

Toward an Early Warning System: Feature Selection Mechanism for Food Fraud from Fishing Activities

Georgios A. Klados
Ubitech,
Chalandri, Greece
gklados@ubitech.eu

Aristeidis Tsitiridis
Ubitech,
Chalandri, Greece
atsitiridis@ubitech.eu

Nikolaos Chachampis
Ubitech,
Chalandri, Greece
nchachampis@ubitech.eu

Konstantinos Perakis
Ubitech,
Chalandri, Greece
kperakis@ubitech.eu

Abstract— Food fraud presents a considerable challenge to consumer safety, economic stability, and the integrity of global food supply chains. This study introduces a Feature Selection Mechanism for Fraud (FSMF) designed to proactively monitor and identify fraudulent activities within food supply chains, with a particular emphasis on the fish supply chain, and more specifically focusing on fishing stage. The system integrates IoT sensor data, environmental factors, and supply chain records, employing machine learning techniques for anomaly detection and feature extraction. The key methodologies used in this work include, Isolation Forest, Autoencoders, ARIMA, in addition to multi-modal data fusion and feature importance selection approaches. To illustrate the proposed system's effectiveness, a case study focusing on the Norwegian whitefish, utilizing Random Forest and XGBoost for vessel classification. Explainable AI techniques, such as SHAP analysis, are implemented to examine the impact of environmental and behavioral features on classification accuracy. The findings reveal a significant enhancement in fraud detection, achieving high precision and recall in differentiating fishing activities from other maritime operations.

Keywords—Food Fraud, Early Warning System, AI, Machine Learning, Data Fusion, Predictive Analytics, Food Adulteration.

I. INTRODUCTION

Food fraud is the intentional deception for economic gain involving food products, and it includes a wide range of illicit activities that undermine the integrity, safety, and authenticity of food [1]. Food fraud has emerged as a critical global issue that threatens consumer trust, public health, and the economic stability of food supply chains. Activities such as adulteration, mislabeling, counterfeiting, and dilution undermine the authenticity and safety of food products, often resulting in substantial financial and reputational damage [1],[2].

Common approaches to food fraud detection are predominantly reactive, identifying fraudulent activities only after products have entered the market or focusing in particular products or screening methods[3], [4], [5], [6]. Existing food fraud detection frameworks and commercial solutions including the Decernis Food Fraud Database¹, HorizonScan², and Foodakai³, largely rely on historical records, expert assessments, and regulatory inspections. They provide valuable food safety intelligence but are limited in their ability to offer real-time predictive capabilities.

The development of a holistic framework that screens different sample chains and throughout the different stages, such as cropping, transportation, packaging and storage, is a key concept to identify possible fraudulent activities in the

food industry. Such a holistic framework can be achieved with an Early Warning System [6], [7] that generates proactive alerts to the food authorities for further investigation.

An Early Warning System can analyze samples from various food supply chains efficiently, handle different data types and identify various fraudulent activities. The development of an effective EWS faces numerous challenges, including the inherent complexity of global food supply chains, the dynamic nature of fraudulent tactics, and regulatory inconsistencies across jurisdictions. Many instances of fraud occur deep within supply networks [2], complicating detection without robust data-sharing infrastructures [8]. Furthermore, fraudsters consistently adapt their strategies to exploit vulnerabilities in existing regulations and detection technologies [8]. An efficient, scalable, and adaptive EWS must continuously learn from emerging fraud patterns and integrate diverse datasets, including market demand trends, environmental factors, and supply chain vulnerabilities, to provide actionable insights to stakeholders. Regulatory fragmentation further complicates the implementation of a unified EWS, as countries enforce varying standards regarding food authenticity, safety, and traceability.

In the past, the research community explored methodologies such as Bayesian networks, Big Data analysis, and AI-driven approaches to identify patterns of fraud [9], [3], [10]. However, these models often depend on media-reported incidents and regulatory recalls, which are fundamentally retrospective. As fraudsters continuously evolve their strategies, there is an urgent need for a system that integrates real-time Internet of Things (IoT) sensor data, AI-based anomaly detection, and robust risk assessment frameworks to issue predictive fraud alerts [10].

The current paper presents Feature Selection Mechanism for Fraud (FSMF) acting as preliminary work towards the development of a holistic Early Warning System architecture. The designed mechanism focuses on the early stages of the data pre-processing, feature and anomaly extraction and has been evaluated on the fishing stages within the Norwegian White Fish supply chain.

II. EARLY WARNING SYSTEM FOR FOOD FRAUD

Fraud detection systems proactively identify, as inferred by the information available to them, data irregularities with great accuracy and speed [11]. Precision in EWSs is of paramount importance in food fraud detection. Each misclassification of fraud attempts whether a false alarm (false positive) or an overlooked risk (false negative), can have significant real-world consequences. At the same time,

¹ <https://www.decernis.com/>

² <https://www.fera.co.uk/>

³ <https://www.foodakai.com/>

unwarranted alerts strain resources and disrupt business operations, eroding trust in the system itself. Even more importantly, failing to detect an actual fraud incident can lead to compromised consumer safety, financial losses, and reputational damage across the supply chain, and ultimately the consumer market itself. Therefore, maintaining high precision also implies that warnings are both well-timed and justified, thus preventing wasteful responses. Moreover, flagging genuine suspicious activities allows EWSs to build stakeholder confidence, optimize resource allocation, and uphold the integrity of the food supply in the long run [12]. Another essential attribute of an EWS is adaptability. This key EWS attribute underpins the system's proactiveness by enabling it to evolve alongside shifting fraud tactics. Regardless of the EWS industry application [13], sustaining relevance against both current and emerging threats demands that adaptability by design be a core EWS requirement.

Fraudulent activities can emerge at any point in the supply chain and at every stage, during production, transport, or packaging, therefore a constant stream of real-time data can enable EWSs to pinpoint supply chain anomalies almost as they occur. Such a rapid feedback loop helps prevent contaminated or misrepresented products from reaching consumers and reduces the window in which fraud can go unnoticed [14], [15]. Recent technological breakthroughs from multimodal machine learning (ML) research [16] spurred synergistic approaches for combating food fraud together with blockchain [17], IoT [18], ML/AI [19], and big data [20]. Combined with advanced analytical techniques, like spectroscopy (e.g., near-infrared analysis) and DNA barcoding [21], rapid on-site testing of food products can be conducted before products reach consumers. This synergy of scientific fields gave rise to more sophisticated architectures adaptive enough to handle the multidimensional nature of emerging datasets efficiently, such as with the implementation of Bayesian Networks [22].

Constructing a precise, adaptive and rapid EWS for the food industry requires the fusion of data-driven intelligence and risk detection methods [23] so that emerging threats can be identified and mitigated before they cause significant financial or health harm. As a precursor to our fully fledged EWS, we have developed a multimodal data ingestion and feature extraction pipeline (FSMF), designed to handle the complexity of fraud detection tasks. This pipeline, which employs IoT data, retrospective industry data, and external datasets, consolidates disparate information into a unified framework and performs feature engineering to highlight subtle anomalies. Acquiring signals from environmental conditions, production metrics, and logistics records, equips the EWS with well-rounded information to detect suspicious events with heightened speed and accuracy. Consequently, the foundational step presented in the Methodology section sets the groundwork for an advanced EWS which will be both scalable and precise enough to address evolving threats across every stage of food production and distribution.

III. METHODOLOGY

Toward the development of a FSMF, we utilize methodologies derived from existing literature. In this

section, we present an overview of the implemented methodologies, which are vital for processing data from diverse sources, particularly those aimed at addressing fraudulent activities in the fishing stage.

A. Architecture

The proposed architecture for the FSMF is structured around the Processing Layer and Data Analysis – Feature & Anomaly Extraction module (Fig. 1).

The bottom level of the architecture consists of the proposed system's input data sources. The input layer will significantly contribute to fraud detection in food supply chains by integrating open-source and internal operational data. While challenges exist regarding data availability, the combination

of diverse sources enhances the system's effectiveness. Even though, the FSMF proposed is meant to be employed in all food supply chains, the primary focus of this work is the fish supply chain, more concretely, the fishing production stage. External data sources like open datasets, including *Open-Meteo*⁴ and datasets from European ministries related e.g. with the annual productions, provide critical contextual

insights. Internal supply chain data, such as IoT sensor readings (temperature, humidity, and storage conditions) and operational records (catch reports, transportation logs, and processing details), are essential for detecting anomalies and ensuring compliance with food safety standards. By merging these diverse datasets, the FSMF aims to improve detection accuracy, bolster fraud prevention efforts, and enhance transparency within the supply chain.

The Processing Layer is tasked with data preparation and integration from diverse sources, including IoT sensor data, supply chain records, and weather and climate open databases. This layer features a Data Consolidation and Data Preprocessing module responsible for addressing challenges like missing values, outlier detection, and statistical analysis using methods such as Isolation Forest and Spectral Residual [10], [18]. Effectively managing outliers is crucial, as they can lead to false positives; however, their removal can also create challenges, including the risk of discarding genuine cases. Furthermore, a Data Fusion module integrates multimodal data sources through concatenation and feature importance evaluation, enhancing the accuracy and robustness of predictive models. With a diversity of data types, the fusion process enhances decision-making

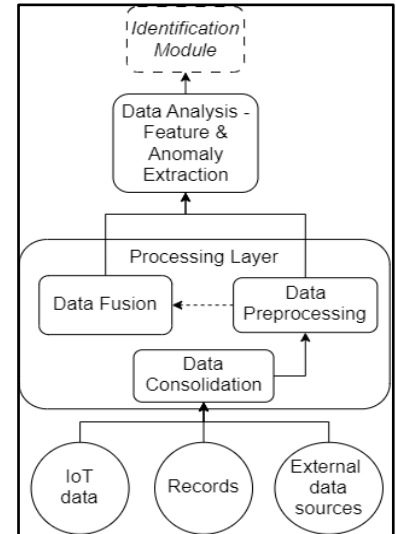


Fig. 1. Early Warning System Conceptual Architecture

⁴<https://open-meteo.com/>

capabilities even more and provides deeper insights into anomalies and potential risks.[8],[9].

The Data Analysis – Feature and Anomaly Extraction module is designed to extract features and detect anomalies to identify potential fraudulent indicators. It analyzes sensor data using both statistical and frequency-based methods. Additionally, geolocation and time-series data are utilized to uncover anomalies that may enhance a suspicious activity identification.

B. Methods

Data Cleaning

During data collection an important milestone is to ensure data are properly cleaned. As different datasets will be loaded into the FSMF, it is essential to prepare them in a format compatible with machine learning models. Specifically, each data source must be examined for duplicates, NaN (missing) values, and irrelevant records. Removing duplicates is necessary to prevent bias, and handling NaN values is equally important. This can be achieved either by removing them if they substantially impact a feature or sample, or by employing imputation techniques [25],[26].

Imputation techniques

Imputation techniques are a collection of methods used to replace NaN values with estimated values based on existing data, so that no additional bias is introduced into the dataset. Two widely used methods are k-nearest neighbors (kNN) imputation and Multiple Imputation by Chained Equations (MICE) [27]. kNN imputation involves projecting data points into the feature space and filling in missing values based on the values of the nearest data points. For numerical values, the missing value is determined as the average of the closest data points. For categorical variables, the value is selected based on the most frequently occurring category among the nearest data points. Conversely, MICE treat the missing value as a function, transforming the problem into a classification or regression task. This process is iteratively repeated to generate new samples in each cycle, with the results combined in the final step.

Outlier detection

This early study focuses on the enhancement of the FSMF within the food supply chain by employing three efficient anomaly detection techniques: Isolation Forest, Elliptic Envelope, and Spectral Residual. These methods are essential for identifying and addressing extreme values and true outliers. If these issues are not properly managed, they can significantly distort data analysis and lead to inaccurate decision-making, which could negatively impact both supply chain efficiency and food safety.

Isolation Forest proves particularly advantageous for detecting extreme deviations in multidimensional datasets, including irregular fluctuations in temperature logs, suspicious shipment movements and sensor-based quality measurements. By filtering out these anomalies, the system ensures that subsequent analyses are informed by reliable data, thereby mitigating the risk of false alarms within the FSMF [13],[14].

The Elliptic Envelope Module utilizes a statistical method based on the hypothesis that supply chain data follows a multivariate Gaussian distribution to identify anomalies. The EE Module effectively detects and flags outliers, such as

atypical supplier behaviors. This approach enhances fraud detection by systematically filtering out statistical anomalies that deviate from expected operational patterns. [13],[14]. Spectral Residual is applied to filter out sudden fluctuations in demand forecasting, sensor readings, and logistical performance metrics. The implementation of SR within a rolling window framework allows for continuous monitoring and the subsequent removal of short-term anomalies that could distort predictive models [28].

The integration of these three anomaly detection techniques allows the system to effectively filter out extreme values and genuine outliers prior to further processing, ensuring the usage of clean and reliable data.

Multi-modal Data Fusion

Detecting food fraud efficiently requires analyzing data from multiple sources, including IoT sensors, digital records, and open data repositories. To develop an FSMF, these diverse data streams must be integrated into a unified framework. Multi-modal data fusion is extremely important in this process, combining structured and unstructured data, improving anomaly detection and the overall decision-making accuracy [24].

In our approach, we utilize Concatenation-Based Fusion, which structures multiple datasets into a unified format, aligning different data sources under a shared reference point. This method efficiently merges data while preserving the integrity of the various data types. For example, in the whitefish supply chain under investigation, data from different types and domains such as ship movements, weather data, and on-board sensors, are considered making data fusion imperative. The utilization of these data implies the need of homogenizing different types of information into a unique, fully aligned dataset.

Other fusion strategies include Operation-Based Fusion, which uses mathematical and statistical operations to analyze interactions and reveal hidden patterns. Furthermore, techniques for Feature Importance Selection, such as Principal Component Analysis (PCA), help simplify datasets by reducing complexity while retaining critical variations. These variations make up the most meaningful features that contribute to anomaly detection [16],[17]. Overall, the integration of data fusion methods, leads to the enhancement of the efficiency and reliability of food supply chain [18],[19].

Feature Extraction

To enhance the efficacy of early warning systems within the food supply chain, we implement a comprehensive set of feature extraction techniques designed to convert raw data into informative representations. These engineered features substantially augment the capability of anomaly detection models to identify extreme values and genuine outliers across diverse data types, including sensor data, geolocation data, and temporal timestamps.

For sensor data, which is sourced from IoT devices tasked with monitoring critical parameters such as temperature and humidity, we apply feature engineering methodologies to emphasize pivotal statistical and spectral characteristics. Key statistical features, including mean, median, standard deviation, variance, skewness, and kurtosis, effectively encapsulate distributional traits. Additionally, range-based features such as minimum, maximum, and moving range

values, yield valuable insights into environmental fluctuations. The application of Fourier Transform Features, notably Fast Fourier Transform (FFT) and real FFT, facilitates the extraction of dominant frequency components, which assist in discerning cyclic patterns and abrupt disruptions [20],[21],[22]. The utilization of non-linear feature embeddings, such as t-SNE, enhancing the model's capacity to differentiate between normative and anomalous patterns[29].

In our study, concerning geolocation data, which encompasses GPS coordinates of shipments, an array of feature extraction methods is utilized to monitor anomalies in transportation. Distance-based features are computed by assessing consecutive distances between recorded GPS coordinates via the Haversine formula. Furthermore, velocity features are extracted by calculating speed based on the distance traveled over temporal intervals, while acceleration and deceleration metrics assist in identifying irregular navigation behaviors. Directional features, including changes in travel direction and variability in turning rates, play a crucial role in detecting routing behavior anomalies [30],[31],[32],[33]. The speed-direction coupling technique quantitatively measures interactions between speed and direction, thus facilitating the detection of unexpected deviations from anticipated transport patterns. Rolling window variability measures, like the standard deviation over a specified time window, serve to monitor fluctuations in vessel's movement patterns over time [34].

Timestamps are instrumental in providing critical temporal context essential for anomaly monitoring within the supply chain. Time decomposition techniques allow for the extraction of time-related features such as hour, day-of-week, month, and year to capture seasonal and periodic trends effectively. The application of sine and cosine transformations considers the cyclical characteristics of time, thus encoding periodic patterns. The inclusion of the time of last event feature, which measures the interval between consecutive timestamps, enables an analysis of shipment delays and operational inconsistencies [33].

The integration of extracted features into various anomaly detection models is expected to enhance their sensitivity to extreme values. To improve reliability, a majority voting mechanism aggregates predictions from multiple models, including Autoencoders, ARIMA, Isolation Forest, and STL, resulting in a more robust classification of anomalies [24],[28].

IV. EVALUATION

The Data collection methodology is based on the Norwegian Whitefish Supply Chain in order to gather data related to fishing vessels' positions. More details are fully described on the next paragraph.

A. Case Study: Norwegian Whitefish Supply Chain data

To assess the effectiveness of the FSMF, a case study was conducted on the Norwegian whitefish supply chain, wherein real-world data was analyzed to identify fraudulent activities such as misreporting of catches and unauthorized fishing in restricted areas. The dataset included vessel tracking data from *Barentswatch*⁵, which provided daily position reports of

vessels, in conjunction with environmental conditions from *Open-Meteo*⁴, offering essential contextual data for analyzing vessel movement patterns. The study encompassed a diverse array of vessel types, including fishing vessels, passenger ships, cargo ships, and tankers. The primary objective of this classification task was to discern fishing behavior based on vessel trajectories and environmental conditions. The fundamental assumption underlying this experiment is that fishing vessels exhibit behavior analogous to that of other vessel types when following a specific course (e.g., a trip to port), while their behavior alters during active fishing operations. By analyzing these movement patterns and integrating multi-modal data sources, the system demonstrated its capability to flag suspicious activities, thereby enhancing fraud detection and ensuring improved compliance within the seafood supply chain.

The final dataset used in this study comprises approximately 307 vessel trajectories, spanning a monitoring period of 24 hours. These trajectories were classified into two main classes as described previously: fishing vessels (n=179) and non-fishing vessels (n=128). Environmental data such as wind speed and direction were obtained from Open-Meteo and were synchronized with vessel data using a time-window alignment strategy using 30 minutes time windows interval. Each vessel trajectory contains geospatial coordinates, speed, directional changes, and corresponding environmental variables. Thus, the final dataset is imbalanced with a 71% ratio.

B. Experimental Setup

The input data are validated by data consolidation and data preprocessing modules concerning the elimination of missing values and carrying out the appropriate format for the processing. As previously mentioned, the dataset includes information from various domains, such as spatial and weather data. The data homogenization is addressed by the fusion module, applying concatenation. To ensure consistency between weather data and vessel positions, we implement weather windows of 30 minutes in duration. This approach helps us mitigate any potential discrepancies and enhances the reliability of our operations. The homogenized dataset led to the data analysis – Feature and Anomaly extraction module. The analysis process incorporated several features, specifically vessel movement patterns, environmental factors, and behavioral interaction features,

TABLE I. PERFORMANCE METRICS

Model		Metrics			
		Accuracy	Precision	Recall	f1
Random Forest	Fishing vessels	0.89	0.87	0.96	0.91
	Non-fishing vessels		0.94	0.80	0.86
XGBOOST	Fishing vessels	0.85	0.82	0.95	0.88
	Non-fishing vessels		0.91	0.71	0.80

each evaluated using mean and standard deviation values. Vessel movement patterns included aspects such as consecutive differences in position, consecutive directional

⁵ <https://www.barentswatch.no/>

TABLE II. CONFUSION MATRIX FOR XGBOOST

		<i>Predicted Labels</i>	
		Fishing vessels	Non-fishing vessels
<i>Actual Labels</i>	Fishing vessels	73	4
	Non-fishing vessels	16	39

TABLE III. CONFUSION MATRIX FOR RANDOM FOREST

		<i>Predicted Labels</i>	
		Fishing vessels	Non-fishing vessels
<i>Actual Labels</i>	Fishing vessels	74	3
	Non-fishing vessels	11	44

changes, speed, and turning rate, as described in the Features Extraction paragraph. Environmental factors encompassed elements like snowfall, wind direction, and wind speed measured at a height of 10 meters. Additionally, behavioral interaction features considered relationships such as speed-direction coupling.

The characteristics of the used dataset in this study establish limitations on the applicability of techniques. In particular, the tree-based models employed, such as Random Forest (RF) and XGBoost (XGB), are especially suited for capturing complex, non-linear feature interactions and accommodating mixed data types. This capability aligns well with the combined environmental and behavioral features present in the dataset. In contrast, distance-based models, such as SVM and kNN, are less effective in these scenarios due to their sensitivity to feature scaling and their limited ability to model non-linear relationships. Additionally, the relatively small size and the nature of the dataset do not support the effective application of deep learning approaches, which typically require larger and more diverse datasets to achieve successful generalization. The hyperparameters optimization was achieved by grid-search using 5 folds cross validation. More specifically, the hyperparameter grid of the Random Forest model included the number of estimators values of 50, 100, and 200, the maximum depth settings of None, 10, 20, and 30, the minimum split samples of 2, 5, and 10, the minimum leaf samples settings of 1, 2, and 4 and bootstrap options of True and False. The optimal performance was achieved with the following parameter configuration: `n_estimators` set to 500, `max_depth` set to 10, and `min_samples_split` set to 5, `min_samples_leaf` to 1 and `bootstrap` to False.

Similarly, the hyperparameter grid search for the XGBoost model explored several parameter values, including the number of estimators values at 100, 300, and 500; the learning rate values of 0.01, 0.1, and 0.2; the maximum depth settings of 3, 5, and 7 and subsample ratios of 0.6, 0.8, and 1.0. The best-performing configuration identified was `n_estimators` set to 100, `learning_rate` at 0.2, `subsample` at 0.6 and `max_depth` set to 5.

Both models were implemented using Python 3.9.2 and trained with Scikit-learn and XGBoost libraries.

C. Performance Metrics

Model performance was evaluated using metrics, such as accuracy, precision, recall, and F1-score, as summarized in Table I. Both the Random Forest (RF) and XGBoost (XGB)

models achieved high accuracy, with RF at 0.89 and XGB at 0.85. Both models showed exceptional recall in identifying non-fishing vessels, with RF achieving 0.96 and XGB achieving 0.95. This indicates that legal maritime activities were accurately recognized, which is essential for effective early-warning systems.

For fishing vessels, RF achieved 0.87 precision, 0.96 recall, and 0.91 F1-score, whereas XGB achieved 0.82 precision, 0.95 recall, and 0.88 F1-score. These results suggest that both models are highly effective in identifying fishing activities, particularly in minimizing false negatives, which is crucial in early-warning contexts to avoid missing instances of potential fraud.

In contrast, performance on the non-fishing vessel class was slightly lower. RF achieved 0.94 precision, 0.80 recall, and 0.86 F1-score, while XGB attained 0.91 precision, 0.71 recall, and 0.80 F1-score. These results show a higher likelihood of false negatives when identifying non-fishing activities, indicating that some legal maritime operations may be incorrectly flagged. While this could lead to unnecessary inspections, the precision values remain high enough to maintain trust in the alerting mechanism.

Finally, from an operational perspective, this trade-off might be acceptable, especially in high-risk supply chains where the costs of undetected fraud exceed the inconvenience of false alerts. Moreover, both models have demonstrated robustness



Fig. 2. xAI for Fishing Activities Identification

in the presence of class imbalance and environmental noise. The models tested here show robustness in navigating these challenges. However, future implementations would benefit from user feedback to calibrate better the false rate.

D. Explainable AI (xAI) and Feature Importance Analysis

To enhance the interpretability of our classification models and gain a deeper understanding of their decision-making processes, we employed Shapley Additive Explanations (SHAP), a well-established technique in Explainable AI (xAI). The SHAP beeswarm plot provides valuable insights into the contribution of each feature to the model's predictions which are illustrated in Fig. 2. Each point represents a specific vessel instance and its position on the x-axis shows the amount of the feature influence to the prediction outcome (positive or negative) and its color indicates the original feature value (red for high, blue for low). The SHAP analysis revealed that wind speed, vessel speed, and directional changes were the most influential predictors. High wind speed and abrupt changes in direction are typically associated with fishing behavior, while consistent high speed may indicate transit rather than fishing. These insights highlight the importance of combining environmental and behavioral

data to enhance classification accuracy and model transparency.

V. DISCUSSION

In the era of AI, it is imperative for both consumers and the food industry to mitigate fraudulent activities within supply chains across various stages. In this context, we propose a FSMF System conceptual architecture, empowered by machine learning methods, specifically focused on the fishing stages within the Norwegian White Fish supply chain. The preliminary results, outlined in the previous section, are encouraging in their potential to prevent illegal fishing activities. However, the most significant impact can be achieved by expanding the current approach to encompass other stages of the specific supply chain, as well as products such as wine, honey, and extra virgin olive oil.

The architecture of the FSMF enables its adaptation to other food supply chains beyond the Norwegian whitefish case study. In the wine industry, the framework can integrate IoT sensor data related to fermentation, storage, and temperature conditions to detect anomalies in production processes.

In the honey supply chain, FSMF can be applied to detect adulteration using NIR spectroscopy or identify mislabeling based on botanical origin through pollen analysis and geolocation data.

Similarly, in the olive oil sector, potential fraud scenarios include inconsistencies between the quantity of harvested olives and the volume of oil produced, mislabeling detected through DNA profiling, or deviations in expected production yields based on field-level data such as tree density and cultivation area.

Across these domains, the core components of FSMF (data fusion, feature extraction, and anomaly detection) remain applicable, requiring only domain-specific calibration of data inputs and feature sets. This flexibility underlines the potential of FSMF to serve as a scalable early warning solution for diverse food fraud scenarios. These supply chains are particularly vulnerable, and the safety of consumers is jeopardized due to these illicit activities.

VI. FUTURE WORK

While the preliminary results from the fishing stage of the Norwegian White Fish case are promising, it is essential to integrate additional data into our experimentation phase and extend its application to other supply chains. We are currently in the process of acquiring data across each stage of the Wine, Honey, and Extra Virgin Olive Oil supply chains. The additional data are detailed described on the *section II*.

To implement EWS adaptation effectively within these supply chains, it is vital to continue exploring and testing the proposed methods. The implementation of the identification module that will demonstrate the end-to-end effectiveness will be included in the next version of the FSMF. This module will analyze detected anomalies, by the Data Analysis – Feature and Anomaly Extraction module, employing advanced risk assessment techniques to classify potential fraud risks based on historical data, probability models, and score-based evaluations. It will generate fraud alerts that connect anomalies to geographical and temporal patterns, enhancing traceability. Additionally, a Probabilistic Support mechanism will utilize these alerts to dynamically adjust risk scores and inform authorities about areas that are particularly

susceptible to fraud. The architecture will be designed for seamless integration with IoT-enabled monitoring systems, regulatory frameworks, and blockchain-based traceability solutions. By incorporating real-time data processing and adaptive learning, the final architecture will improve fraud detection and facilitate proactive decision-making in food supply chain management.

ACKNOWLEDGMENT

This work was funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency (REA). Neither the European Union nor the granting authority can be held responsible for them. This work was conducted as part of the Watson project, funded by the European Union's Horizon Europe research and innovation program, under grant agreement No. 101084265.

REFERENCES

- [1] L. Kyrgiakos, Malak Hazimeh, Marios Vasileiou, C. Kleisiari, G. Klefodimos, and G. Vrontzos, 'The Food Fraud Landscape: A Brief Review of Food Safety and Authenticity', *The 17th International Conference of the Hellenic Association of Agricultural Economists*, 2024, doi: 10.3390/proceedings2024094006.
- [2] Konstantinos Giannakas and Amalia Yiannaka, 'Food Fraud: Causes, Consequences, and Deterrence Strategies', *Annual Review of Resource Economics*, vol. 15, no. 1, pp. 85–104, Oct. 2023, doi: 10.1146/annurev-resource-101422-013027.
- [3] Y. Yakar and K. Karadağ, 'Identifying olive oil fraud and adulteration using machine learning algorithms', *Quim. Nova*, vol. 45, pp. 1245–1250, Jan. 2023, doi: <https://doi.org/10.21577/0100-4042.20170948>.
- [4] 'Spatial-Temporal Event Analysis as a Prospective Approach for Signalling Emerging Food Fraud-Related Anomalies in Supply Chains - PMC'. Accessed: Feb. 21, 2025. [Online]. Available: <https://pmc.ncbi.nlm.nih.gov/articles/PMC9818448/>
- [5] W. Angie Abia, 'Food Fraud Detection: The Role of Spectroscopy Coupled with Chemometrics', *J Nutr Diet Manage*, vol. 1, no. 1, pp. 1–7, Sep. 2023, doi: 10.59462/JNDM.1.1.103.
- [6] 'Anomaly Score-Based Risk Early Warning System for Rapidly Controlling Food Safety Risk - PMC'. Accessed: Feb. 21, 2025. [Online]. Available: <https://pmc.ncbi.nlm.nih.gov/articles/PMC9316538/>
- [7] N. Naem, 'Artificial Intelligence based techniques for authenticity of food products in Food fraud', *International Journal for Electronic Crime Investigation*, vol. 8, no. 3, Art. no. 3, Sep. 2024, doi: 10.54692/ijeci.2024.0803199.
- [8] L. Manning *et al.*, 'Food fraud prevention strategies: Building an effective verification ecosystem', *Comprehensive Reviews in Food Science and Food Safety*, vol. 23, no. 6, p. e70036, 2024, doi: 10.1111/1541-4337.70036.
- [9] J. M. Soon and I. R. Abdul Wahab, 'A Bayesian Approach to Predict Food Fraud Type and Point of Adulteration', *Foods*, vol. 11, no. 3, Art. no. 3, Jan. 2022, doi: 10.3390/foods11030328.
- [10] Suhaili Othman, Nidhi Rajesh Mavani, Mohd Azlan Hussain, Norliza Abd Rahman, and Jarinah Mohd Ali, 'Artificial intelligence-based techniques for adulteration and defect detections in food and agricultural industry: A review', *Journal of Agriculture and Food Research*, vol. 12, pp. 100590–100590, Jun. 2023, doi: 10.1016/j.jafr.2023.100590.
- [11] R. Sharma *et al.*, 'Rapid and sensitive approaches for detecting food fraud: A review on prospects and challenges', *Food Chemistry*, vol. 454, p. 139817, Oct. 2024, doi: 10.1016/j.foodchem.2024.139817.
- [12] N. Liu, Y. Bouzembrak, L. M. van den Bulk, A. Gavai, L. J. van den Heuvel, and H. J. P. Marvin, 'Automated food safety early warning system in the dairy supply chain using machine learning', *Food Control*, vol. 136, p. 108872, Jun. 2022, doi: 10.1016/j.foodcont.2022.108872.
- [13] K. G. Dastidar, O. Caelen, and M. Granitzer, 'Machine Learning Methods for Credit Card Fraud Detection: A Survey', *IEEE Access*, vol. 12, pp. 158939–158965, 2024, doi: 10.1109/ACCESS.2024.3487298.

- [14] E. Zuo *et al.*, 'Anomaly Score-Based Risk Early Warning System for Rapidly Controlling Food Safety Risk', *Foods*, vol. 11, no. 14, Art. no. 14, Jan. 2022, doi: 10.3390/foods11142076.
- [15] W. Mu *et al.*, 'Making food systems more resilient to food safety risks by including artificial intelligence, big data, and internet of things into food safety early warning and emerging risk identification tools', *Comprehensive Reviews in Food Science and Food Safety*, vol. 23, no. 1, p. e13296, 2024, doi: 10.1111/1541-4337.13296.
- [16] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, 'Multimodal Machine Learning: A Survey and Taxonomy', Aug. 01, 2017, *arXiv: arXiv:1705.09406*. doi: 10.48550/arXiv.1705.09406.
- [17] V. Tsoukas, A. Gkogkidis, A. Kampa, G. Spathoulas, and A. Kakarountas, 'Enhancing Food Supply Chain Security through the Use of Blockchain and TinyML', *Information*, vol. 13, no. 5, Art. no. 5, May 2022, doi: 10.3390/info13050213.
- [18] P. W. Khan, Y.-C. Byun, and N. Park, 'IoT-Blockchain Enabled Optimized Provenance System for Food Industry 4.0 Using Advanced Deep Learning', *Sensors*, vol. 20, no. 10, p. 2990, May 2020, doi: 10.3390/s20102990.
- [19] S. Othman, N. R. Mavani, M. A. Hussain, N. A. Rahman, and J. Mohd Ali, 'Artificial intelligence-based techniques for adulteration and defect detections in food and agricultural industry: A review', *Journal of Agriculture and Food Research*, vol. 12, p. 100590, Jun. 2023, doi: 10.1016/j.jafr.2023.100590.
- [20] C. Jin *et al.*, 'Big Data in food safety- A review', *Current Opinion in Food Science*, vol. 36, pp. 24–32, Dec. 2020, doi: 10.1016/j.cofs.2020.11.006.
- [21] A. Vinothkanna, O. I. Dar, Z. Liu, and A.-Q. Jia, 'Advanced detection tools in food fraud: A systematic review for holistic and rational detection method based on research and patents', *Food Chemistry*, vol. 446, p. 138893, Jul. 2024, doi: 10.1016/j.foodchem.2024.138893.
- [22] Y. Bouzembrak *et al.*, 'Data driven food fraud vulnerability assessment using Bayesian Network: Spices supply chain', *Food Control*, vol. 164, p. 110616, Oct. 2024, doi: 10.1016/j.foodcont.2024.110616.
- [23] J. Spink, D. L. Ortega, C. Chen, and F. Wu, 'Food fraud prevention shifts the food risk focus to vulnerability', *Trends in Food Science & Technology*, vol. 62, pp. 215–220, Apr. 2017, doi: 10.1016/j.tifs.2017.02.012.
- [24] D. Lahat, T. Adali, and C. Jutten, 'Multimodal Data Fusion: An Overview of Methods, Challenges, and Prospects', *Proc. IEEE*, vol. 103, no. 9, pp. 1449–1477, Sep. 2015, doi: 10.1109/JPROC.2015.2460697.
- [25] M. Alwateer, E.-S. Atlam, M. M. A. El-Raouf, O. A. Ghoneim, and I. Gad, 'Missing Data Imputation: A Comprehensive Review', *Journal of Computer and Communications*, vol. 12, no. 11, Art. no. 11, Oct. 2024, doi: 10.4236/jcc.2024.1211004.
- [26] P.-O. Côté, A. Nikanjam, N. Ahmed, D. Humeniuk, and F. Khomh, 'Data Cleaning and Machine Learning: A Systematic Literature Review', May 31, 2024, *arXiv: arXiv:2310.01765*. doi: 10.48550/arXiv.2310.01765.
- [27] Nuraddeen Usman, Ema Utami, and Anggit Dwi Hartanto, 'Comparative Analysis of Elliptic Envelope, Isolation Forest, One-Class SVM, and Local Outlier Factor in Detecting Earthquakes with Status Anomaly using Outlier', *2023 International Conference on Computer Science, Information Technology and Engineering (ICCoSITE)*, Feb. 2023, doi: 10.1109/iccosite57641.2023.10127748.
- [28] Tianming Xie, Qifa Xu, Cuixia Jiang, Zhiwei Gao, and Xiangxiang Wang, 'A Robust Anomaly Detection Model for Pumps Based on the Spectral Residual With Self-Attention Variational Autoencoder', *IEEE Transactions on Industrial Informatics*, 2024, doi: 10.1109/tii.2024.3381790.
- [29] H. Liu *et al.*, 'Using t-distributed Stochastic Neighbor Embedding (t-SNE) for cluster analysis and spatial zone delineation of groundwater geochemistry data', *Journal of Hydrology*, vol. 597, p. 126146, Jun. 2021, doi: 10.1016/j.jhydrol.2021.126146.
- [30] A. N. Himaya and M. Sano, 'Course-Keeping Performance of a Container Ship with Various Draft and Trim Conditions under Wind Disturbance', *Journal of Marine Science and Engineering*, vol. 11, no. 5, Art. no. 5, May 2023, doi: 10.3390/jmse11051052.
- [31] H. Duan, F. Ma, L. Miao, and C. Zhang, 'A semi-supervised deep learning approach for vessel trajectory classification based on AIS data', *Ocean & Coastal Management*, vol. 218, p. 106015, Mar. 2022, doi: 10.1016/j.ocecoaman.2021.106015.
- [32] D. Luo, P. Chen, J. Yang, X. Li, and Y. Zhao, 'A New Classification Method for Ship Trajectories Based on AIS Data', *Journal of Marine Science and Engineering*, vol. 11, no. 9, Art. no. 9, Sep. 2023, doi: 10.3390/jmse11091646.
- [33] P. Kraus, C. Mohrdieck, and F. Schwenker, 'Ship classification based on trajectory data with machine-learning methods', in *2018 19th International Radar Symposium (IRS)*, Jun. 2018, pp. 1–10. doi: 10.23919/IRS.2018.8448028.
- [34] S. Baeg and T. Hammond, 'Ship Type Classification Based on The Ship Navigating Trajectory and Machine Learning'.