat_ice	float	Average fat percentage of the goat milk in the icebowls that were poured into the compartment
avg_sheep_milk_fat_ice	float	Average fat percentage of the sheep milk in the icebowls that were poured into the compartment
min_goat_milk_fat_ice	float	Minimum fat percentage of the goat milk in the icebowls that were poured into the compartment
min_sheep_milk_fat_ice	float	Minimum fat percentage of the sheep milk in the icebowls that were poured into the compartment
max_goat_milk_fat_ice	float	Maximum fat percentage of the goat milk in the icebowls that were poured into the compartment
max_sheep_milk_fat_ice	float	Maximum fat percentage of the sheep milk in the icebowls that were poured into the compartment
avg_goat_milk_protein_ice	float	Average protein percentage of the goat milk in the icebowls that were poured into the compartment
avg_sheep_milk_protein_ice	float	Average protein percentage of the sheep milk in the icebowls that were poured into the compartment
min_goat_milk_protein_ice	float	Minimum protein percentage of the goat milk in the icebowls that were poured into the compartment
min_sheep_milk_protein_ice	float	Minimum protein percentage of the sheep milk in the icebowls that were poured into the compartment
max_goat_milk_protein_ice	float	Maximum protein percentage of the goat milk in the icebowls that were poured into the compartment
max_sheep_milk_protein_ice	float	Maximum protein percentage of the sheep milk in the icebowls that were poured into the compartment
collection_month	int(11)	Month of the collection
collection_weekday	int(11)	Weekday of the collection
area_longitude	float	Longitude value representing the area that was covered during this route

area_latitude	float	Latitude value representing the area that was covered during this route
fraud_comp	int(11)	Fraud presence indicator, as obtained by the raw data
fraud_probability	float	Probability of fraud, as calculated by the deep learning model
fraud_indicator	int(11)	Fraud presence indicator, as calculated by the deep learning model

water_fraud_prediction_data

The Table 5.11 containing the data related to water fraud prediction, as obtained by the deep learning model. Each row represents the contents of a single compartment of a company truck on a given day. The table is also enhanced with the actual fraud value obtained by the blockchain, as a way to validate the predictions inferenced.

Table 5.11 Water fraud prediction data

Field Name	Type	Description
collection_date	datetime	Date that the compartment was filled
compartment_id	varchar(255)	Compartment id of interest
total_quantity_of_goat_milk	float	Total quantity of goat milk in the compartment
total_quantity_of_sheep_milk	float	Total quantity of sheep milk in the compartment
area	varchar(255)	Geographical area of the milk collection
ph_comp	float	Ph of the compartment
temperature_comp	float	Temperature of the compartment
goat_percentage_comp	float	Percentage of goat milk in the compartment
avg_goat_milk_ph_ice	float	Average ph of the goat milk on the compartment
avg_sheep_milk_ph_ice	float	Average ph of the sheep milk in the compartment
avg_goat_milk_temperature_ice	float	Average temp of the goat milk in the compartment
avg_sheep_milk_temperature_ice	float	Average temp of the sheep milk in the compartment
avg_goat_milk_fat_ice	float	Average fat percentage of the goat milk in the icebowls that were poured into the compartment
avg_sheep_milk_fat_ice	float	Average fat percentage of the sheep milk in the icebowls that were poured into the compartment
min_goat_milk_fat_ice	float	Minimum fat percentage of the goat milk in the icebowls that were poured into the compartment
min_sheep_milk_fat_ice	float	Minimum fat percentage of the sheep milk in the icebowls that were poured into the compartment

max_goat_milk_fat_ice	float	Maximum fat percentage of the goat milk in the icebowls that were poured into the compartment
max_sheep_milk_fat_ice	float	Maximum fat percentage of the sheep milk in the icebowls that were poured into the compartment
avg_goat_milk_protein_ice	float	Average protein percentage of the goat milk in the icebowls that were poured into the compartment
avg_sheep_milk_protein_ice	float	Average protein percentage of the sheep milk in the icebowls that were poured into the compartment
min_goat_milk_protein_ice	float	Minimum protein percentage of the goat milk in the icebowls that were poured into the compartment
min_sheep_milk_protein_ice	float	Minimum protein percentage of the sheep milk in the icebowls that were poured into the compartment
max_goat_milk_protein_ice	float	Maximum protein percentage of the goat milk in the icebowls that were poured into the compartment
max_sheep_milk_protein_ice	float	Maximum protein percentage of the sheep milk in the icebowls that were poured into the compartment
collection_month	int(11)	Month of the collection
collection_weekday	int(11)	Weekday of the collection
area_longitude	float	Longitude value representing the area that was covered during this route
area_latitude	float	Latitude value representing the area that was covered during this route
fraud_comp	int(11)	Fraud presence indicator, as obtained by the raw data
fraud_probability	float	Probability of fraud, as calculated by the deep learning model
fraud_indicator	int(11)	Fraud presence indicator, as calculated by the deep learning model

CoordinatesTable

Table 5.12 containing aggregated geospatial data for all 3 types of frauds detected on a given day. The data are used in order to construct geospatial charts that depict the fraud distribution on the map.

Table 5.12 Aggregated geospatial data

Field Name	Type	Description
batch_date	datetime	Date of the batch collection
WaterFraud	bigint(21)	Water fraud indicator
GoatFraud	bigint(21)	Goat fraud indicator
CowFraud	bigint(21)	Cow fraud indicator

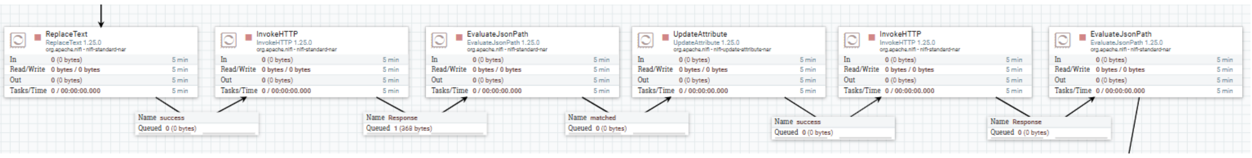
Area	varchar(255)	Geographical area of the milk collection
area_longitude	float	Longitude value representing the area that was covered during this route
area_longitude_wat	float	Longitude value representing the area that was covered during this route used for water fraud
area_longitude_cow	float	Longitude value representing the area that was covered during this route used for cow fraud
area_longitude_goat	float	Longitude value representing the area that was covered during this route used for goat fraud
area_latitude	float	Latitude value representing the area that was covered during this route
area_latitude_wat	float	Latitude value representing the area that was covered during this route used for water fraud
area_latitude_cow	float	Latitude value representing the area that was covered during this route used for cow fraud
area_latitude_goat	float	Latitude value representing the area that was covered during this route used for goat fraud

5.5.1.2 Data Acquisition

Data is retrieved from the blockchain network and fed throughout the implemented pipeline using Apache NiFi.

To extract the data from the blockchain network, the provided Web API is being used. First, a login HTTP request is sent with the respective credentials in order to acquire an authorization token, which is being used in all subsequent interactions with the API. Once the token is received, it is incorporated into the actual HTTP request that extracts the data from the API.

The NiFi architecture for this part of the pipeline is depicted in Figure 5-19.



ReplaceText
ReplaceText 1.25.0
org.apache.nifi - nifi-standard-nar
In 0 (0 bytes) 5 min
Read/Write 0 bytes / 0 bytes 5 min
Out 0 (0 bytes) 5 min
Tasks/Time 0 / 00:00:00.000 5 min

Name success
Queued 0 (0 bytes)

InvokeHTTP
InvokeHTTP 1.25.0
org.apache.nifi - nifi-standard-nar
In 0 (0 bytes) 5 min
Read/Write 0 bytes / 0 bytes 5 min
Out 0 (0 bytes) 5 min
Tasks/Time 0 / 00:00:00.000 5 min

Name Response
Queued 1 (368 bytes)

EvaluateJsonPath
EvaluateJsonPath 1.25.0
org.apache.nifi - nifi-standard-nar
In 0 (0 bytes) 5 min
Read/Write 0 bytes / 0 bytes 5 min
Out 0 (0 bytes) 5 min
Tasks/Time 0 / 00:00:00.000 5 min

Name matched
Queued 0 (0 bytes)

UpdateAttribute
UpdateAttribute 1.25.0
org.apache.nifi - nifi-update-attribute-nar
In 0 (0 bytes) 5 min
Read/Write 0 bytes / 0 bytes 5 min
Out 0 (0 bytes) 5 min
Tasks/Time 0 / 00:00:00.000 5 min

Name success
Queued 0 (0 bytes)

InvokeHTTP
InvokeHTTP 1.25.0
org.apache.nifi - nifi-standard-nar
In 0 (0 bytes) 5 min
Read/Write 0 bytes / 0 bytes 5 min
Out 0 (0 bytes) 5 min
Tasks/Time 0 / 00:00:00.000 5 min

Name Response
Queued 0 (0 bytes)

EvaluateJsonPath
EvaluateJsonPath 1.25.0
org.apache.nifi - nifi-standard-nar
In 0 (0 bytes) 5 min
Read/Write 0 bytes / 0 bytes 5 min
Out 0 (0 bytes) 5 min
Tasks/Time 0 / 00:00:00.000 5 min

Figure 5-19 Modules comprising the data acquisition sub-flow.

The main modules utilized are:

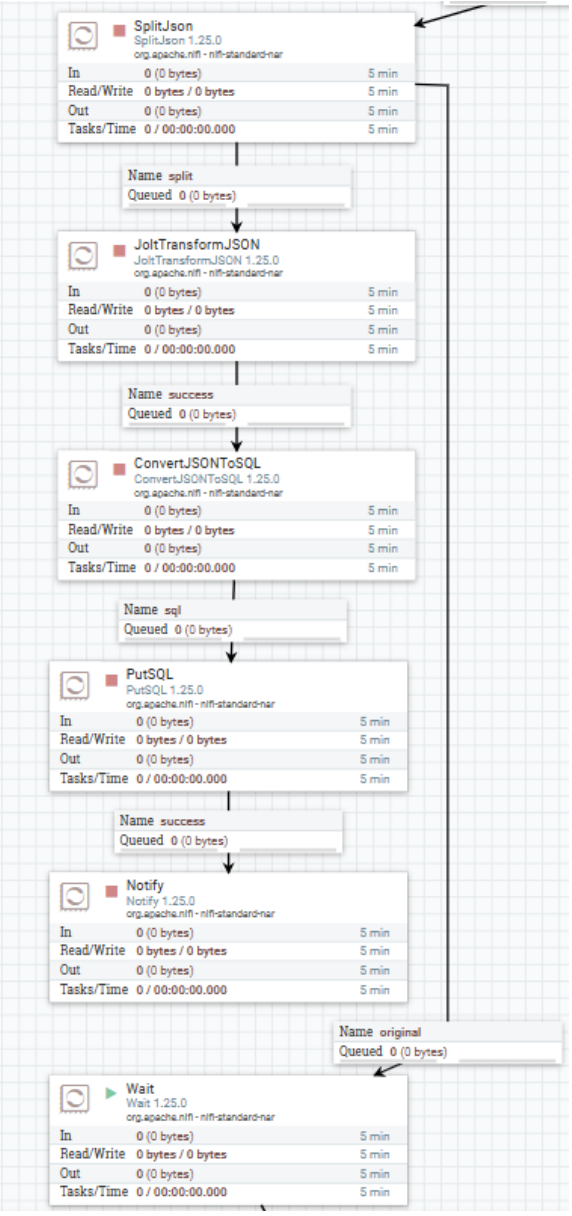
- Replace Text: Module that prepares the initial POST HTTP request for authentication
- InvokeHTTP: Module that executes the HTTP request and obtains the response with the Bearer token
- EvaluateJSONPath: Module that extracts the Bearer token from the HTTP response
- UpdateAttribute: Module that incorporates the Bearer token into the next GET HTTP for data retrieval
- InvokeHTTP: Module that executes the HTTP request and obtains the response with the data

5.5.1.3 Data Ingestion

Once the raw data is extracted from the blockchain using the Web API, the next stage of the data pipeline is activated, which performs any potentially necessary transformations, as well as the ingestion into the database of the Predictive Analytics system.

The data is received in a JSON array format, which is handled in an element-based approach. Each record is treated as a separate entry and an SQL query is constructed on its values. After the queries are executed, the data is inserted into the *milk_info* table of the database.

The NiFi architecture for this part of the pipeline is displayed in Figure 5-20.



SplitJson
SplitJson 1.25.0
org.apache.nifi - nifi-standard-nar
In 0 (0 bytes) 5 min
Read/Write 0 bytes / 0 bytes 5 min
Out 0 (0 bytes) 5 min
Tasks/Time 0 / 00:00:00.000 5 min

Name split
Queued 0 (0 bytes)

JoltTransformJSON
JoltTransformJSON 1.25.0
org.apache.nifi - nifi-standard-nar
In 0 (0 bytes) 5 min
Read/Write 0 bytes / 0 bytes 5 min
Out 0 (0 bytes) 5 min
Tasks/Time 0 / 00:00:00.000 5 min

Name success
Queued 0 (0 bytes)

ConvertJSONToSQL
ConvertJSONToSQL 1.25.0
org.apache.nifi - nifi-standard-nar
In 0 (0 bytes) 5 min
Read/Write 0 bytes / 0 bytes 5 min
Out 0 (0 bytes) 5 min
Tasks/Time 0 / 00:00:00.000 5 min

Name sql
Queued 0 (0 bytes)

PutSQL
PutSQL 1.25.0
org.apache.nifi - nifi-standard-nar
In 0 (0 bytes) 5 min
Read/Write 0 bytes / 0 bytes 5 min
Out 0 (0 bytes) 5 min
Tasks/Time 0 / 00:00:00.000 5 min

Name success
Queued 0 (0 bytes)

Notify
Notify 1.25.0
org.apache.nifi - nifi-standard-nar
In 0 (0 bytes) 5 min
Read/Write 0 bytes / 0 bytes 5 min
Out 0 (0 bytes) 5 min
Tasks/Time 0 / 00:00:00.000 5 min

Name original
Queued 0 (0 bytes)

Wait
Wait 1.25.0
org.apache.nifi - nifi-standard-nar
In 0 (0 bytes) 5 min
Read/Write 0 bytes / 0 bytes 5 min
Out 0 (0 bytes) 5 min
Tasks/Time 0 / 00:00:00.000 5 min

Figure 5-20 Modules comprising the raw data ingestion sub-flow.

The main modules utilized are:

- SplitJson: Module that splits the JSON array with the incoming data into multiple elements, in order to be handled element-by-element.
- ConvertJSONtoSQL: Module that constructs the respective SQL query for each record described by the JSON element.
- PutSQL: Module that executes the INSERT SQL query against the database, ingesting the record.

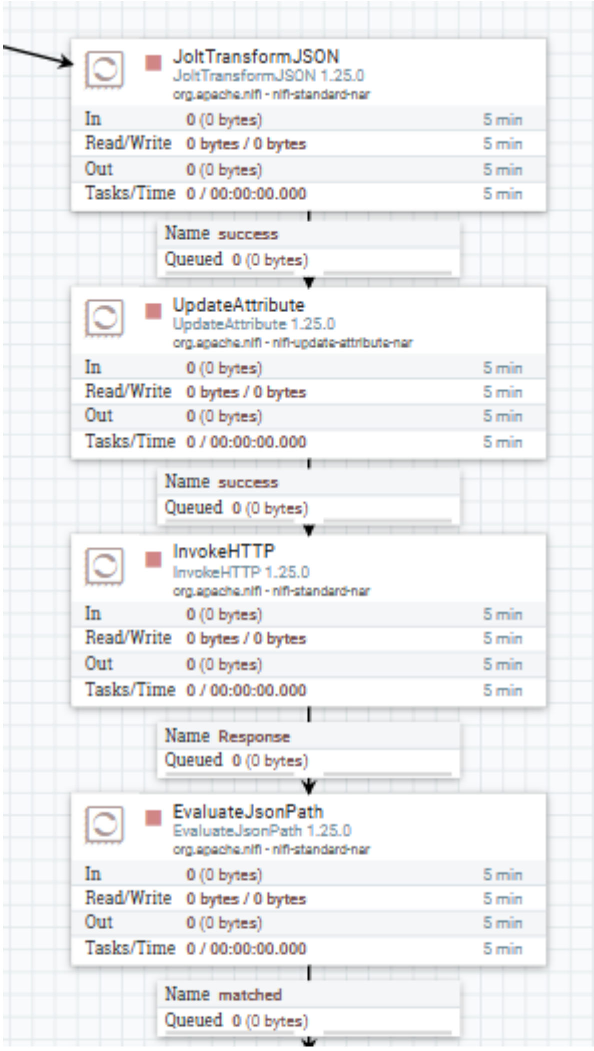
- Wait/Notify: Modules ensuring that all the records of a single JSON array have been ingested before moving on to the next step, in order to guarantee the atomicity of the operations.

5.5.2 Predictions Inference

Apart from ingesting the raw data into the database, the AI-enabled predictions must also be generated, that will serve as the dataset, upon which the final results are calculated and presented to the end users.

The deep learning model used for the predictions inference is also deployed as a Web Service and similarly accessed via its respective Web API. The JSON array containing the raw data is forwarded into the model API and a JSON response is received, containing aggregated info for all types of fraud being examined (cow, goat, water). Next, the response is fed into 3 parallel sub-flows, each for updating the respective table which contains the fraud prediction data for a specific type of fraud (*cow_fraud_prediction_data, goat_fraud_prediction_data, water_fraud_prediction_data*).

The NiFi architecture for this part of the pipeline is described in Figure 5-22:



JoltTransformJSON
JoltTransformJSON 1.25.0
org.apache.nifi - nifi-standard-nar

In	0 (0 bytes)	5 min
Read/Write	0 bytes / 0 bytes	5 min
Out	0 (0 bytes)	5 min
Tasks/Time	0 / 00:00:00.000	5 min

Name success
Queued 0 (0 bytes)

UpdateAttribute
UpdateAttribute 1.25.0
org.apache.nifi - nifi-update-attribute-nar

In	0 (0 bytes)	5 min
Read/Write	0 bytes / 0 bytes	5 min
Out	0 (0 bytes)	5 min
Tasks/Time	0 / 00:00:00.000	5 min

Name success
Queued 0 (0 bytes)

InvokeHTTP
InvokeHTTP 1.25.0
org.apache.nifi - nifi-standard-nar

In	0 (0 bytes)	5 min
Read/Write	0 bytes / 0 bytes	5 min
Out	0 (0 bytes)	5 min
Tasks/Time	0 / 00:00:00.000	5 min

Name Response
Queued 0 (0 bytes)

EvaluateJsonPath
EvaluateJsonPath 1.25.0
org.apache.nifi - nifi-standard-nar

In	0 (0 bytes)	5 min
Read/Write	0 bytes / 0 bytes	5 min
Out	0 (0 bytes)	5 min
Tasks/Time	0 / 00:00:00.000	5 min

Name matched
Queued 0 (0 bytes)

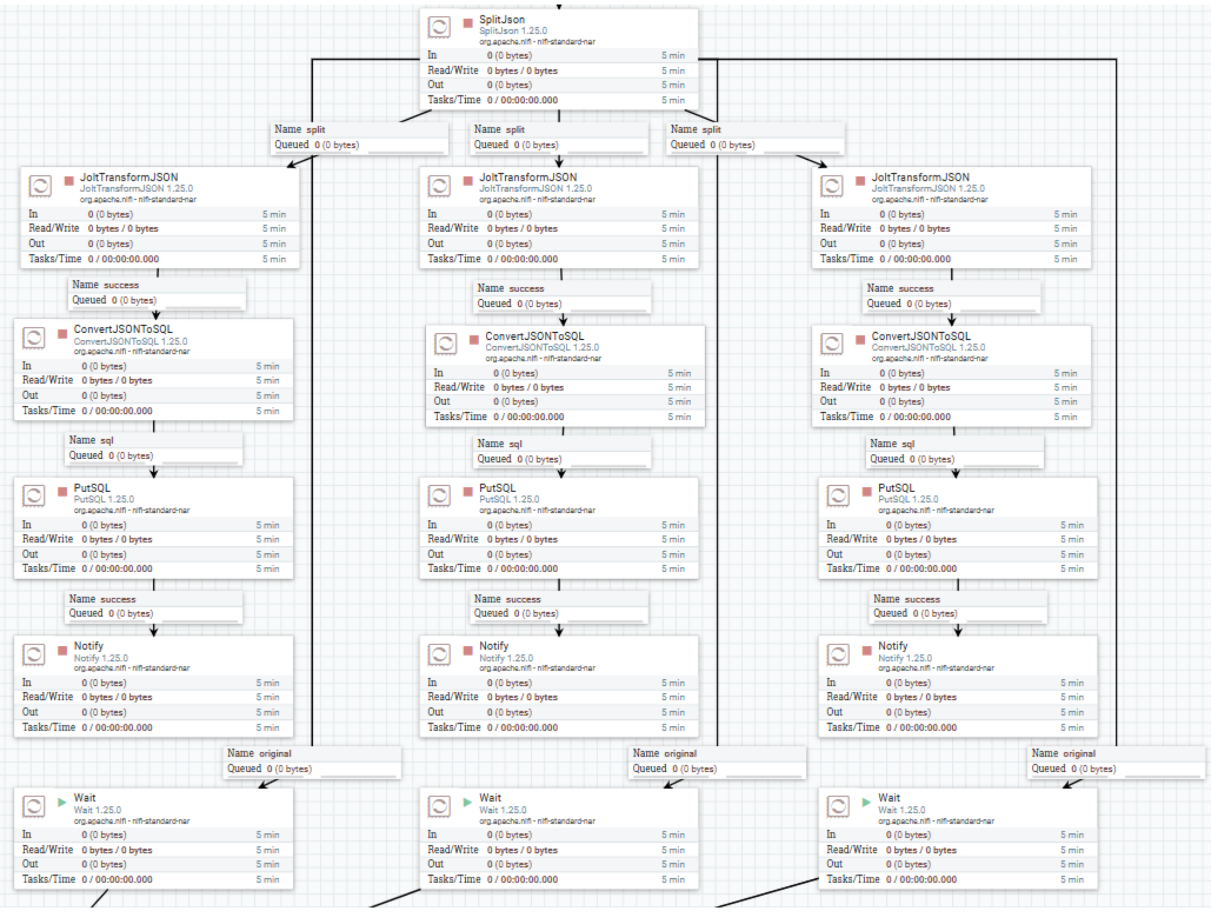


Figure 5-21 Modules comprising the predictions inference & fraud data storage sub-flow.

The main modules utilized are:

- JoltTransformJson: Module that prepares the JSON array with the raw data, in order to be forwarded to the Deep Learning Model API
- InvokeHTTP: Module that executes the HTTP request against the Deep Learning Model API and obtains the response with the predictions
- SplitJSON: Module that splits the JSON array with the prediction data into multiple elements, in order to be handled element-by-element. It also forwards the data into the 3 parallel sub-flows that update the respective fraud tables
- ConvertJSONtoSQL: Module that constructs the respective SQL query for each record described by the JSON element
- PutSQL: Module that executes the INSERT SQL query against the database, ingesting the record
- Wait/Notify: Modules ensuring that all the records of a single JSON array have been ingested before moving on to the next step, in order to guarantee the atomicity of the operations

5.5.2.1 Update of data between tables

Following up on the defined database schema, certain tables need to store pieces of information that are generated in a different part of the pipeline. Therefore, once all the data has been initially ingested, we need to update the tables by joining them with each other. For that purpose, a final part of the pipeline has been constructed, which performs these updates once all previous tables have been filled with data.

The Nifi architecture for this part of the pipeline is depicted in Figure 5-22.

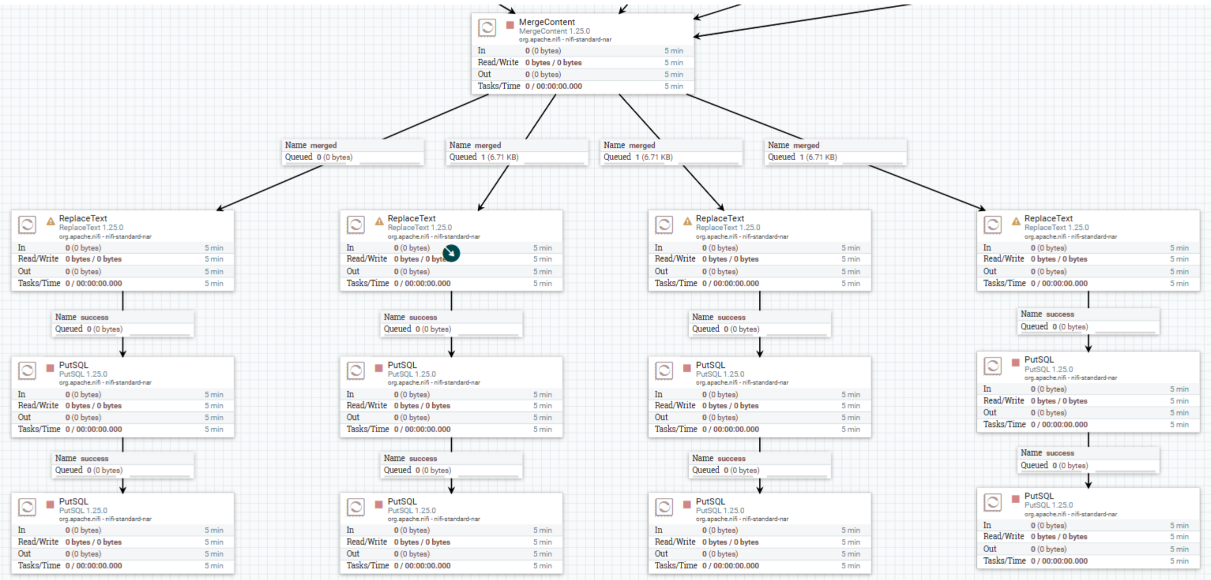


Figure 5-22 Modules comprising the sub-flow for data update and consolidation.

The main modules utilized are:

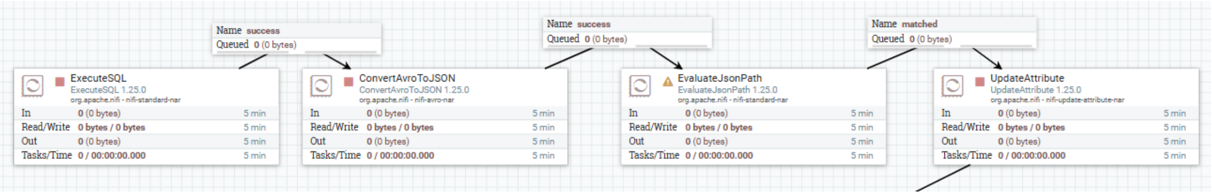
- MergeContent: Module that waits for the ingestion in all tables to be completed, before proceeding with updates between them
- PutSQL: Modules that perform the updates between the tables by executing the respective JOIN SQL queries

5.5.3 Offsetting and continuous syncing between systems

To ensure a seamless integration of systems across the ALLIANCE platform, as well as to provide results of high business value to the end users, it is essential that the Food Fraud Prevention with Predictive Analytics database is continuously in sync with the Blockchain Network.

This is achieved by pinging the Blockchain Web API in regular time intervals and retrieving the new data produced by the pilot as soon as possible upon their addition. In order to identify the new data added, a column containing the creation timestamp of each record is being used as an offset. By keeping the max value of this column and only retrieving data with larger values in the respective field, it is guaranteed that the newly acquired records will be the most recent ones and the risk of creating duplicate entries is eliminated.

The NiFi architecture for this part of the pipeline is depicted in Figure 5-23.



Name success
Queued 0 (0 bytes)

Name success
Queued 0 (0 bytes)

Name matched
Queued 0 (0 bytes)

ExecuteSQL
ExecuteSQL 1.25.0
org.apache.nifi - nifi-standard-nar
In 0 (0 bytes) 5 min
Read/Write 0 bytes / 0 bytes 5 min
Out 0 (0 bytes) 5 min
Tasks/Time 0 / 00:00:00.000 5 min

ConvertAvroToJSON
ConvertAvroToJSON 1.25.0
org.apache.nifi - nifi-avro-nar
In 0 (0 bytes) 5 min
Read/Write 0 bytes / 0 bytes 5 min
Out 0 (0 bytes) 5 min
Tasks/Time 0 / 00:00:00.000 5 min

EvaluateJsonPath
EvaluateJsonPath 1.25.0
org.apache.nifi - nifi-standard-nar
In 0 (0 bytes) 5 min
Read/Write 0 bytes / 0 bytes 5 min
Out 0 (0 bytes) 5 min
Tasks/Time 0 / 00:00:00.000 5 min

UpdateAttribute
UpdateAttribute 1.25.0
org.apache.nifi - nifi-update-attribute-nar
In 0 (0 bytes) 5 min
Read/Write 0 bytes / 0 bytes 5 min
Out 0 (0 bytes) 5 min
Tasks/Time 0 / 00:00:00.000 5 min

Figure 5-23 Modules comprising the continuous syncing sub-flow.

The main modules utilized are:

- ExecuteSQL: Module that extracts the max value of the column containing the creation timestamp of each raw data record in our database, which is subsequently used in the GET HTTP request in order to identify the new data added to the Blockchain

- EvaluateJsonPath: Module that prepares the max timestamp value to be incorporated into the HTTP request

- UpdateAttribute: Module that constructs the new HTTP request using the max value extracted from the database

5.6 Demonstration of System Functionalities

Apache Superset platform was used for data visualisation and dashboard creation purposes. Apache Superset™ is an open-source data exploration and visualisation platform.

On *Apache Superset*, six main dashboards were created under the following names: *Timeseries - Historical*, *Timeseries - Forecasts*, *Clustering*, *Geospatial Visualisations*, *Probabilities* and *Model Performance*. These dashboards are visual interfaces that show key information, data and metrics through charts, graphs and tables.

5.6.1 Dashboard and Charts creation on Superset

Firstly, a connection is needed to be established between Superset and *Alliance_DB* database to be able to query and visualize data from it, since Superset doesn't have a storage layer to store data. Once the data source is configured, the user can select specific tables, called Datasets in Superset that will be exposed in Superset for querying. The schema used here was *olympos_sc*.

Superset has a thin semantic layer that can store two types of computed data:

1. Virtual metrics: The user can write SQL queries that aggregate values from multiple columns and make them available as columns for visualization in Explore.

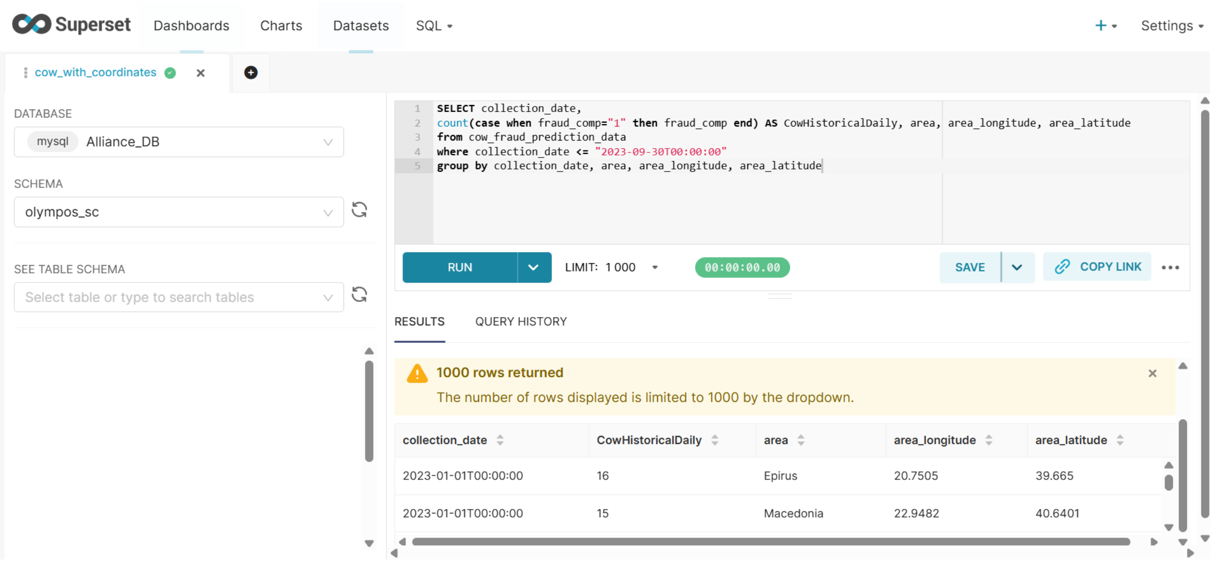
2. Virtual calculated columns: The user can write SQL queries that customize the appearance and behaviour of a specific column.

Superset has 2 main interfaces for exploring data:

1. Explore: No-code builder. The user selects a dataset, selects the chart, customizes the appearance and publishes.

2. SQL Lab: SQL IDE for cleaning, joining and preparing data for Explore workflow.

Mainly, SQL Lab was used in order to create the appropriate datasets that were needed to create the charts presented in the six dashboards. For example, the query shown in the following picture was used to generate the dataset for the ‘Cow - Map’ chart in the "Geospatial Visualisations" dashboard.



```
SELECT collection_date,
count(case when fraud_comp="1" then fraud_comp end) AS CowHistoricalDaily, area, area_longitude, area_latitude
from cow_fraud_prediction_data
where collection_date <= "2023-09-30T00:00:00"
group by collection_date, area, area_longitude, area_latitude
```

1000 rows returned
The number of rows displayed is limited to 1000 by the dropdown.

collection_date	CowHistoricalDaily	area	area_longitude	area_latitude
2023-01-01T00:00:00	16	Epirus	20.7505	39.665
2023-01-01T00:00:00	15	Macedonia	22.9482	40.6401

Figure 5-24 - "Cow - Map" query

Based on the above query results, the following chart was created.

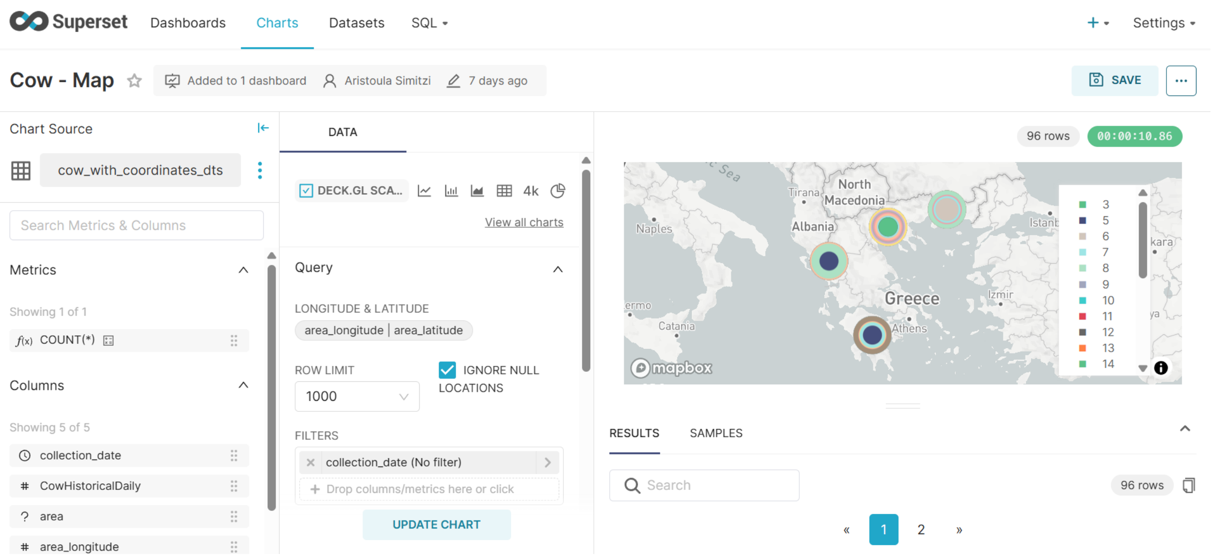


Figure 5-25 - "Cow - Map" chart creation

The chart was then added to the *Geospatial Visualisations* dashboard. Accordingly, all other charts were created and then added to the respective dashboards.

5.6.1.1 Timeseries - Historical Dashboard

In this dashboard, timeseries of historical data are presented in five visualisations. These timeseries of historical data charts show how the daily number of frauds changes over time. In particular, the horizontal axis represents time, and the vertical axis represents the daily sum of incidents. In each visual, the threshold of 30 incidents per day is plotted with a red horizontal line.

In the following visualisation the historical data for cow fraud are shown for the time period 1/1/2023 to 30/09/2023.

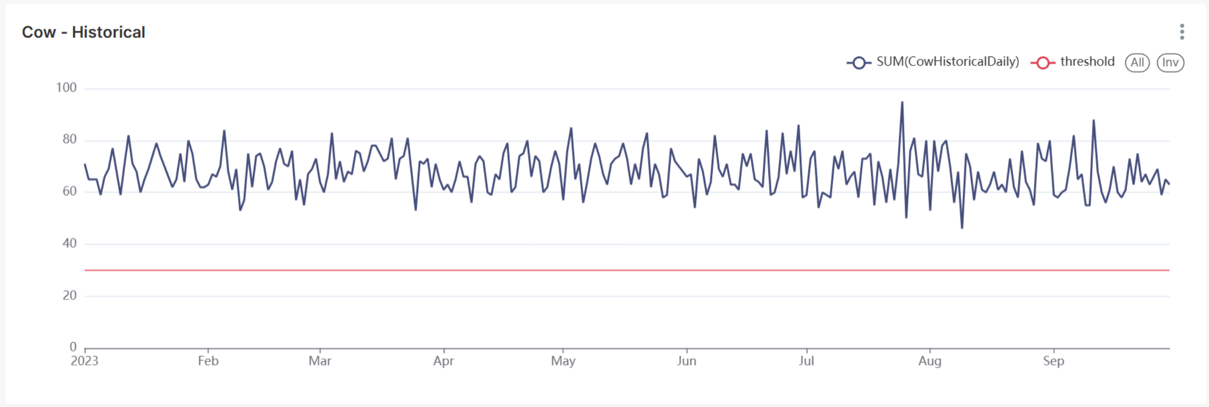


Figure 5-26 Cow fraud - Historical timeseries

Likewise, in the visualisation below the historical data for goat fraud are shown for the time period 1/1/2023 to 30/09/2023.

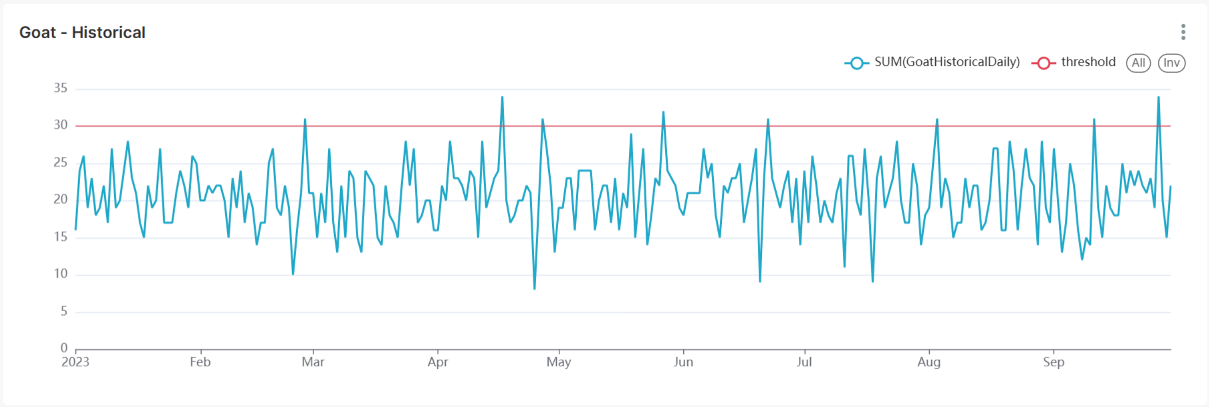


Figure 5-27 Goat fraud - Historical timeseries

Next, the historical data for water fraud are shown for the time period 1/1/2023 to 30/09/2023.

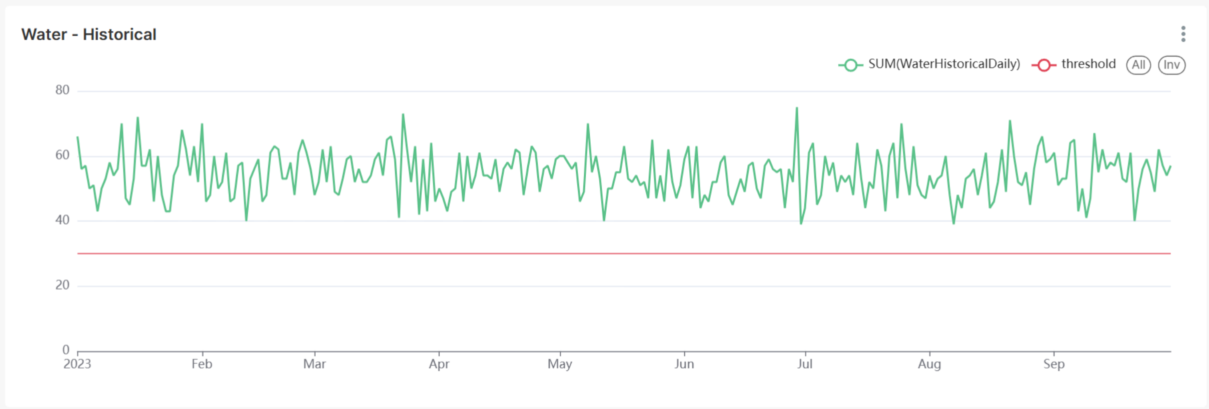


Figure 5-28 Water fraud - Historical timeseries

In the following visualisation the historical data for all three types of fraud, cow, goat and water are shown together for the time period 1/1/2023 to 30/09/2023.

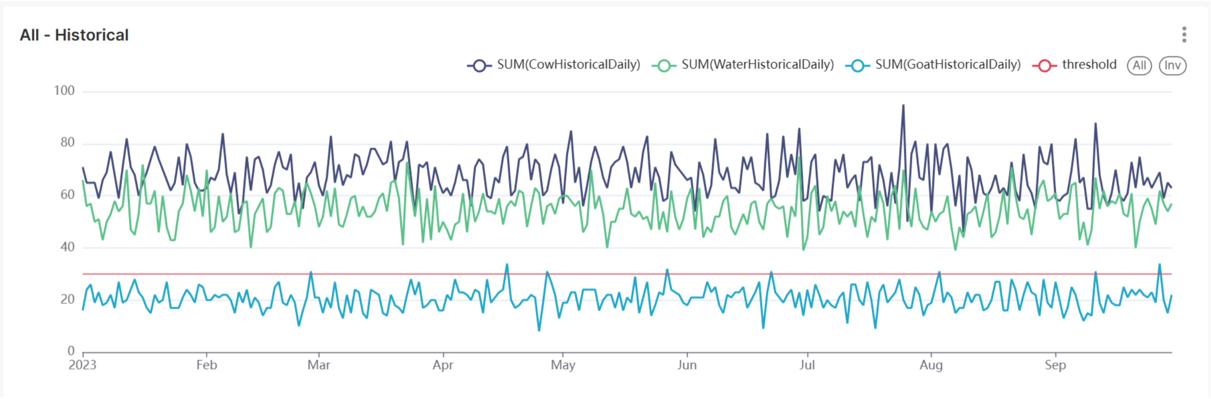


Figure 5-29 All types of fraud - Historical timeseries

The following visualisation shows the daily sum of the three types of fraud, cow, goat and water for the time period 1/1/2023 to 30/09/2023.

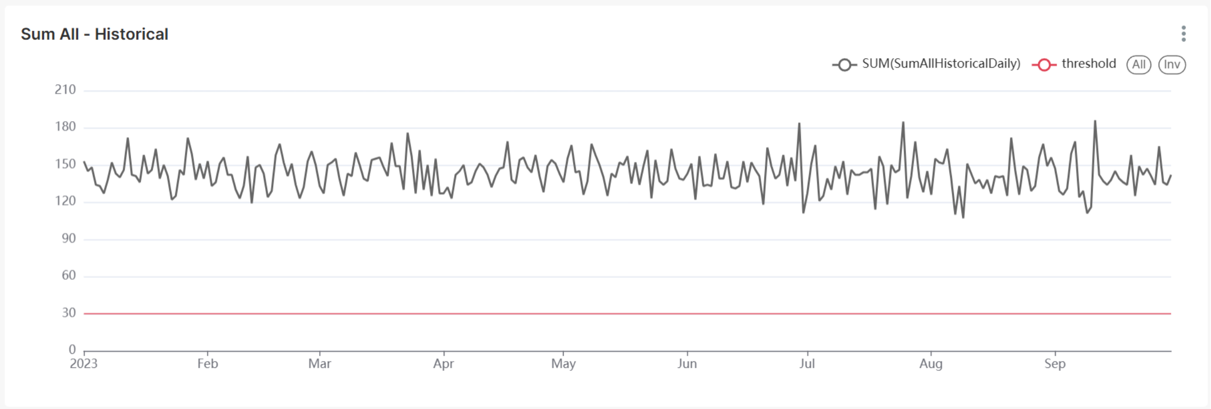


Figure 5-30 Daily sum of all types of fraud - Historical timeseries

On the left side of the dashboard, there is the option to select the date filter. The date filter allows users to dynamically narrow down the time period for which data is displayed and analysed. More specifically, this filter serves as an interactive feature that enables users to focus on specific timeframes relevant to their analysis, identify trends and gain deeper insights from the visualisations. After selecting the relevant date or time period, the user needs to press the "Apply filters" button. Then, all visualisations are automatically updated. In case the user needs to clear the selected filter, they need to press the "Clear all" button.

Filters
+ ADD/EDIT FILTERS
APPLY FILTERS
CLEAR ALL

Date

2023-04-01 ≤ col < 2023...

Figure 5-31 Date filter - Historical timeseries

5.6.1.2 Timeseries - Forecasts Dashboard

In this dashboard, timeseries of predictive model forecasts based on historical data are presented in five visualisations. In these charts, the horizontal axis represents time, and the vertical axis represents the daily sum of incidents. In each visual, the threshold of 30 incidents per day is plotted with a red horizontal line.

In the following visualisation the historical data for cow fraud are shown for the time period 1/1/2023 to 30/09/2023 along with the data that are predicted for the time period 1/10/2023 to 31/12/2023.

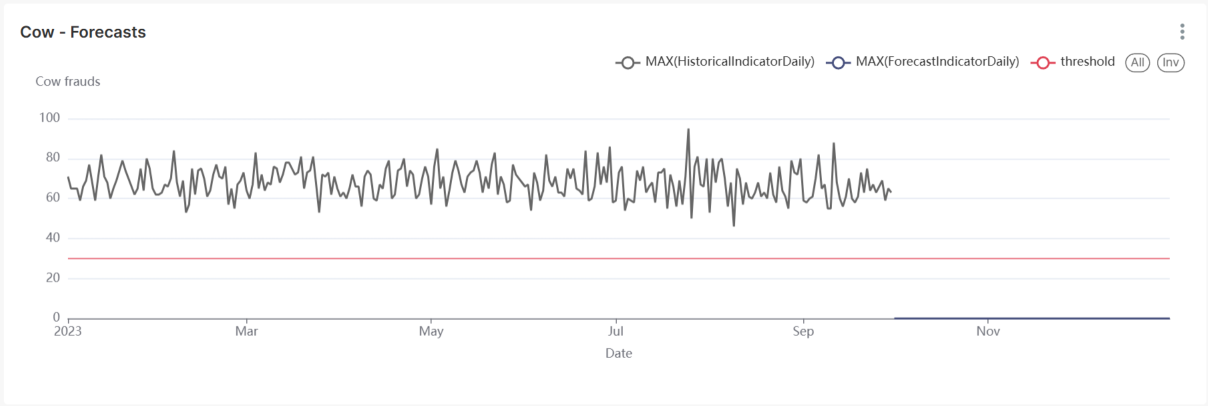


Figure 5-32 Cow fraud - Forecasts timeseries

Likewise, in the visualisation below the historical data for goat fraud are shown for the time period 1/1/2023 to 30/09/2023 along with the forecasts for the time period 1/10/2023 to 31/12/2023.

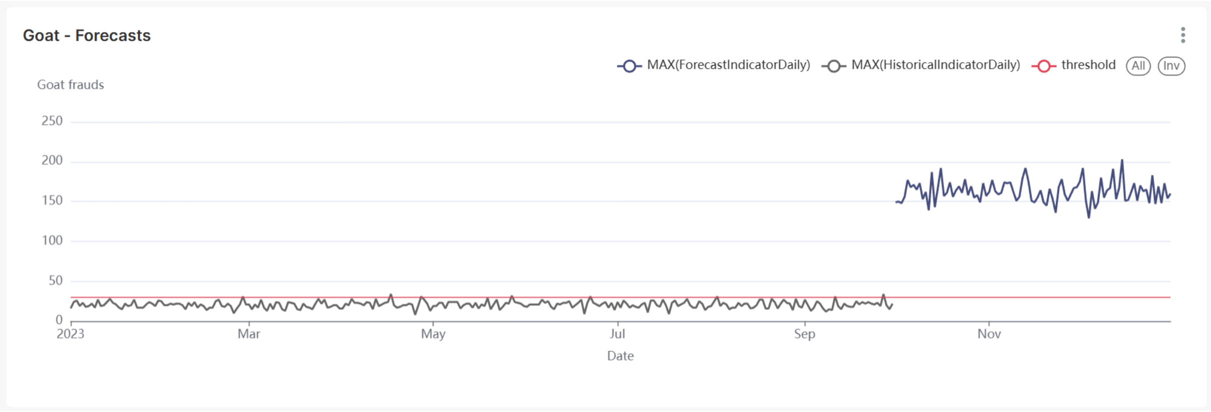


Figure 5-33 Goat fraud - Forecasts timeseries

Next, the historical data for water fraud are shown for the time period 1/1/2023 to 30/09/2023 and the forecasts for the time period 1/10/2023 to 31/12/2023.

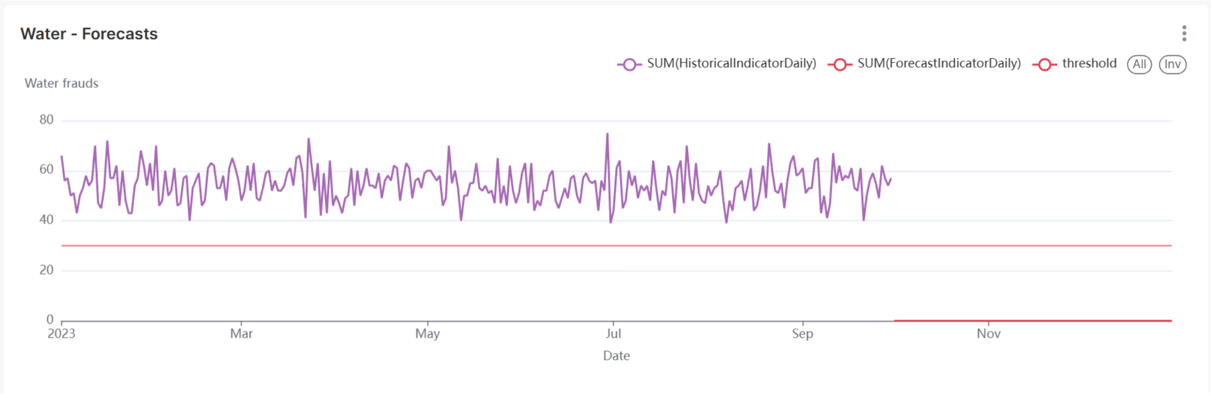


Figure 5-34 Water fraud - Forecasts timeseries

In the following visualisation the historical data for all three types of fraud, cow, goat and water are shown together for the time period 1/1/2023 to 30/09/2023 with the respective daily forecasts for the time period 1/10/2023 to 31/12/2023.

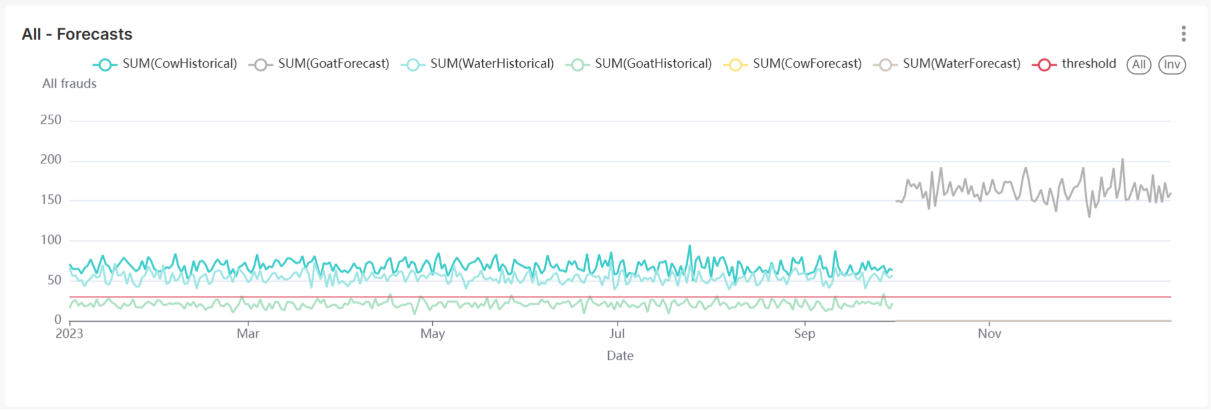


Figure 5-35 All types of fraud - Forecasts timeseries

The following visualisation shows the daily sum of the three types of fraud, cow, goat and water for the time period 1/1/2023 to 30/09/2023 and the daily sum of the forecasted values for the time period 1/10/2023 to 31/12/2023.

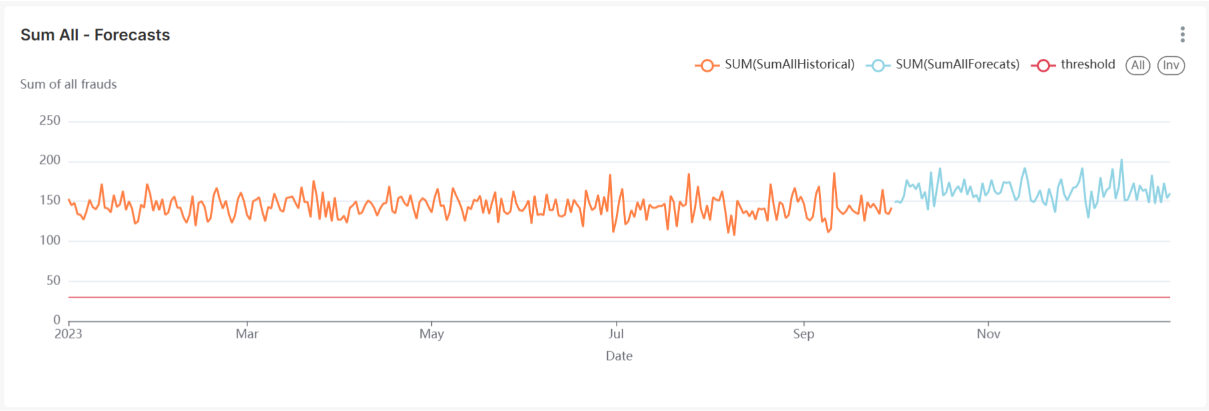
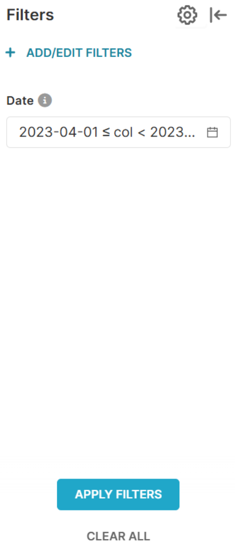


Figure 5-36 Daily sum of all types of fraud - Forecasts timeseries

On the left side of the dashboard, there is the option to select the date filter. The date filter allows users to dynamically narrow down the time period for which data is displayed and analysed. More specifically, this filter serves as an interactive feature that enables users to focus on specific timeframes relevant to their analysis, identify trends and gain deeper insights from the visualisations. After selecting the relevant date or time period, the user needs to press the "Apply filters" button. Then, all visualisations are automatically updated. In case the user needs to clear the selected filter, they need to press the "Clear all" button.



2023-04-01 ≤ col < 2023...

Figure 5-37 Date filter - Forecasts timeseries

5.6.1.3 Clustering Dashboard

In this dashboard, the results of cluster analysis are presented in five charts. Clustering analysis is a machine learning technique that was used to group food suppliers together based on their features.

The following chart visualizes the results of Principal Component Analysis in a scatter plot. Principal Component Analysis (PCA) is a dimensionality reduction technique that transforms a dataset with many correlated variables into a smaller set of uncorrelated variables called principal components. Through this visual the user can observe the clusters of data points appearing in the reduced dimensional space.

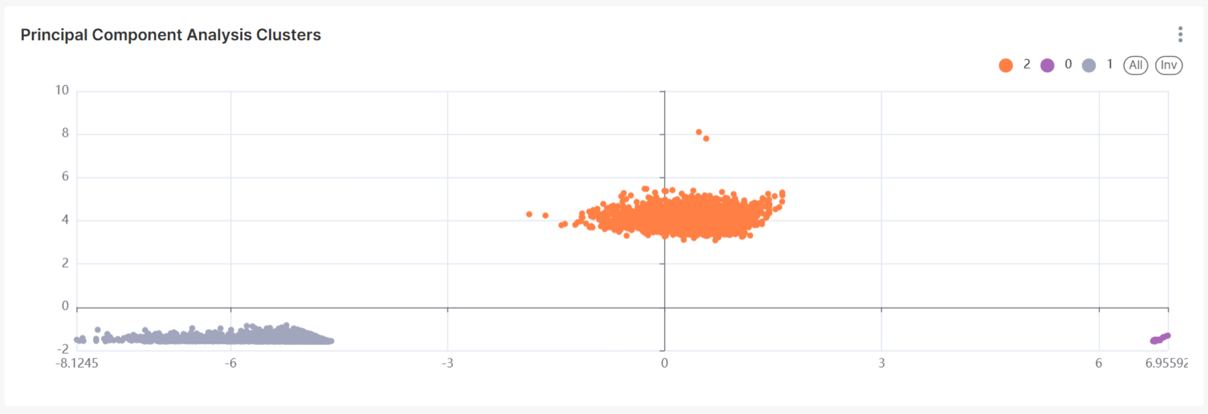


Figure 5-38 Principal Component Analysis Clusters - Clustering

The next visualisation the average quantities of goat and sheep milk are shown in two bars per cluster.

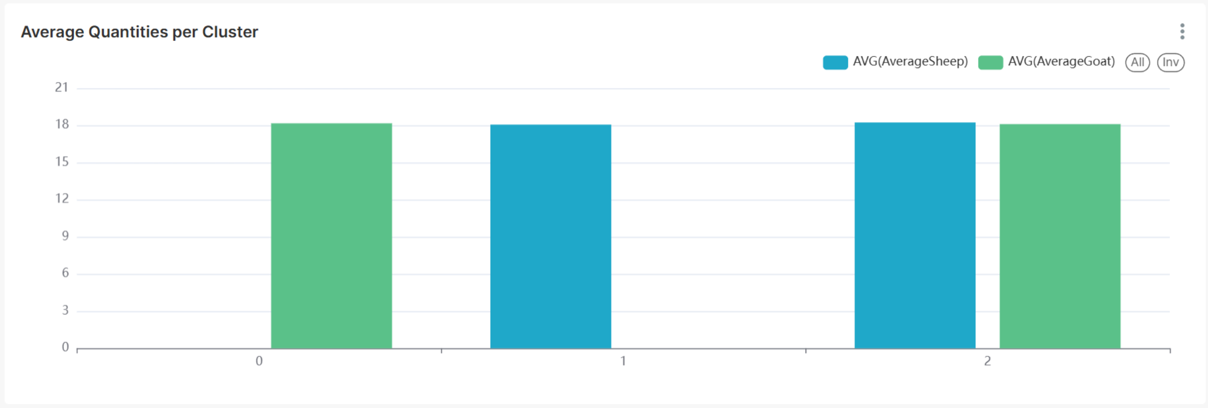


Figure 5-39 Average Quantities per Cluster - Clustering

The following visualisation shows the average numbers of water, goat and cow fraud incidents in sheep milk and also the average numbers of water and cow fraud incidents in goat milk, per cluster.

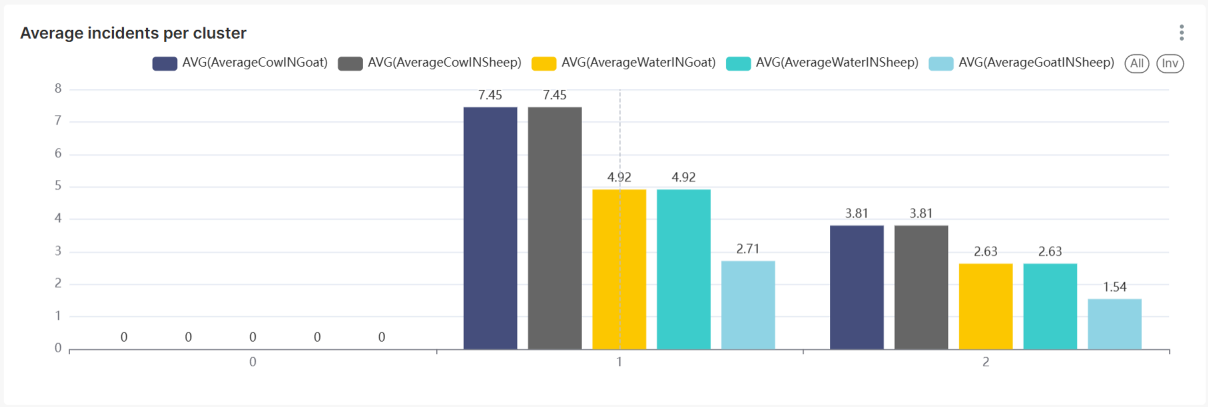
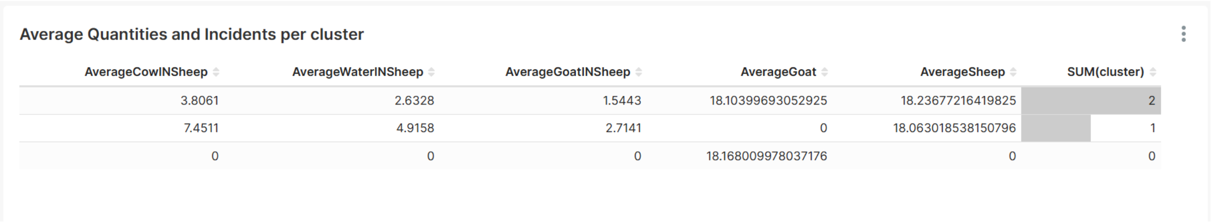


Figure 5-40 Average incidents per cluster - Clustering

The next table presents the average number of water, goat and cow incidents in sheep milk and also the average quantities of goat and sheep milk, per cluster.



Average Quantities and Incidents per cluster

AverageCowINSheep	AverageWaterINSheep	AverageGoatINSheep	AverageGoat	AverageSheep	SUM(cluster)
3.8061	2.6328	1.5443	18.10399693052925	18.23677216419825	2
7.4511	4.9158	2.7141	0	18.063018538150796	1
0	0	0	18.168009978037176	0	0

Figure 5-41 Average Quantities and Incidents per cluster - Clustering

Lastly, the following visualisation is a radar chart where the axes correspond to the average number of water, goat and cow incidents in sheep milk and also the average quantities of goat and sheep milk, per cluster. This chart is a valuable visualization technique for understanding and communicating the results of clustering analysis by providing a visual profile of each cluster and facilitating comparisons across multiple variables.

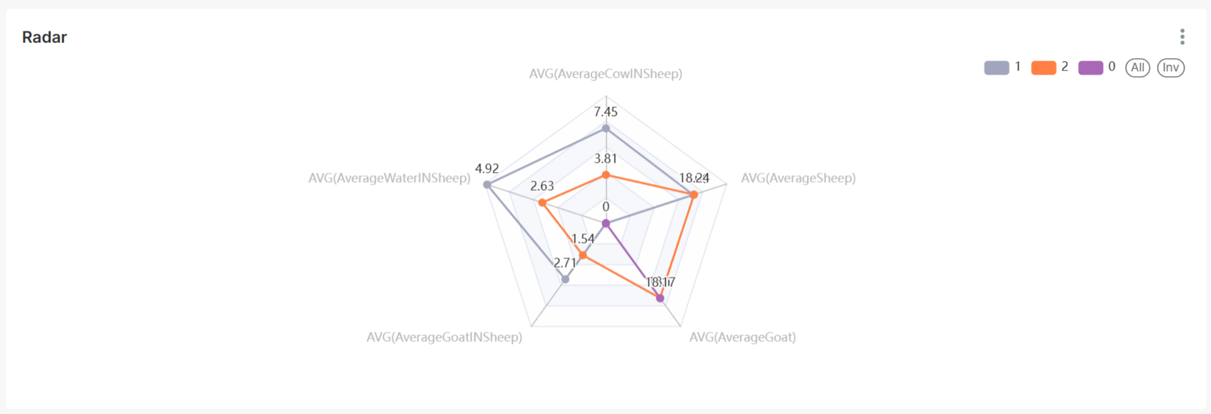


Figure 5-42 Radar - Clustering

5.6.1.4 Geospatial Visualisations Dashboard

In this dashboard, the geospatial distribution of food fraud incidents is mapped in four visuals.

On the right side of each map chart, there is a panel where the colours used for each point are matched to the actual number of incidents of the respective area.

The following map shows the daily incidents of cow fraud per area.

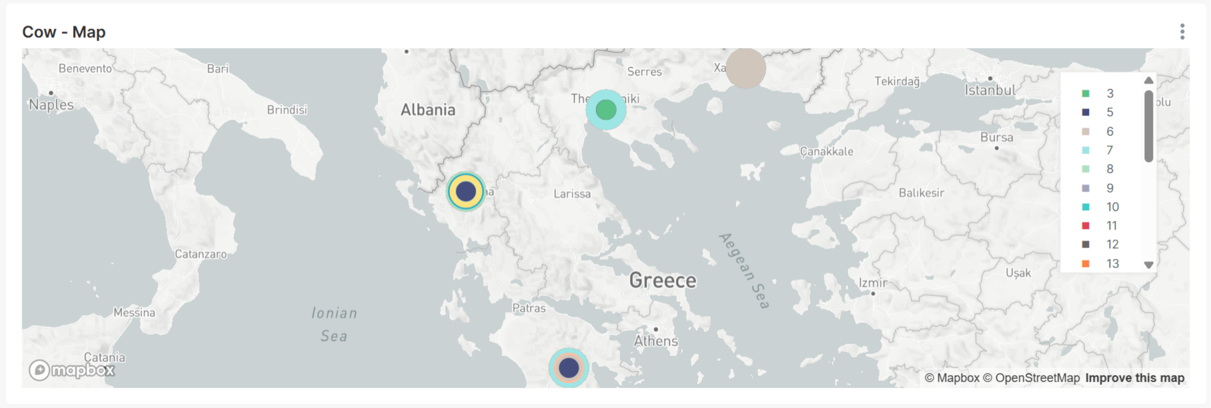


Figure 5-43 Cow fraud - Geospatial visualisations

Similarly, the below map shows the daily incidents of goat fraud per area.

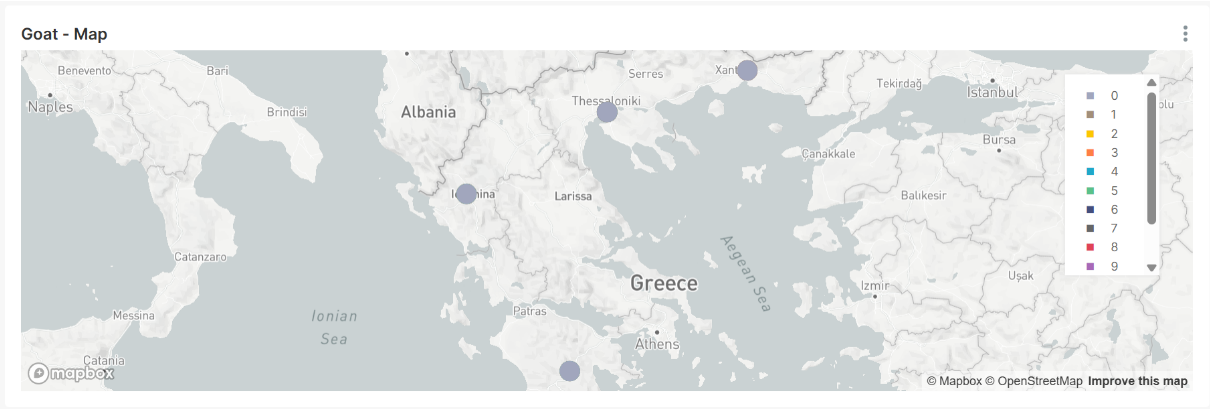


Figure 5-44 Goat fraud - Geospatial visualisations

The next map shows the daily incidents of water fraud per area.

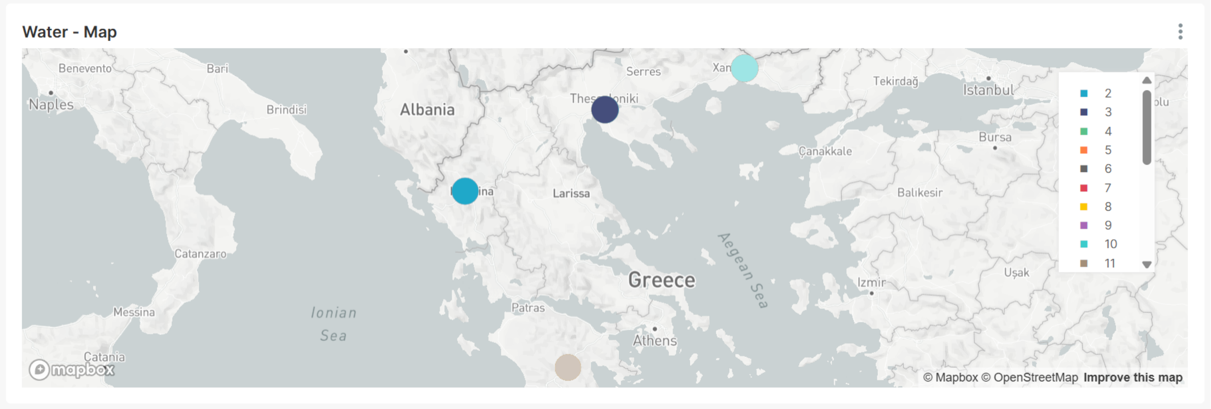


Figure 5-45 Water fraud - Geospatial visualisations

Moreover, in the following map the sum of incidents of all fraud types are visualized.

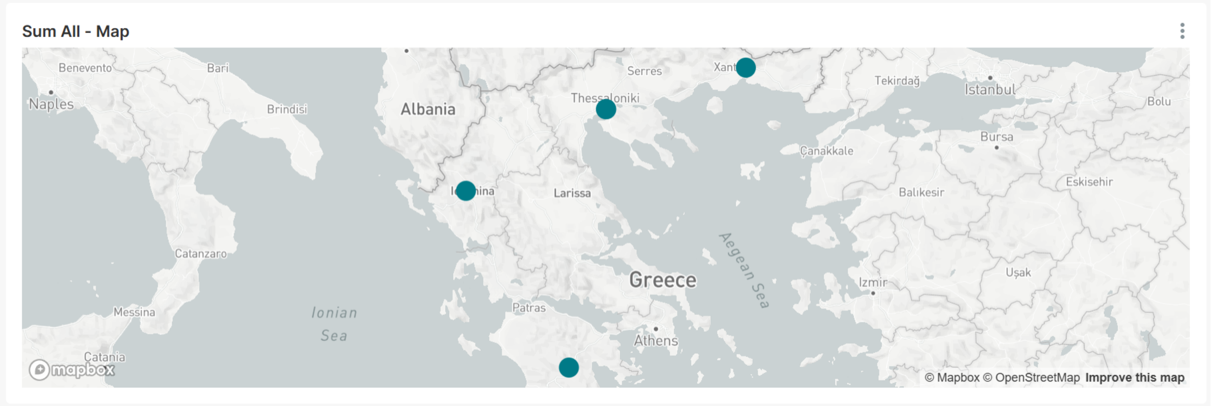
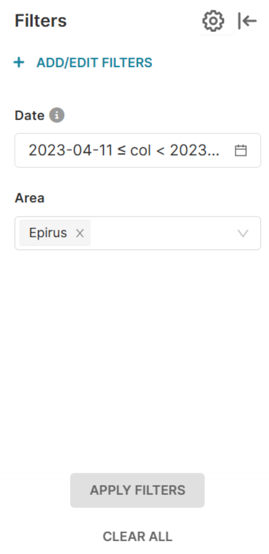


Figure 5-46 Sum of all types of fraud - Geospatial visualisations

On the left side of the dashboard, there is the option to select the date and the area filters. The date filter allows users to dynamically narrow down the time period for which data is displayed and analysed. The area filter can be selected in case the user needs to visualize data for only few areas in the map. After selecting the relevant dates or/and areas, the user needs to press the "Apply filters" button. Then, all visualisations are automatically updated. In case the user needs to clear the selected filters, they need to press the "Clear all" button.



Filters

Date: 2023-04-11 ≤ col < 2023...

Area: Epirus ×

Figure 5-47 Date and Area filters - Geospatial visualisations

5.6.1.5 Probabilities Dashboard

This dashboard consists of nine charts, three of which are probability distribution charts and the rest six are scatter plots.

Diagonally, three probability distribution charts, one for each type of fraud, are presented. Probability distribution charts visually communicate the likelihood of different outcomes of a variable, here of fraud.

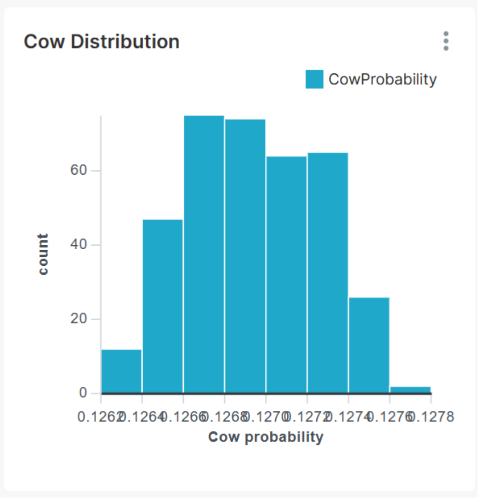


Figure 5-48 Cow Distribution chart - Probabilities

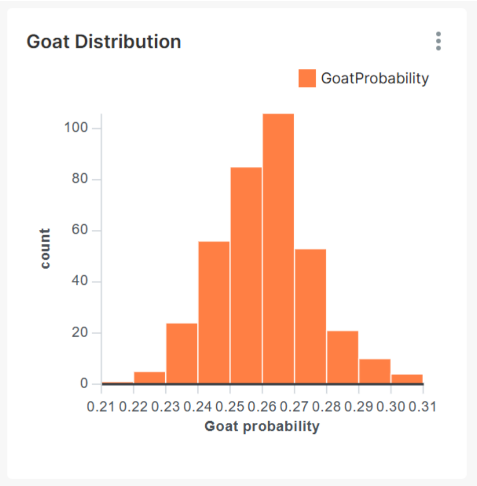


Figure 5-49 Goat Distribution - Probabilities

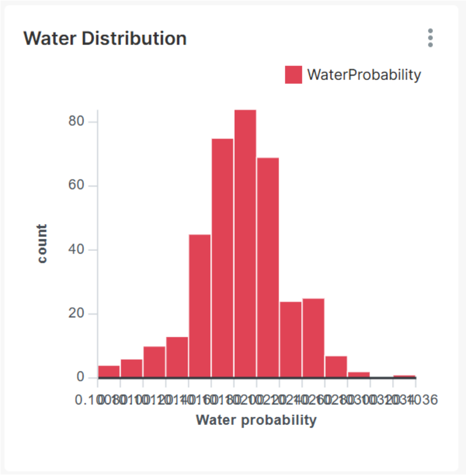


Figure 5-50 Water Distribution - Probabilities

Moreover, six probability scatter plots are shown in this dashboard. These are a type of scatter plot where both axes represent probabilities. The purpose of these plots is typically to visualize relationships involving probabilistic data.

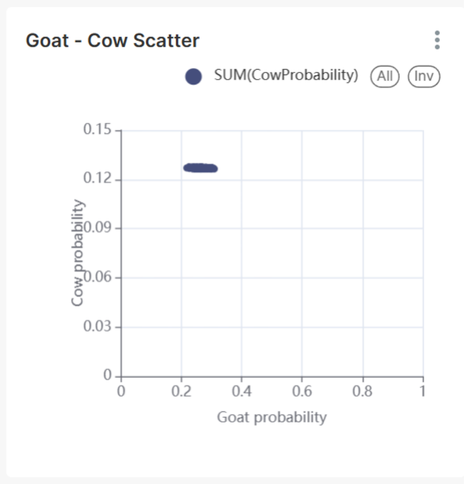


Figure 5-51 Goat - Cow Scatter plot - Probabilities

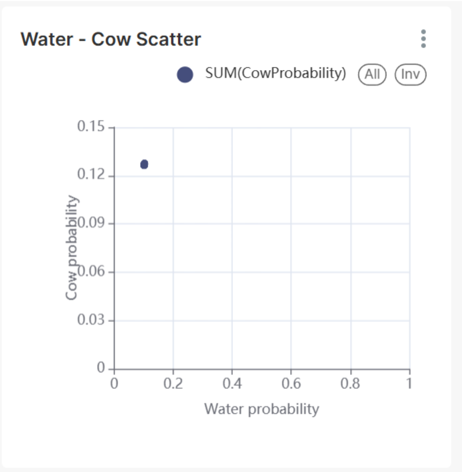


Figure 5-52 Water - Cow Scatter plot - Probabilities

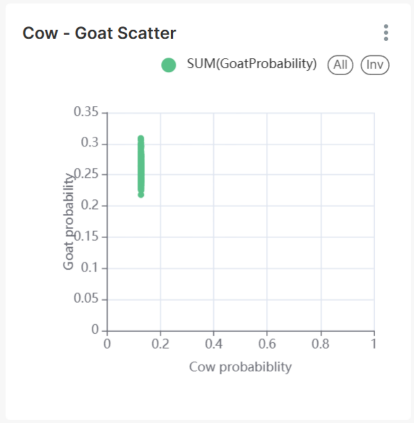


Figure 5-53 Cow - Goat Scatter plot - Probabilities

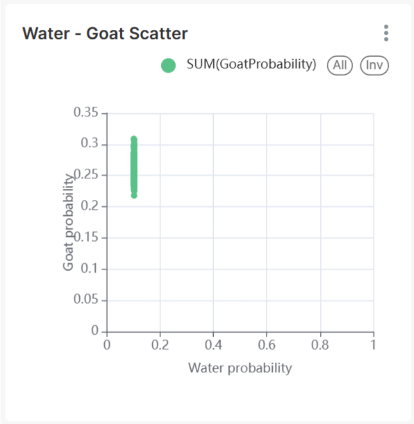


Figure 5-54 Water - Goat Scatter plot - Probabilities

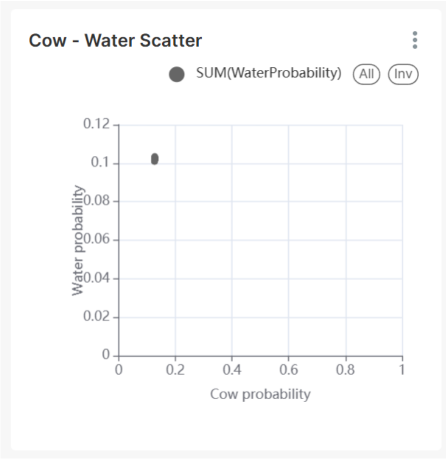


Figure 5-55 Cow - Water Scatter plot - Probabilities

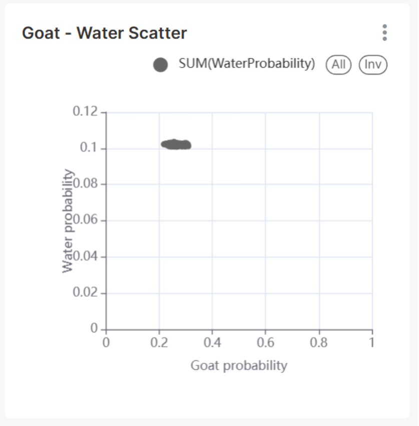


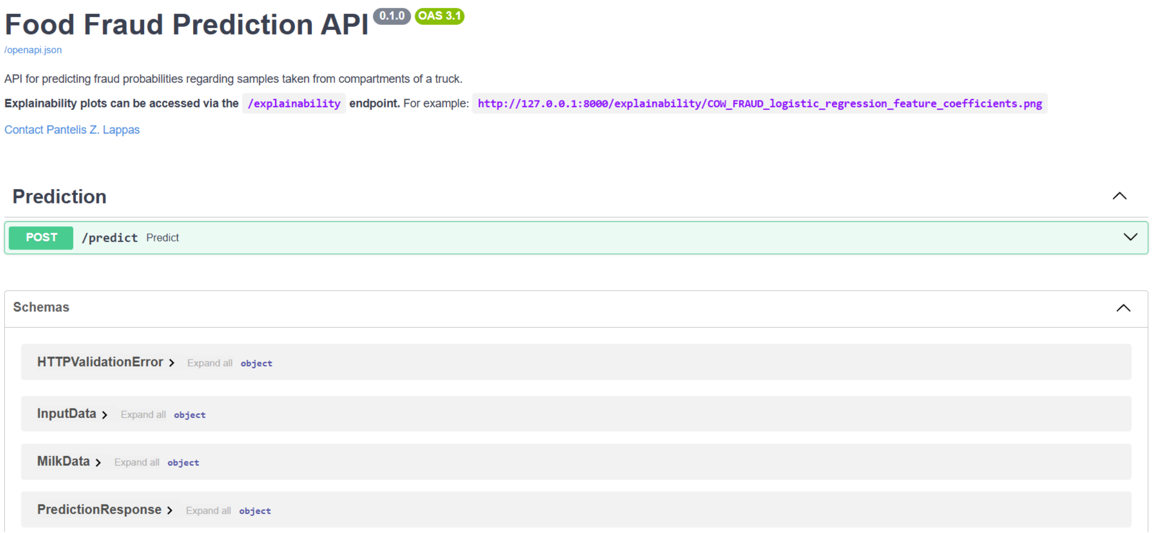
Figure 5-56 Goat - Water Scatter plot - Probabilities

5.6.1.6 Model Performance Dashboard

The *Model Performance Dashboard* allows users to evaluate how well the deployed prediction services are functioning in practice. It focuses on the three specific types of fraud detected by the system (i.e., cow, water, and goat fraud types) by summarizing key outcomes in a confusion matrix for each. These matrices display counts of true positives, true negatives, false positives, and false negatives, offering an immediate sense of where the model is performing well and where it may be making errors. Alongside these, additional metrics such as accuracy, area AUC and so on are automatically calculated and visualized. These values help interpret how reliable the model is under real usage, especially when the actual outcome is known.

5.7 Service Deployment and Access

In order to make the developed prediction and clustering models easily accessible and usable, both were deployed as web services. These services enable external applications to interact with the models via standardized HTTP requests. Users can send input data in JSON format to specify endpoints and receive model predictions or cluster assignments in real-time. The service can be accessed via the following URL: Food Fraud Prediction API - Swagger UI for predictions and Milk Supply Clustering API - Swagger UI for clustering assignments.



Food Fraud Prediction API 0.1.0 OAS 3.1

/openapi.json

API for predicting fraud probabilities regarding samples taken from compartments of a truck.

Explainability plots can be accessed via the `/explainability` **endpoint.** For example: `http://127.0.0.1:8000/explainability/COW_FRAUD_logistic_regression_feature_coefficients.png`

Contact Pantelis Z. Lappas

Prediction

POST `/predict` Predict

Schemas

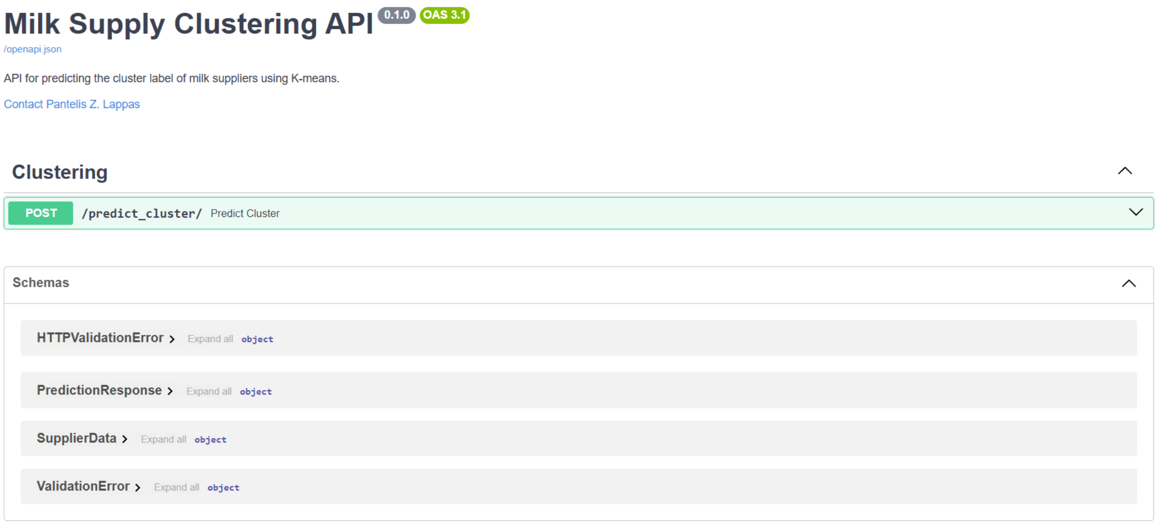
HTTPValidationError › Expand all object

InputData › Expand all object

MilkData › Expand all object

PredictionResponse › Expand all object

Figure 5-57 Prediction API



Milk Supply Clustering API 0.1.0 OAS 3.1

/openapi.json

API for predicting the cluster label of milk suppliers using K-means.

Contact Pantelis Z. Lappas

Clustering

POST `/predict_cluster/` Predict Cluster

Schemas

HTTPValidationError › Expand all object

PredictionResponse › Expand all object

SupplierData › Expand all object

ValidationError › Expand all object

Figure 5-58 Clustering API

The prediction service accepts input in the form of a JSON object containing a list of milk collection records under the data field (see Table 5.13). Each record includes detailed information such as the collection date, supplier ID, milk type, quantity, fat and protein content, truck and route details, sample measurements (pH, temperature), and fraud detection indicators (cow, water, and goat fraud markers). For instance, a typical request may consist of multiple entries for goat milk samples collected on a specific date, with associated laboratory and logistic information. Upon receiving the request, the service processes the features and returns a JSON response containing the predicted fraud risks in terms of probabilities and indicators, as well as path for getting useful plots towards explainable AI (see Table 5.14).

Table 5.13 Example of a JSON request (prediction service)

JSON Request

{
 "data": [
  {

JSON Request

```
    "collection_date": "2024-04-03",
    "supplier_id": "S000001",
    "milk_type": "goat",
    "quantity": 37.56,
    "fat": 5.17,
    "protein": 3.62,
    "truck_plate": "TP0362",
    "route": "Macedonia",
    "compartment_id": "TP0362_C19",
    "icebowl_id": "IB3518",
    "sample_barcode_comp": "SBC9696",
    "sample_barcode_ice": "SBI4523",
    "pH_comp": 4.55,
    "pH_ice": 4.69,
    "temperature_comp": 4.27,
    "temperature_ice": 3.04,
    "cow_fraud_comp": 0,
    "cow_fraud_ice": 0,
    "water_fraud_comp": 1,
    "water_fraud_ice": 1,
    "goat_fraud_comp": 0,
    "goat_fraud_ice": 0,
    "goat_percentage_comp": 50.16,
    "goat_percentage_ice": 45.53,
    "createdAt": "2025-02-24T13:14:22.295Z",
    "updatedAt": "2025-02-24T13:14:22.522Z"
   },
   {
    "collection_date": "2024-04-03",
    "supplier_id": "S000001",
    "milk_type": "goat",
    "quantity": 37.56,
    "fat": 5.17,
    "protein": 3.62,
    "truck_plate": "TP0362",
    "route": "Macedonia",
    "compartment_id": "TP0362_C19",
    "icebowl_id": "IB1313",
    "sample_barcode_comp": "SBC9696",
    "sample_barcode_ice": "SBI4523",
    "pH_comp": 4.55,
    "pH_ice": 4.69,
    "temperature_comp": 4.27,
    "temperature_ice": 3.04,
    "cow_fraud_comp": 0,
    "cow_fraud_ice": 0,
    "water_fraud_comp": 1,
    "water_fraud_ice": 1,
    "goat_fraud_comp": 0,
    "goat_fraud_ice": 0,
    "goat_percentage_comp": 50.16,
    "goat_percentage_ice": 45.53,
    "createdAt": "2025-02-24T13:14:22.295Z",
    "updatedAt": "2025-02-24T13:14:22.522Z"
   }
 ]
 }
```

Table 5.14 Example of a JSON response (prediction service)

JSON response

```
{
  "prediction_results": [
    {
      "collection_date": "2024-05-03",
      "compartment_id": "TP0307_C12",
      "total_quantity_of_goat_milk": 0.0,
      "total_quantity_of_sheep_milk": 37.22,
      "area": "Peloponnese",
      "ph_comp": 6.66,
      "temperature_comp": 8.0,
      "goat_percentage_comp": 73.96,
      "avg_goat_milk_ph_ice": 0.0,
      "avg_sheep_milk_ph_ice": 6.89,
      "avg_goat_milk_temperature_ice": 0.0,
      "avg_sheep_milk_temperature_ice": 6.0,
      "avg_goat_milk_fat_ice": 0.0,
      "avg_sheep_milk_fat_ice": 8.02,
      "min_goat_milk_fat_ice": 0.0,
      "min_sheep_milk_fat_ice": 8.02,
      "max_goat_milk_fat_ice": 0.0,
      "max_sheep_milk_fat_ice": 8.02,
      "avg_goat_milk_protein_ice": 0.0,
      "avg_sheep_milk_protein_ice": 6.16,
      "min_goat_milk_protein_ice": 0.0,
      "min_sheep_milk_protein_ice": 6.16,
      "max_goat_milk_protein_ice": 0.0,
      "max_sheep_milk_protein_ice": 6.16,
      "collection_month": 5,
      "collection_weekday": 4,
      "area_longitude": 22.364,
      "area_latitude": 37.5081,
      "cow_fraud_probability": 0.12478804636903411,
      "goat_fraud_probability": 0.7586403658245512,
      "water_fraud_probability": 0.10058381870038577,
      "cow_fraud_indicator": 0,
      "goat_fraud_indicator": 1,
      "water_fraud_indicator": 0
    },
    {
      "collection_date": "2024-06-20",
      "compartment_id": "TP0362_C12",
      "total_quantity_of_goat_milk": 40.46,
      "total_quantity_of_sheep_milk": 0.0,
      "area": "Macedonia",
      "ph_comp": 7.44,
      "temperature_comp": 6.57,
      "goat_percentage_comp": 94.09,
      "avg_goat_milk_ph_ice": 7.44,
      "avg_sheep_milk_ph_ice": 0.0,
      "avg_goat_milk_temperature_ice": 5.55,
      "avg_sheep_milk_temperature_ice": 0.0,
      "avg_goat_milk_fat_ice": 4.65,
      "avg_sheep_milk_fat_ice": 0.0,
      "min_goat_milk_fat_ice": 4.65,
      "min_sheep_milk_fat_ice": 0.0,
      "max_goat_milk_fat_ice": 4.65,
      "max_sheep_milk_fat_ice": 0.0,
      "avg_goat_milk_protein_ice": 4.4,
```

JSON response

```
      "avg_sheep_milk_protein_ice": 0.0,
      "min_goat_milk_protein_ice": 4.4,
      "min_sheep_milk_protein_ice": 0.0,
      "max_goat_milk_protein_ice": 4.4,
      "max_sheep_milk_protein_ice": 0.0,
      "collection_month": 6,
      "collection_weekday": 3,
      "area_longitude": 22.9482,
      "area_latitude": 40.6401,
      "cow_fraud_probability": 0.12799668078256268,
      "goat_fraud_probability": 0.00029253731343283583,
      "water_fraud_probability": 0.11015270491543602,
      "cow_fraud_indicator": 0,
      "goat_fraud_indicator": 0,
      "water_fraud_indicator": 0
    },
    {
      "collection_date": "2024-07-08",
      "compartment_id": "TP0307_C14",
      "total_quantity_of_goat_milk": 0.0,
      "total_quantity_of_sheep_milk": 70.2,
      "area": "Peloponnese",
      "ph_comp": 5.77,
      "temperature_comp": 3.95,
      "goat_percentage_comp": 94.83,
      "avg_goat_milk_ph_ice": 0.0,
      "avg_sheep_milk_ph_ice": 5.55,
      "avg_goat_milk_temperature_ice": 0.0,
      "avg_sheep_milk_temperature_ice": 2.16,
      "avg_goat_milk_fat_ice": 0.0,
      "avg_sheep_milk_fat_ice": 6.58,
      "min_goat_milk_fat_ice": 0.0,
      "min_sheep_milk_fat_ice": 6.58,
      "max_goat_milk_fat_ice": 0.0,
      "max_sheep_milk_fat_ice": 6.58,
      "avg_goat_milk_protein_ice": 0.0,
      "avg_sheep_milk_protein_ice": 5.25,
      "min_goat_milk_protein_ice": 0.0,
      "min_sheep_milk_protein_ice": 5.25,
      "max_goat_milk_protein_ice": 0.0,
      "max_sheep_milk_protein_ice": 5.25,
      "collection_month": 7,
      "collection_weekday": 0,
      "area_longitude": 22.364,
      "area_latitude": 37.5081,
      "cow_fraud_probability": 0.12582760614241714,
      "goat_fraud_probability": 0.7576257069834803,
      "water_fraud_probability": 0.09697224769092724,
      "cow_fraud_indicator": 0,
      "goat_fraud_indicator": 1,
      "water_fraud_indicator": 0
    }
  ],
  "explainability_plots": {
    "COW_FRAUD_contribution_box_plot":
"Explainability/COW_FRAUD_contribution_box_plot.png",
    "COW_FRAUD_feature_contribution_heatmap":
"Explainability/COW_FRAUD_feature_contribution_heatmap.png",
    "COW_FRAUD_logistic_regression_feature_coefficients":
"Explainability/COW_FRAUD_logistic_regression_feature_coefficients.png",
```

JSON response

```
    "COW_FRAUD_mean_contribution_bar_plot":
"Explainability/COW_FRAUD_mean_contribution_bar_plot.png",
    "COW_FRAUD_probability_evolution_plot":
"Explainability/COW_FRAUD_probability_evolution_plot.png",
    "GOAT_FRAUD_feature_importance_based_on_shap_values":
"Explainability/GOAT_FRAUD_feature_importance_based_on_shap_values.png",
    "GOAT_FRAUD_shap_bar_plot": "Explainability/GOAT_FRAUD_shap_bar_plot.png",
    "GOAT_FRAUD_shap_summary_plot":
"Explainability/GOAT_FRAUD_shap_summary_plot.png"
  }
}
```

The clustering service accepts a JSON payload containing detailed milk supply data for individual suppliers, which helps categorize them into relevant clusters based on their milk quality and fraud incidents (see Table 5.15). This comprehensive input allows the clustering service to analyse patterns and categorize suppliers into distinct groups. In particular, the JSON response includes the supplier ID along with the assigned cluster number (see Table 5.16).

Table 5.15 Example of a JSON request (clustering service)

JSON request

```
[
  {
    "supplier_id": "S001",
    "avg_quantity_goat_milk": 18.1,
    "avg_quantity_sheep_milk": 2.6,
    "max_quantity_goat_milk": 20.5,
    "max_quantity_sheep_milk": 3.1,
    "min_quantity_goat_milk": 15.2,
    "min_quantity_sheep_milk": 1.9,
    "avg_goat_milk_fat_ice": 4.2,
    "avg_sheep_milk_fat_ice": 5.1,
    "max_goat_milk_fat_ice": 4.8,
    "max_sheep_milk_fat_ice": 5.5,
    "min_goat_milk_fat_ice": 3.9,
    "min_sheep_milk_fat_ice": 4.7,
    "avg_goat_milk_protein_ice": 3.5,
    "avg_sheep_milk_protein_ice": 4.0,
    "max_goat_milk_protein_ice": 3.8,
    "max_sheep_milk_protein_ice": 4.3,
    "min_goat_milk_protein_ice": 3.1,
    "min_sheep_milk_protein_ice": 3.7,
    "avg_goat_milk_ph_ice": 6.5,
    "avg_sheep_milk_ph_ice": 6.6,
    "max_goat_milk_ph_ice": 6.8,
    "max_sheep_milk_ph_ice": 6.9,
    "min_goat_milk_ph_ice": 6.2,
    "min_sheep_milk_ph_ice": 6.3,
    "avg_goat_milk_temperature_ice": 4.5,
    "avg_sheep_milk_temperature_ice": 4.6,
    "max_goat_milk_temperature_ice": 5.0,
    "max_sheep_milk_temperature_ice": 5.1,
    "min_goat_milk_temperature_ice": 3.9,
    "min_sheep_milk_temperature_ice": 4.0,
    "number_of_cow_fraud_incidents_detected_in_sheep_milk": 0,
    "number_of_water_fraud_incidents_detected_in_sheep_milk": 1,
    "number_of_goat_fraud_incidents_detected_in_sheep_milk": 0,
    "number_of_cow_fraud_incidents_detected_in_goat_milk": 0,
    "number_of_water_fraud_incidents_detected_in_goat_milk": 1
  }
]
```

Table 5.16 Example of a JSON response (clustering service)

JSON response

```
[
 {
  "supplier_id": "S001",
  "cluster": 2
 }
]
```

5.8 Conclusions and Next Steps

ALLIANCE led to the successful development and deployment of a complete milk quality assessment and fraud detection platform. Two RESTful API services were built: one for real-time fraud prediction based on sample data, and another for clustering suppliers according to risk and quality profiles. FastAPI was used to expose these services, enabling easy and scalable access. Furthermore, to support monitoring analysis, Apache Superset was implemented, offering dynamic dashboards that make it easy to visualize key metrics, track model outputs, and detect trends over time. At the same time, Apache NiFi orchestrated the data flows between components, automating ingestion, processing, and service triggering.

To facilitate model development and testing, high-quality synthetic datasets were created, mimicking real-world complexity and variability. Special focus was given to model explainability throughout the project. Techniques like SHAP were used to interpret the contribution of each feature to predictions, while logistic regression coefficients offered a transparent, easily understandable view of the factors influencing outcomes.

It is worth mentioning that the combined architecture demonstrates the potential for an end-to-end, scalable solution for milk quality control and supplier risk assessment in operational/production environments. Looking ahead, the focus should now shift to integrating real-world datasets to replace synthetic examples, allowing the models to capture genuine supplier behaviour and fraud patterns. Additionally, launching pilot programs with dairy cooperatives or quality control partners will be critical to evaluating system performance under operational conditions.

6 Consumer Demand Assessment and Strengthening

6.1 Introduction

The objective of Task 3.6 is to assess and strengthen consumer demand. In this task, our primary focus was on investigating consumers' purchase intentions regarding products tracked by blockchain technology. This deliverable builds upon the approach described in D3.2. Specifically, Theory of Planned Behaviour was used for predicting consumers purchase intentions, and the collected data were analysed using structural equation modelling. To avoid repetition, we do not elaborate on the theoretical framework or methodology in this section. Detailed information about the data analysis is provided in the following sections.

6.2 Experimental design and implementation

6.2.1 Data Collection and (pre-)processing

The tool used for data collection in this study was an online questionnaire developed on the LimeSurvey platform. The LimeSurvey tool (https://www.limesurvey.org/) is a popular free and open-source online survey tool providing a web interface for creating surveys, managing users and participants, collecting responses and exporting data for analysis (Nyumba et al., 2022).

3500 questionnaires were completed in six countries, including Italy, Greece, Spain, France, Croatia, and Serbia. 500 respondents completed the questionnaire for each case study. Our objective is to explain product purchasing intentions using the extended Theory of Planned Behavior (TPB) rather than investigate the decision of whether or not to consume the product. Therefore, all respondents met the criteria of having consumed the product and being responsible for food purchasing within their households. In addition, the participants in the study were between the ages of 18 and 70. The age range of 18 to 70 years was chosen for this study to include a wide range of adult consumers who are legally capable of making their own purchase decisions and are likely to use new technologies. A quota sampling was used to guarantee representation across various demographic groups. Specifically, quotas were set for gender and age groups.

The questionnaire was initially developed in English, then translated into the respective local languages, and further tailored for each case study to reflect the specific product and country context. The questionnaires were distributed in December 2024 and January 2025 through a consumer panel of a market research institute. These panels give researchers access to various populations in several countries, allowing them to control samples and target particular demographics. Because of their larger participant base, which enables faster data gathering and wider generalizability, they are ideal for studies that need targeted or widespread sampling (Moss et al., 2023). It is noteworthy that this study was carried out in compliance with ethical guidelines and was approved by the German Association for Experimental Economic Research e.V. (GfeW), with approval number Invoice E-2024-12-10-000963. The following section presents the questionnaire used in the case study of organic pasta as an example.

1.Gender

- Male
- Female
- More

- I prefer not to answer

2. Birth year

3. Region of residence

4. Education

- Elementary school
- Secondary school
- Degree
- Postgraduate

5. Occupation

- Student
- Employed / Self-employed
- Not employed / Unemployed / Inactive population
- Retired
- Other condition

6. Family members (including you) (insert number)

7. How many people under the age of 18 are in your household? (insert number)

8. Thinking about the food purchases that are made in the family, who takes care of them?

- Almost always she
- You in particular but, to a lesser extent, also someone else
- Somebody else in particular and sometimes even you
- Only someone else

9. Which of these answers best describes the economic situation of your household?

- I have to be very careful about what I spend, sometimes my income is not enough for necessary purchases
- With a little prudence, I can, from time to time, afford some small luxuries
- We don't have financial problems and when I feel like buying something I do it
- I prefer not to answer

10. What is your knowledge of food traceability?

- I have a thorough knowledge of food traceability
- I have a basic understanding of food traceability
- I've heard the term but I don't know what it is
- I have never heard of food traceability

Traceability

Agri-food traceability is a system that monitors the entire journey of a food product and has been mandatory throughout the European Union since 2005. It ensures the quality and integrity of food at all stages of the supply chain, from production to consumption, through the detailed recording of the production process by each operator involved. It also applies to certified products, such as organic farming designations of origin. All organizations that handle food are obliged to adopt an internal traceability system that ensures the minimum requirements of:

- Tracking the entire life of the product by recording relevant events,
- tracing the product, that is, being able to go back through the food product production process to find the causes of a quality and safety problem discovered at a later stage.

11.What is your knowledge of food traceability systems based on blockchain technology?

- have in-depth knowledge of blockchain technology for tracking systems
- have a basic understanding of blockchain technology for tracking systems
- I've heard the term blockchain, but I don't know what it is
- have never heard of blockchain technology

Blockchain

Blockchain technology refers to the use of innovative technologies to help manage food traceability information. Blockchain provides a single, secure, transparent and unalterable record of the food supply chain, ensuring greater accuracy, trust and accountability in tracking product information from farm to fork. Once traceability information is recorded, the information cannot be changed

The questionnaire focuses on quality certifications such as organic, protected geographical indication (PGI), and protected designation of origin (PDO). Below is an explanation of these characteristics.

- Organic farming is an agricultural production system defined and regulated at the EU level by Regulation (EU) No. 2018/848. It does not use synthetic chemicals (fertilizers, herbicides, insecticides, fungicides) to fertilize the soil, control weeds, animal pests and plant diseases; it also prohibits the use of genetically modified organisms (GMOs). Resorts to traditional, essentially preventive practices, selecting local disease-resistant species and intervening with appropriate cultivation techniques
- The names of products registered as PDOs are those that have the strongest connection to the place of production. Every part of the production, processing and preparation process must take place in the specific region.
- The PGI emphasizes the relationship between the specific geographical region and the name of the product when a particular quality, reputation or other characteristic is essentially attributable to its geographical origin. For most products, at least one of the stages of production, processing or preparation takes place in the region.

12.How often do you purchase these types of pasta?

Always Often Sometimes Rarely Never

3.1. Conventional dough

3.2. Organic pasta

3.3. PDO/PGI Pasta

13. On a scale of 1 to 5, please indicate your opinion of organic products Very negative | Negative | Neither positive nor negative Positive | | Very positive

14.Intentions

Indicate your level of agreement/disagreement with the following statements:

(Totally disagree | Moderately disagree | Neutral | Moderately agree | Totally agree)

- When blockchain-traceable pasta becomes available, I intend to purchase it
- When blockchain-traceable paste becomes available, I will look for it and consider buying it
- When blockchain-traceable paste becomes available, I will be inclined to buy it

15. Subjective rules

Indicate your level of agreement/disagreement with the following statements: (Totally disagree | Moderately disagree | Neutral | Moderately agree | Totally agree)

- I would buy tracked pasta with the support of blockchain technology because my partner, family and friends would approve of this choice
- I would buy pasta tracked through blockchain technology because scientists say it is beneficial
- I would buy pasta tracked through blockchain technology because the media (TV radio social) is supportive

16.PBC

Indicate your level of agreement/disagreement with the following statements: (Totally disagree | Moderately disagree | Neutral | Moderately agree | Totally agree)

- feel able to easily find food products tracked by blockchain in stores
- think it is easy to use apps or online tools to verify food traceability using blockchain
- think it is easy for me to follow the food production chain thanks to blockchain

17. Attitude

Indicate your level of agreement/disagreement with the following statements: (Totally disagree | Moderately disagree | Neutral | Moderately agree | Totally agree)

- With the use of blockchain, pasta traceability information is more secure
- The origin of the paste tracked with the support of blockchain technology is always transparent
- Paste information with blockchain technology support is more authentic

18. Confidence in quality certifications

Indicate your level of agreement/disagreement with the following statements: (Totally disagree | Moderately disagree | Neutral | Moderately agree | Totally agree)

- Companies always comply with quality certification standards
- Companies provide consumers with transparent information about quality certification
- Certified quality product information is always true
- Traceability information is always reliable

19. Attitude toward technology

Indicate your level of agreement/disagreement with the following statements: (Totally disagree | Moderately disagree | Neutral | Moderately agree | Totally agree)

- I am optimistic about the innovative impact of technology
- I feel comfortable becoming familiar with the technology

- I believe that the adoption of technology can generate significant improvement in security and information

20. Where do you usually buy pasta?

- Supermarket
- Local market
- Online
- Specialty stores
- Organic store
- Agricultural cooperative
- Other (specify)

21. Using a scale of 0 to 10, how likely are you to recommend the product Organic Pasta to your family/friends/acquaintances?

0 1 2 3 4 5 6 7 8 9 10

22. You can choose from:

One package of organic pasta (500g) tracked by regular tracking system at a cost of €2.40

A package of organic pasta (500g) at PREMIUM PRICE, with the words "tracked with blockchain technology" on the label

- I would buy the package of organic pasta (500g) tracked with regular tracking system for €2.40
- I would buy the package of organic pasta (500g) tracked with blockchain technology for PREMIUM PRICE

23. When you buy pasta, do you look for information about the production process? (e.g., bronze-drawn and/or slowly dried at low temperatures)

- Never
- Rarely
- Sometimes
- Often
- Always

6.2.2 Data Analysis

6.3 Key results

Results from case studies on olive oil, pasta, feta cheese, fava beans, honey, potatoes, and raspberries in Italy, Greece, Spain, France, Croatia, and Serbia respecitively indicate that consumers in these countries have a moderate understanding of food traceability systems. A relatively small percentage of respondents (between 9.6% and 24.4% of those surveyed) show a more in-depth understanding, while more respondents have a basic knowledge. Additionally, consumers' understanding of traceability systems based on blockchain is lower than their knowledge of traceability.

According to the results of the hypothesis test, consumers' intention to use traceability systems for the food products under study is consistently influenced by their attitude toward technology and subjective norms (social influence). It means that when consumers have favourable opinions about technology and believe that products tracked by blockchain are socially acceptable, they are more likely to buy products tracked by blockchain technology.

However, Attitudes Towards Blockchain-based traceability and Perceived Behavioural Control (PBC) show different effects across product categories, suggesting that these factors may be more influenced by product-specific characteristics.

Interestingly, consumers' intention to buy products tracked by blockchain technology is not statistically significantly influenced by their trust in quality certification, except for potatoes and raspberries, where there is a slight but negative impact. This suggests that established quality certifications are either ignored or perhaps seen as inadequate for certain products, which would prevent further adoption of traceability.

Regarding willingness to pay for blockchain label, the findings indicate that, for all products under study, consumers prefer conventional traceability systems (for example, QR) to blockchain-based traceability. Specifically, between roughly 67% and 84% of consumers preferred a traditional traceability system, which was free, while only 16% to 33% said they preferred to pay a premium for products that use blockchain technology in their manufacturing process. It can be due the lack of familiarity of consumers with blockchain technology.

Consumers' loyalty related to the products was evaluated based on a 0 to 10 scale. Customers were asked if they would recommend the product- highlighting its specific quality attribute-such as, organic pasta to their friends or family. The results show that they were generally very loyal to these products, and a significant portion of consumers stated they are very likely or most likely to recommend them to others. This is especially true for PDO feta cheese, PDO Arilje raspberries, and PGI Lika potatoes.

Overall, these findings suggest that while traceability is valued, the adoption and premium pricing of blockchain-based traceability systems may be limited by low consumer awareness and understanding, highlighting the need for targeted education and communication strategies to enhance acceptance and perceived value.

6.4 Validation and final implementation

6.4.1 Assessment of consumer perception and behaviour

In this study, we investigated the factors that influence consumers' intention to purchase products tracked by blockchain technology. The results highlight important factors influencing consumer behaviour and offer practical implications for marketers and policymakers seeking to promote the adoption of blockchain technology in the food industry.

The positive impact of attitudes toward technology (TEC) shows that people who have a positive attitude toward technology are more willing to purchase products tracked by blockchain. This result underlines the importance of technology awareness and offering educational initiatives. This finding is consistent with the results of studies by Lin et al. (2021), Dang & Tran (2020), Contini et al. (2023). This presents a valuable opportunity for companies to develop marketing strategies that display the transparency, security and innovation of blockchain technology. In this way, companies can build consumer confidence and encourage wider adoption. For instance, informing customers about how blockchain guarantees product authenticity and traceability may attract tech-oriented individuals who appreciate innovation and openness in their food choices.

The study found that Attitudes Towards Blockchain (ATB) has significant influence on purchase intention for all case studies except of Raspberries. This result contradicts the findings of previous studies by Dang & Tran (2020) and Prisco et al. (2022). These studies also found that general attitudes towards a product do not always translate into purchase behaviour, especially in contexts where consumers do not fully understand or appreciate the perceived benefits. However, this finding is in line with the results of Dionysis et al. (2022), who postulated that a positive attitude towards traceability and transparency in the food industry is a good predictor of purchase intention. The divergence in results may be attributed to contextual differences or the presence of blockchain technology features that consumers have not yet fully understood. Even if consumers favour the concept of traceability, this does not necessarily mean they are motivated to buy Raspberries with blockchain traceability. This suggests a disconnect between attitudes and actions, with consumer attitudes not always translating into actual purchasing behaviour. Further research could explore how this gap can be bridged by linking blockchain traceability to more directly perceived benefits such as food safety, quality assurance and environmental sustainability.

Perceived Behaviour control (PBC)was identified as an important predictor of purchase intention for Olive oil, pasta, Fava beans, Potatoes, and Raspberries suggesting that consumers who believe they have the ability and resources to identify and use blockchain-traceable products are significantly more likely to express a purchase intention. This finding is consistent with the results of studies by Lin et al. (2021), Dang & Tran (2020), Contini et al. (2023) and Prisco et al. (2022), which have shown that PBC plays a central role in influencing consumer intentions, especially in the context of new technology adoption. This result suggest that consumers are more likely to purchase products that are easy to access and use. Therefore, companies should prioritise the development of blockchain-based tracking tools that are user-friendly and accessible. For example, companies can develop simple apps or digital tools that allow consumers to effortlessly check the authenticity of products.

In addition, the results of this study show that subjective norms play a crucial role in consumers' intention to purchase all the products under study that are tracked by blockchain, namely olive oil, pasta, fava beans, feta cheese, potatoes, honey, and raspberries. This result is consistent with the theory of planned behaviour, which postulates that the approval and support of significant others, e.g. family, friends and social networks, can strongly influence a person's behavioural intentions (Ajzen, 1991). The results regarding Subjective Norms suggest that social acceptance can be one of the effective drivers of consumer purchase intent for the products being tracked by blockchain. As a result, marketing strategies could successfully cause consumer interest by incorporating social proof, such as recommendations from experts, influencers, and leaders in the food industry. Furthermore, the implementation of educational programs that share information about the advantages of blockchain technology, supported by credible individuals like scientists and experts in food safety, may serve to further solidify societal norms surrounding the purchase of such products.

Concerning trust in quality certification, although the effect of trust in quality certification for olive oil, pasta, feta cheese, fava beans, and honey was not statistically significant which is aligns with the result reported by Contini et al. (2023), it had negative and statistically significant in case studies of potatoes and raspberries. The negative effect of trust in quality certification to consumers intention to buy potatoes and raspberries tracked by blockchain suggest that established quality certifications are either ignored or perhaps seen as inadequate for certain products, which would prevent further accepting of traceability. The lack of emphasis on the role of trust suggests that consumers may not perceive blockchain technology as a natural extension of existing quality certification systems. An alternative explanation is that respondents may have a high level of trust in traditional certifications but do not perceive the value of blockchain technology as being enhanced by them. This emphasises the need for clear communication about how blockchain can complement and enhance quality certification by providing additional layers of transparency and authenticity beyond traditional systems.

6.4.2 Lessons learned and policy implication

This study provides new insights into the factors influencing consumers' intention to buy blockchain-labelled products in Italy, Spain, Greece, France, Croatia, and Serbia. The findings suggest that successful marketing strategies should focus on educating consumers about the benefits of blockchain, simplifying the user experience, and leveraging social influences to drive the adoption of blockchain-based traceability. The findings suggest that successful marketing strategies should focus on educating consumers about the benefits of blockchain, simplifying the user experience, and leveraging social influences to drive the adoption of blockchain-based traceability.

The study highlights that while attitudes towards the technology and subjective norms positively influence consumer purchase intentions, attitudes towards blockchain-based traceability and perceived behavioural control vary depending on the product. This suggests that there are still gaps in consumer knowledge and perceptions, emphasizing the need for clearer communication about the practical benefits of blockchain technology.

These findings have important implications for both policymakers and producers in the agri-food sector. For policymakers, the positive association between attitudes toward technology and the intention to purchase blockchain-tracked products underscores the necessity of fostering technological awareness and digital literacy among consumers. Public policy initiatives should prioritize educational campaigns and outreach programs that elucidate the benefits of blockchain technology, particularly its contributions to transparency, security, and traceability in the food supply chain. Such measures may facilitate broader public acceptance and adoption of innovative traceability systems. Furthermore, the findings suggest that subjective norms-namely, the influence of peers, experts, and authoritative figures-play a significant role in shaping consumer behavior. Policymakers can leverage this by engaging credible stakeholders, such as scientists and food safety authorities, in communication strategies to reinforce positive societal attitudes toward blockchain-enabled traceability.

For producers, the results highlight the importance of integrating user-friendly and accessible blockchain-based tracking solutions. Since perceived behavioral control (PBC) is a significant predictor of purchase intention, companies should invest in the development of intuitive digital tools, such as mobile applications or QR code systems, that enable consumers to easily verify product authenticity and traceability. Marketing strategies should emphasize the unique benefits of blockchain, including enhanced transparency and authenticity, to appeal to technologically inclined consumers. Additionally, the study suggests that the impact of trust in traditional quality certifications on the adoption of blockchain-tracked products is limited or even negative in certain cases (e.g., potatoes and raspberries). This indicates a potential disconnect between established certification schemes and emerging technological solutions. Producers should therefore focus on clearly communicating how blockchain technology complements and extends existing quality assurance systems, rather than merely replicating traditional certification approaches.

In summary, the findings advocate for a coordinated approach in which policymakers advance educational and regulatory frameworks to support blockchain adoption, while producers prioritize consumer-oriented, accessible solutions and transparent communication. Such efforts are essential to bridging the gap between consumer attitudes and actual purchasing behaviour, ultimately fostering greater acceptance and diffusion of blockchain technology within the agri-food sector. In summary, this study contributes to the literature on consumer behaviour towards new food technologies by providing a framework for the use of blockchain technology to meet consumer expectations in the food industry. While blockchain can potentially increase trust in existing quality signals, the challenge is effectively communicating its benefits to consumers. By recognising the importance of social norms, attitudes towards technology and perceived behavioural control, stakeholders can promote transparency, accountability and sustainability in the agri-food industry, creating a more efficient and competitive environment.

6.5 Conclusion

To conclude, this study identifies several key factors influencing consumers' intentions to purchase food products tracked by blockchain technology, including attitudes toward technology, attitudes toward blockchain, perceived behavioural control, and subjective norms. The findings show that positive attitudes toward technology and social influences generally encourage the purchase of blockchain-tracked products, while the impact of attitudes toward blockchain and perceived behavioural control varies depending on the product category. Additionally, trust in traditional quality certifications does not consistently increase purchase intention for blockchain-tracked products; in some cases, such as with potatoes and raspberries, higher trust in these certifications was associated with a decrease in consumers' intention to buy blockchain-tracked items. These findings suggest the need for targeted educational initiatives to raise consumer awareness about the benefits of blockchain technology, as well as the development of user-friendly traceability tools. Policymakers and industry stakeholders should collaborate to clearly communicate how blockchain can complement existing quality certifications and address specific consumer concerns. Future research could further explore how to bridge the gap between positive attitudes and actual purchasing behaviour, ensuring that blockchain-based solutions are both accessible and valued by consumers. Ultimately, fostering trust, transparency, and digital literacy will be essential for the successful integration of blockchain technology in the agri-food sector.

7 Conclusions

7.1 Recap of Key Achievements

The DNA-based authentication and traceability tool has significantly increased the accuracy of EVOO variety classification, especially through the discovery and validation of a new biomarker. The model's performance significantly improved highlighting its potential for reliable traceability in EVOO FSC. Both portable NIR and HSI spectroscopy models were successfully validated and demonstrated their capacity to detect fraudulent mixtures in PGI Asturian faba beans (detection of faba beans from Bolivia). In addition, preliminary model for identification of faba beans from different plots was also developed, although more samples are needed to enhance their reliability. A modular prototype of the ALLIANCE Digital Knowledge Base has been delivered, capable of ingesting and displaying structured knowledge about food fraud cases and prevention tools. Furthermore, the successful development and deployment of a comprehensive milk quality assessment and fraud detection platform showcases the robust scientific infrastructure behind ALLIANCE's tools, showcasing their potential for reliable, real-world application across various food sectors. Last but not least, the consumers demand assessment study identified key psychological and contextual factors that significantly influence consumer intentions to purchase blockchain solutions.

7.2 Limitations and challenges

Field trials revealed that the performance of the NIR-based spectroscopy tool can be affected by factors such as environmental conditions, sample handling, and calibration inconsistencies. Variables such as temperature fluctuations, calibration drift, and inconsistencies in sample preparation affected the tool's accuracy. Despite efforts, no significant improvements were found. Nevertheless, the huge dataset generated during the first round of demonstration will help the ongoing model optimisation.

Additionally, data availability always remains challenging task for the model development. As justified in section 5.2, the Food Fraud Predictive Analytics system was primarily built using synthetic datasets to effectively demonstrate the system's functionalities and capabilities. The dataset is structured around multiple variables related to the supply chain and quality control of feta cheese, with a focus on fraud detection. However, real data is necessary in order to ensure the models robustness and accuracy in operational environments, and this has been planned for the upcoming period (during pilots in WP4).

7.3 Final reflections and recommendations

ALLIANCE has established a good foundation towards digitalising the FSC processes and fraud prevention through the delivery of innovative tool that presented in this deliverable. In the coming months, end users and key stakeholders will validate the effectiveness of these solutions. Additionally, improvements will be made to enhance the system's responsiveness and its capacity to detect and address fraud across the entire food supply chain.