Journal of Food Safety and Food Quality
Archiv für Lebensmittelhygiene

J. Food Saf. Food Qual. **2025**; 76(2): 39242
https://doi.org/10.31083/JFSFQ39242

Original Research

Explainable AI and Voting Ensemble Model to Predict the Results of Seafood Product Importation Inspections

Saksonita Khoeurn[1], Kyunghee Lee[2], Wan-Sup Cho[1,2,*]

[1]BigDataLabs Co., Ltd, 28644 Cheongju, Chungcheongbuk-do, Republic of Korea
[2]Department of Management Information Systems, Chungbuk National University, 28644 Cheongju, Chungcheongbuk-do, Republic of Korea
*Correspondence: wscho@cbnu.ac.kr (Wan-Sup Cho)
Academic Editor: Corinna Kehrenberg
Submitted: 24 March 2025 Revised: 18 April 2025 Accepted: 27 April 2025 Published: 17 June 2025

Abstract

Background: As the volume of imported food flowing into South Korea rapidly increases due to the expansion of free trade agreements, improving inspection efficiency through artificial intelligence technology emerges as a critical task, particularly as time and cost expenditures for safety inspections conducted by the Korean Ministry of Food and Drug Safety concurrently increase rapidly. The lack of a generalizable machine learning model for predicting the safety of food for human consumption constitutes a significant challenge for policymakers and responsible authorities. **Methods**: This study developed an effective classification model for predicting non-conformance in customs inspection of imported seafood products. To address the severe class imbalance inherent in the inspection data, we applied class weight-based cost-sensitive learning and adopted an ensemble approach combining Decision Trees (DT), Random Forests (RF), Logistic Regression (LR), and Naive Bayes (NB) models. **Results**: Performance evaluation demonstrated that the soft voting ensemble technique achieved superior predictive performance in identifying non-conformance cases, with a recall of 75.57% and an Area Under the Curve (AUC) of 87.49%, significantly outperforming the hard voting method's recall of 44.32% and AUC of 72.07%. Through SHapley Additive exPlanations (SHAP) analysis, we confirmed that various characteristics, including exporting country ratio, major product category, overseas manufacturer ratio, importer ratio, and seasonal variation, exerted substantial influence on the models' decisions. **Conclusion**: Notably, the Naive Bayes model component provided a more comprehensive analysis for identifying non-conformance by considering multiple dimensions and potential seasonality. This research guide for predicting seafood product import inspection results contributes to enhancing inspection efficiency for securing the safety of imported aquatic products.The proposed methodology demonstrates potential applicability to other regulatory inspection domains confronting similar data imbalance challenges.

Keywords: border inspection; decision trees; ensemble learning; explainable artificial intelligence; food safety management

1. Introduction

In recent decades, the global seafood trade has expanded dramatically, with international seafood imports reaching unprecedented volumes to meet growing consumer demand worldwide [1]. This expansion has been accompanied by increasing concerns regarding food safety, as seafood products are particularly susceptible to various contaminants, including heavy metals, pathogens, and unauthorized additives [2]. Regulatory authorities around the world have implemented rigorous inspection systems at entry points to ensure that imported seafood meets established safety standards. However, the effectiveness of these inspection systems faces significant challenges due to the vast volume of imports and the relatively low incidence of non-conformity, approximately 0.2%, as observed in our dataset.

Food safety is deeply affected by the diversity of foods and their raw materials. The increasing trade openness of the global economy has resulted in increased food imports, emphasizing the significance of food risk management in protecting consumer health. Predictions and early warning are crucial to ensure food safety; in particular, food inspection prior to entry into the consumer market is a significant step in ensuring good food quality.

In this context, the term non-conformity is used to denote inspection failures resulting from violations of established safety and quality criteria. These include microbiological hazards, chemical residues, and labeling defects, which are evaluated according to the regulatory standards enforced by the Korean Ministry of Food and Drug Safety [3].

The identification of key factors influencing non-conformity in imported seafood represents a critical area for research, as it could potentially enhance the efficiency and effectiveness of inspection protocols. Previous studies have examined various aspects of seafood safety, including the prevalence of specific contaminants [4,5], geographical variations in compliance rates [6], and seasonal fluctuations in detection rates [7]. However, there has been little research on the use of proactive inspections for high-risk food predictions as part of the border control for imported foods, with the exceptions of the United States (US) and the European Union (EU). Furthermore, there remains a significant gap in understanding how multiple factors—such as

seafood type, distribution method, import timing, country of origin, and characteristics of importers and exporters—collectively influence the likelihood of non-conformity. Furthermore, governmental implementations such as the risk prediction-based imported food inspection system developed by the Korean Ministry of Food and Drug Safety [8] have established procedural precedents through logistic modeling approaches incorporating multidimensional risk factors, though such systems exhibit methodological constraints when addressing extreme classification imbalance and provide limited explainability mechanisms for specialized product categories.

The inherent challenge in developing predictive models for seafood inspection outcomes lies in the severe class imbalance within the available data. With non-conformity rates of approximately 0.2% (796 non-conforming cases from 389,389 total observations in our study), traditional classification algorithms often fail to accurately identify the minority class (non-conforming samples), instead favoring the overwhelming majority class (conforming samples) [9]. This imbalance problem is particularly problematic in regulatory contexts where the cost of missing a non-conforminity product (false negative) significantly outweighs the cost of additional inspection for a compliant product (false positive).

Despite several studies focused on using machine learning for predicting the results of inspections, there is a lack of information regarding the model decision and explainability, such as that found in the areas of water quality [10,11] and healthcare [12,13]. To address this gap, our study employs SHapley Additive exPlanations (SHAP) [14] to provide principled and interpretable insights into the model's decision-making process.

This study addresses these challenges by developing and evaluating multiple classification models designed to predict non-conformity in imported seafood products while effectively handling the class imbalance problem. Specifically, we implement and compare four distinct classification algorithms—Decision Tree (DT), Random Forest (RF), Logistic Regression (LR), and Naive Bayes (NB)—and employ both hard and soft voting ensemble mechanisms to determine the most effective approach. Our methodology incorporates specialized data preprocessing techniques to mitigate the effects of class imbalance, including strategic sampling methods and feature engineering tailored to the unique characteristics of seafood import data.

The primary objectives of this research are: (1) to identify the key factors that significantly influence non-conformity in imported seafood products; (2) to develop an effective classification model that accurately identifies potentially non-conformant products despite the severe class imbalance; and (3) to provide actionable insights for regulatory authorities to optimize inspection protocols and resource allocation. By achieving these objectives, this study aims to contribute to the enhancement of food safety systems while simultaneously reducing unnecessary inspection burdens on compliant imports.

The results demonstrated that a soft voting ensemble approach achieved superior performance in identifying non-conformity cases, with a recall of 75.57% and Area Under the Curve (AUC) of 87.49%, outperforming the hard voting method. SHAP analysis confirmed that exporting country ratio, major product category, manufacturer ratio, importer ratio, and seasonal variations significantly influenced model decisions providing comprehensive analysis by considering multiple risk dimensions and seasonality. This research not only enhances seafood inspection efficiency but also offers a methodological framework that can be readily applied to other product categories within the food sector facing similar class imbalance challenges.

The remainder of this paper is organized as follows. Section 2 provides a review of related literature on data sampling methods, ensemble learning, and explainable artificial intelligence. Section 3 describes the materials and methods, including data sources, preprocessing, model construction, and evaluation. Section 4 presents the experimental results and interpretability analysis. Section 5 discusses the implications of the findings. Finally, Section 6 concludes the study and outlines directions for future research.

2. Literature Review

This section presents a comprehensive overview of the theoretical foundations and prior research relevant to the prediction of food inspection outcomes. It introduces essential concepts including data sampling strategies for addressing class imbalance, ensemble learning techniques with an emphasis on voting mechanisms, and explainable artificial intelligence (XAI).

The objective is to establish a solid conceptual and methodological framework for the approaches employed in this study. Furthermore, this section identifies existing research gaps, particularly the limited application of interpretable machine learning models in the domain of food safety inspections, thereby highlighting the novelty and necessity of this study.

2.1 Data Sampling Method

Data sampling methods are necessary in machine learning and data analysis as they allow imbalanced datasets to be addressed, wherein one class has significantly fewer instances than the others. In this study, the non-conformity class contains significantly fewer samples than the conformity class. Consequently, this imbalance of datasets can lead to biased model performance and poor generalization, particularly in the context of classification tasks. Data sampling methods, therefore, aim to balance the class distribution by oversampling the minority class, undersampling the majority class, or generating synthetic samples [15].

Synthetic minority oversampling (SMOTE) is a popular and effective data sampling method [16]. SMOTE

works by selecting samples in a feature space that are close together, drawing a line between the samples and drawing a new sample at a point along that line. More specifically, a random example from the minority class is selected first. For this example, the k of its nearest neighbors is determined (typically, k = 5). A randomly chosen neighbor is selected, and a synthetic example is created in the feature space at a randomly chosen point between the two examples.

SMOTE techniques are also known to be selective. For example, numerous SMOTE extensions exist for oversampling methods. One popular method is the borderline-SMOTE, which involves selecting misclassified instances of the minority class, such as the k-nearest neighbor (KNN) classification model [17]. Instead of randomly generating new synthetic examples for the minority class, the borderline-SMOTE method generates synthetic examples only along the decision boundary between the two classes. In addition to the KNN model, another approach known as borderline-SMOTE Support Vector Machine (SVM) or SMOTE-SVM was introduced using the SVM algorithm to identify misclassifications on the decision boundary [18].

2.2 Voting Ensemble Model

A voting ensemble is a machine learning ensemble model that combines predictions from multiple models. This technique can be used to improve the model performance, ideally outperforming any single model in the ensemble. A voting ensemble combines the predictions from multiple models. This method is suitable for classification; during classification, the predictions for each label are added, and the label with the most votes is predicted [19].

Two approaches are available to predict the majority votes for classification, namely hard voting and soft voting. As shown in Fig. 1, hard voting entails adding up all the predictions for each class label and predicting the class label with the most votes. Meanwhile, soft voting averages the predicted probabilities for each class label and predicts the class label with the greatest probability [20].

2.3 Explainable Artificial Intelligence

Artificial intelligence (AI) methods have achieved unprecedented levels of performance in solving complex computational tasks, making them vital for the future development of human society. In recent years, the sophistication of AI-powered systems has increased, rendering them almost devoid of human intervention in terms of their design and deployment. However, as black-box machine learning (ML) models become increasingly used in practice, the demand for transparency has increased, and the explanations supporting the output of the model become crucial. As humans are hesitant to adopt techniques that are not directly interpretable, tractable, or trustworthy, there is a requirement for ethical AI. In addition, although it is customary to consider that focusing solely on performance leads to unclear systems, improving the understanding of a system can lead to the correction of its deficiencies. For example, enhanced interpretability can improve ML models by ensuring impartiality during decision making, thereby providing robustness by highlighting potential adversarial perturbations, and ensuring that only meaningful variables infer the output. To avoid limiting the effectiveness of current AI systems, eXplainable AI (XAI) proposes creating a suite of ML techniques that produce more explainable models while maintaining a high learning performance. XAI draws insights from the social sciences and from the psychology of explanation to encourage humans to understand, trust, and effectively manage emerging generations of AI partners [21]. In this study, SHAP was applied to interpret the prediction of seafood non-conformity enhancing model transparency and regulatory usability.

Among the various XAI techniques reported to date, SHAP is a powerful and widely used technique, which provides a principled and model-agnostic approach to explain the predictions of machine-learning models. It is based on the cooperative game theory and the concept of Shapley values, which originated in the field of economics. SHAP assigns a fair and consistent contribution score to each feature in a prediction, quantifying its impact on the model's output. The core idea behind SHAP is the consideration of all possible feature combinations and computation of the differences in predictions when a specific feature is included or excluded, thereby capturing its individual effects. By averaging these differences over all possible combinations, the SHAP values provide a global explanation for the entire dataset. Additionally, SHAP values can be applied at the individual level, offering local explanations for each prediction, thereby rendering them highly valuable for understanding model behavior on a case-by-case basis.

3. Materials and Methods

This section presents the methodological framework employed to develop and assess the machine learning models for predicting seafood product import inspection outcomes as illustrated in Fig. 2. The study process is described in detail, including data acquisition, preprocessing procedures, feature engineering, and the construction of classification models. Particular attention is given to handling class imbalance using resampling techniques, selecting statistically significant features, and implementing ensemble learning approaches through hard and soft voting mechanisms. Furthermore, model evaluation metrics and explainable AI techniques are introduced to ensure both the reliability and interpretability of the predictive results.

3.1 Data Sources

This study primary used imported food declaration data provided by Korean Ministry of Food and Drug Safety [3]. The dataset consists of a comprehensive range of product information, including dates, import/export companies,

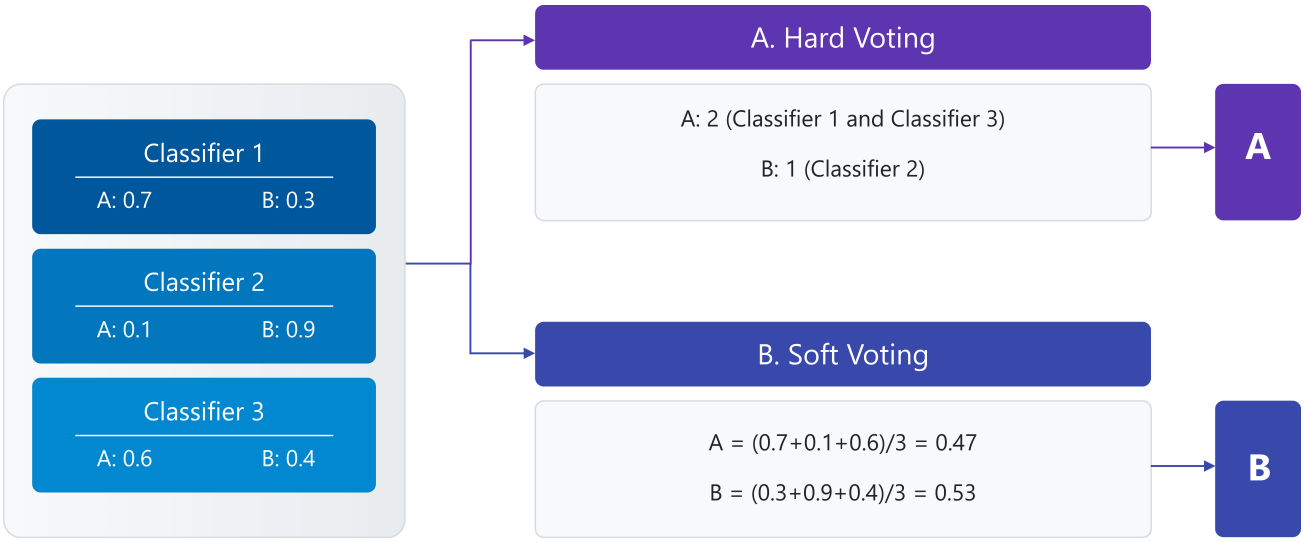

Fig. 1. Voting ensemble techniques explanation. Comparison of hard and soft voting ensemble techniques where hard voting uses majority rule while soft voting averages prediction probabilities across multiple classifiers.

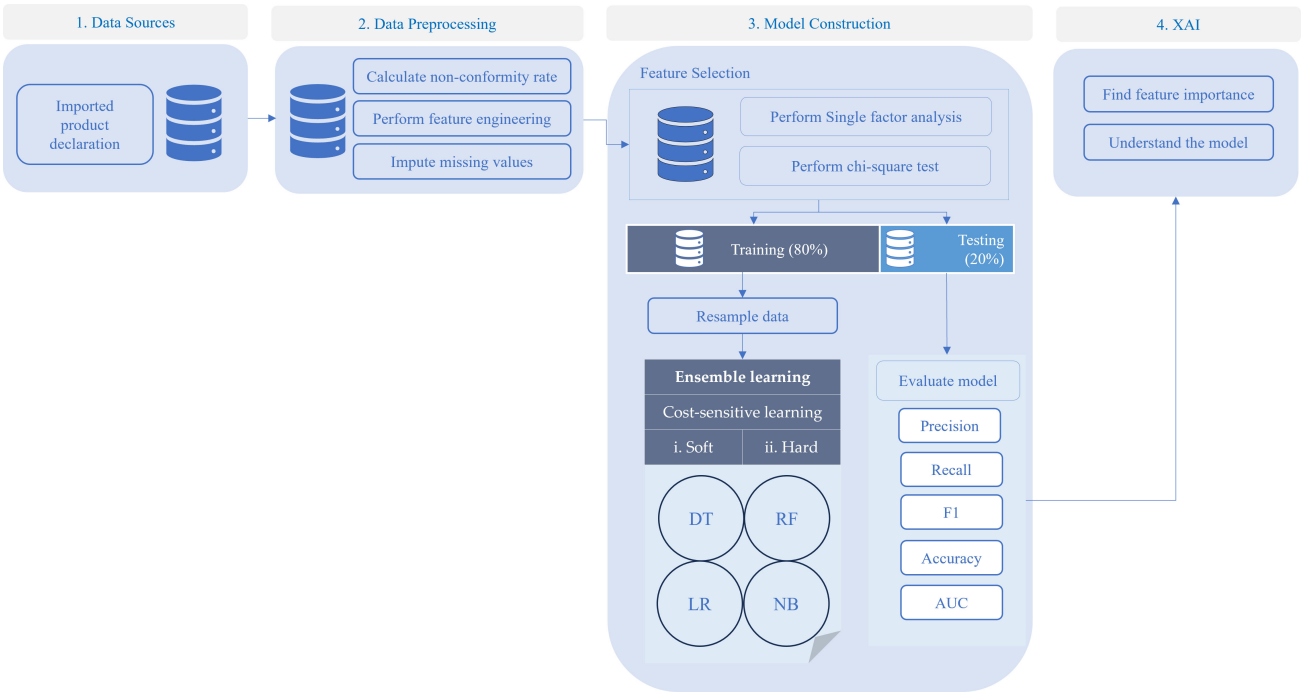

Fig. 2. Workflow of the four-phase methodology. The study process includes four phases. Machine learning models used in the study are Decision Tree (DT), Random Forest (RF), Logistic Regression (LR), and Naive Bayes (NB). AUC, Area Under the Curve; XAI, explainable artificial intelligence.

detailed product specifications, import weights, distribution methods, processing results, and inspection types. The received datasets ranged from 2018 to 2021. The total dataset size is 389,389, with 388,593 and 796 samples in the conformity and non-conformity classes, respectively.

3.2 Data Preprocessing

To ensure sufficient data quality and structure for training, the data preprocessing phase was divided into three essential parts such as non-conformity rate calculations, feature engineering, and missing value imputation. Table 1 presents the data attributes and derived metrics resulting from the non-conformity rate calculations and the feature engineering process. The report receipt dates were converted into months, weeks, and seasons. Spring ranges from March to May, Summer ranges from June to August, and Winter ranges from December to February.

Table 1. Description of variables.

Original variable	Derived variable	Description
Receipt date	Month	Represents the year and month when the report was received.
	Week	Represents the year and week when the report was received.
	Season	Spring, Summer, Fall or Winter.
Import shipper	Import Shipper	The importer.
	Import Shipper Failed Ratio	The previous failed ratio of the relevant importer.
Exporting country	Exporting Country	The country from which the product is being exported.
	Continent	The continent of the exporting country.
	Exporting Country Failed Ratio	The previous failed ratio of the relevant exporting country.
	Continent Failed Ratio	The previous failed ratio of the relevant continent.
Overseas manufacturer	Overseas Manufacturer	The foreign company responsible for producing the goods.
	Overseas Manufacturer Failed Ratio	The previous failed ratio of the relevant overseas manufacturer.
Exporter	Exporter	The company or party that is responsible for exporting the goods from the originating country.
	Exporter Ratio	The previous failed ratio of the relevant exporter.
Major product category	Major Product Category	The major product category of the goods.
	Major Product Category Failed Ratio	The previous failed ratio of the relevant major product category.
Sub product category	Sub Product Category	The sub-product category of the goods.
	Sub Product Category Ratio	The previous failed ratio of the relevant sub-product category.
Product name	Product Name	The specific name or description of the products being imported/exported.
	Product Name Failed Ratio	The previous failed ratio of the relevant product name.
Total net weight	Total Net Weight	The total net weight of the products being imported/exported.
Distribution method	Distribution Method	The distribution method.
Type of inspection	Type of Inspection	The type of inspection conducted on the products.
Processing result	Processing Result	The outcome or result of the inspection of the shipment.

The hit rate of each attribute, including non-conformity, was determined based on a range of variables, including the import shipper, exporting country, continent, overseas manufacturer, exporter, major product category, sub-product category, and product name. These attributes were individually utilized to calculate their respective hit rates, providing valuable insights into the occurrence of non-conformities and irregularities associated with each specific attribute. The non-conformity rate (Π) can be calculated using the following formula:

$$\Pi(\text{Non-conformity rate of variable}) = \frac{N(\text{Non-conforming instances of variable})}{N(\text{Total instances of variable})} \quad (1)$$

where N(variable) is the total number of instances in which the variable is the same as the specific value of interest (the value for which the non-conformity rate must be calculated), while N(Non-conformities of variable) is the number of instances in which the variable is the same as the specific value of interest, and represents the non-conformities. After completing the calculation and feature engineering, the missing values were input as zero.

3.3 Model Construction

3.3.1 Feature Selection

The preprocessed data were subjected to a two-stage variable selection process, in which attributes were designated for inclusion in the model construction. In the first stage, a single-factor analysis was performed to identify factors that had statistically significant relationships with conformity or non-conformity during inspections. Different statistical tests were adopted depending on the variable type. The continuous variables, such as the total net weight, were analyzed using the ANOVA test. In contrast, the remainder of the variables, namely the categorical variables, were analyzed using the chi-squared test [22].

3.3.2 Spliting of the Data into Training and Testing Datasets

To acquire the optimal models, perform model validation, and evaluate the model performance, the dataset was split into two groups—80% for training and 20% for testing. After feature selection was constructed, dataset used for modeling were divided into the test and training datasets and were subsequently oversampled to balance the training data as shown in Fig. 3.

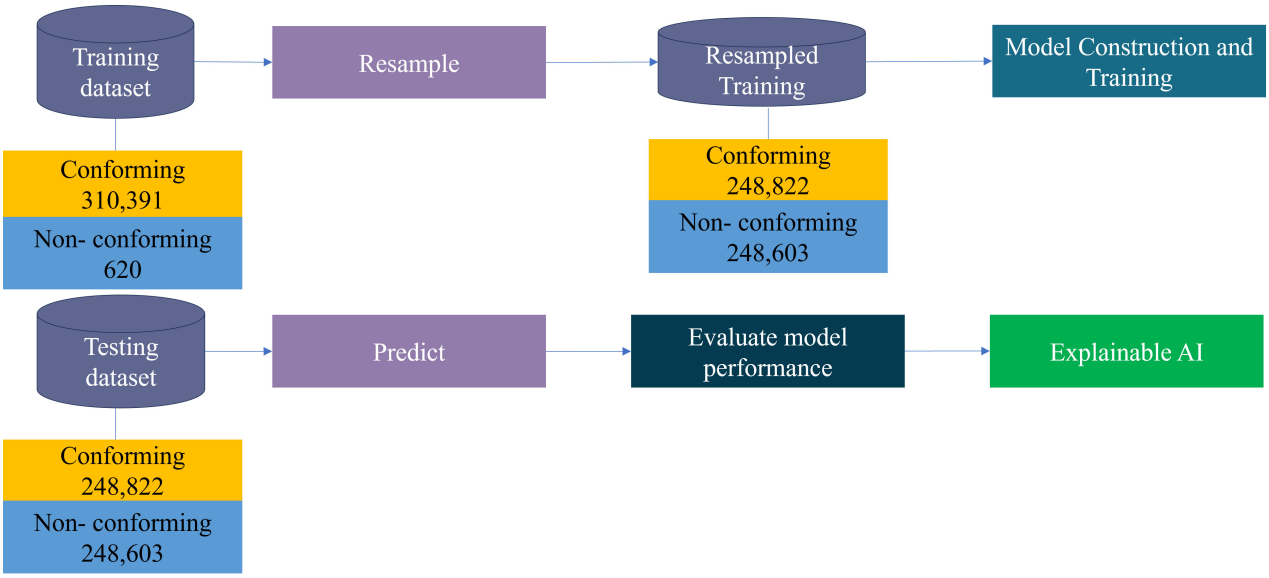

Fig. 3. Flowchart for the prediction model. Data preparation flow for balancing the highly imbalanced seafood inspection dataset through resampling before model training and evaluation.

The main purpose of resampling was to enhance the discriminatory ability of the model rather than to learn erroneous samples. Moreover, the test dataset deviated from the original data if sampling was performed before splitting the data. Consequently, the model learned noise from the data, resulting in inaccurate predictions. The dataset was oversampled using SVM-SMOTE-SVM, and the resampled training dataset was employed during model training to establish the most suitable model for the testing dataset and XAI. The resampled data for training consisted of 497,425 data points for the conform class and 248,603 for the non-conform class.

3.3.3 Modelling

Four types of models, namely, Decision Tree (DT) [23], Random Forest (RF) [24], Logistic Regression [25], and Naive Bayes (NB) [26], were used to create ensemble models for model training. Because non-conformity is considered a minority class, this study used cost-sensitive learning that considers the costs of different misclassifications [27]. Using the balanced method, the class weights were inversely proportional to the class frequencies in the training dataset. Using class weights, the model learned to minimize the total cost, not just the number of misclassifications. This can be beneficial in situations where misclassification costs are unevenly distributed. As mentioned above, two types of ensemble models were used in this study, namely soft and hard voting models. The performances of the soft and hard voting models were compared using both the resampled validation and the test datasets. XAI was then used to check the feature importance of each model (i.e., DT, RF, LR, and NB).

The seafood inspection classification model presented in Fig. 4 constitutes a taxonomic framework designed to predict binary inspection outcomes through the integration of 27 predictor variables strategically organized into seven distinct categories with the detailed predictors and response variable. These categories—comprising Entity Identification Variables (n = 4), Product Characterization Variables (n = 3), Quantitative Assessment Variables (n = 2), Logistical-Procedural Variables (n = 2), Temporal-Chronological Variables (n = 6), Geographical Variables (n = 1), and Ratio Metrics (n = 9)—establish a comprehensive analytical foundation for the voting classifier mechanism. The model employs SHAP analysis to derive feature importance rankings through both global assessments and local explanations, thereby facilitating interpretability of the classification outputs. A representative sample from our dataset is presented in **Supplementary Table 1**, illustrating the 27 predictor variables used in the classification model.

3.3.4 Model Evaluation

The model performance were measured and validated using a confusion matrix and model predictive performance indicators to select the optimal model and evaluate its performance. A confusion matrix was structured using the entries listed in Table 2, and the necessary predictive performance indicators were calculated using the numbers of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN).

Predictive performance indicators included the area under the curve (AUC), the positive predictive value (PPV) (also known as precision), the F1 score, the recall, and the accuracy (ACR), which are defined in detail below.

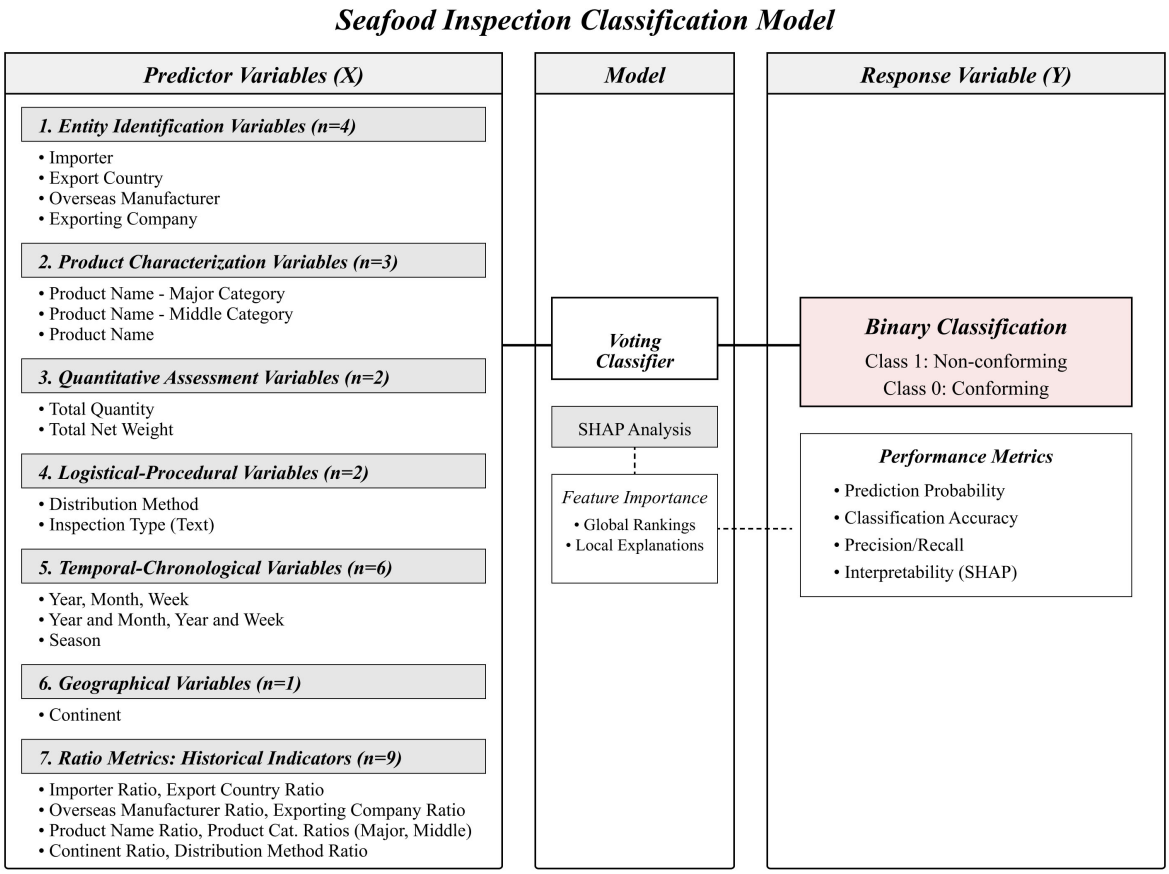

Fig. 4. Seafood inspection classification model. Illustrating the taxonomic organization of predictor variables and their relationship to the binary classification outcome. SHAP, SHapley Additive exPlanations.

Table 2. Definitions of entry types in the confusion matrix.

Entry type	Definition
True Positive (TP)	Predicted inspection result for the product by model classification: non-conformity; actual inspection result: non-conformity.
False Positive (FP)	Predicted inspection result for the product batch by model classification: non-conformity; actual inspection result: conformity.
True Negative (TN)	Predicted inspection result for the product batch by model classification: conformity; actual inspection result: conformity.
False Negative (FN)	Predicted inspection result for the product by model classification: conformity; actual inspection result: non-conformity.

Table 3. Performances of the ensemble models.

Voting method	ACR	Recall	PPV	F1	AUC	TN	FP	TP	FN
Soft voting	99.35%	**75.57**%	22.32%	34.46%	**87.49**%	77,239	463	133	43
Hard voting	**99.69**%	44.32%	**35.62**%	**39.49**%	72.07%	77,561	141	78	98

Note: Bold indicates the higher value for metrics where higher is preferable. ACR, accuracy rate; PPV, positive predictive value.

- The accuracy rate (ACR) evaluates the model's overall capacity to differentiate between conformity and non-conformity samples or the ability to accurately classify samples as conformity. However, owing to the lower proportion of non-conformities in our data, there was an imbalance in the samples considered herein. Because of its higher capacity for discriminating conformities, ACR may show bias in predicting conformities. To over-

come this problem, the recall and PPV indicators have received greater attention during the evaluation of model performance. The ACR can be calculated using (2):

$$ACR = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

- The recall or sensitivity is the proportion of samples correctly labeled as non-conformity out of all non-conformity samples, as shown in (3):

$$Recall = \frac{TP}{FN + TP} \quad (3)$$

- The positive predictive value (PPV), also known as precision, is the proportion of samples that the model classifies as non-conformity out of all samples, and is otherwise referred to as the non-conformity rate. The PPV can be calculated using (4):

$$PPV = \frac{TP}{TP + FP} \quad (4)$$

- The F1 score, defined as the harmonic mean of the recall and PPV indicators, becomes crucial when dealing with imbalanced data. Higher TP values correlate with higher F1 scores, and the F1 score can be calculated using (5):

$$F1 = \frac{2 \cdot PPV \cdot Recall}{PPV + Recall} \quad (5)$$

- The model's classification accuracy can be measured from the area under the receiver operating characteristic (ROC) curve (AUC), wherein a larger AUC denotes a higher accuracy. More specifically, AUC = 1 represents a great classifier, 0.5 < AUC < 1 represents a model that outperforms random guessing, AUC = 0.5 represents a model that is similar to random guessing but lacks classification capacity, and AUC <0.5 represents a classifier that performs worse than random guessing.

According to the explanation above, recall and AUC scores play an important role in the model evaluation. The higher scores show that the higher chance model can correctly identify the non-conformity class.

3.4 Explainable AI

To understand the decisions made by the models, the Shapley approach was employed to determine features having a larger effect on the model's prediction of conformity or non-conformity. Shapley is a widely used interpretability technique that assigns importance values to each feature based on its impact on the model's predictions. By analyzing these important values, this study aims to gain insights into the underlying factors driving the model's conformity or non-conformity predictions. In addition, this approach allows researchers to identify potential biases or inconsistencies in the decision-making process of a model.

This study explored all models incorporated into the ensemble models to determine the common importance features. By comparing the important features across all ensemble models, this study aimed to identify features that consistently had a significant impact on the model's predictions. This analysis provided a more robust understanding of the key factors driving the decision-making process of the model, whilst also helping validate the reliability of the ensemble models.

4. Results

4.1 Comparisons of the Ensemble Model Performance

Four different models, namely NB, DT, RF, and LR models, combined with class-weight cost-sensitive learning, were used to create both hard-voting and soft-voting ensemble models. These models were applied to forecast the inspection outcome after training and were used to predict the test data. The testing dataset contained 778,878 points of data consisting of 176 non-conformity classes and 77,702 non-conformity classes. Table 3 lists the performances of the various ensemble models.

With 75.57% of the votes, the soft voting method outperformed the hard voting method (44.32% of the votes) in terms of the recall score. In addition, soft voting received a higher AUC score of 87.49%, whereas hard voting only received a score of 72.07%. While soft voting received 99.35% in terms of its ACR score, hard voting received a slightly higher 99.69%. A similar result was observed for the PPV score, with hard voting receiving 35.62% of the vote and soft voting receiving 22.32%. Furthermore, hard voting received a higher F1 score of 39.49% than 34.46% for soft voting. These results indicate that overall, the soft voting method exhibited a superior performance in terms of the recall and AUC scores, while the hard voting method outperformed the soft voting method in terms of the ACR, PPV, and F1 scores.

The goal of this study was to determine the best combination of recall and AUC scores to accurately detect non-conformity data, and based on this goal, soft voting produced superior results in predicting the inspection results. It therefore appears that soft voting is a suitable approach for accurately detecting non-conformity data and predicting inspection results with high recall and AUC scores.

4.2 Identification of the Importance of Features Using XAI

4.2.1 Global Importance Feature

The SHAP values of each model used in the ensemble model were calculated to assess the significance of the features in model decision-making. Subsequently, the frequently prioritized attributes that influenced the model classification were identified. However, SHAP values are de-

signed to work with additive models, such as linear models and tree-based models (e.g., decision trees, random forests, and gradient boosting machines). As the importance of the various features in these models is clearly defined, SHAP can provide thorough explanations for each prediction. Meanwhile, the NB model is a probabilistic classification algorithm built on the Bayes theorem and based on the assumption of feature independence. The term "naive" describes the belief that each feature is conditionally independent of the feature assigned a class label [28]. Consequently, different scores were obtained.

Thus, the SHAP values presented in Fig. 5A were used to calculate the feature importance, which corresponds to the DT model. The rank of each feature's influence on the model is represented by the feature importance plots shown in Fig. 5B,D, which correspond to the RF and LR models, respectively. In addition, the influence-scaled score from each feature and class is represented by the heatmap produced for the NB model (see Fig. 5C). The result in Fig. 5 clearly demonstrate that the features with a greater influence on a model's choice include the exporting country ratio, the major category, the overseas manufacturer ratio, and the importer ratio. This suggests that the decision made by the model is more influenced by features with higher values.

Furthermore, the NB model (Fig. 5C) shows that the non-conformity decision of the model is highly dependent on the week and month of the year, the middle category of the product name and its ratio, the overseas manufacturer and importer, the exporting company, the product name and its ratio, and the export country. These factors play crucial roles in determining the non-conformity decisions made by the NB model. By considering various aspects, such as the week and month of the year, any potential seasonal variations that may impact product conformity were also considered. Moreover, factors such as the middle category of the product name and its ratio, overseas manufacturer, importer, exporting company, product name, and its ratio, and the exporting country allow the analysis of the various dimensions that could contribute to non-conformity.

4.2.2 SHAP Local Explanation

While global feature importance provides insights into model behavior across the entire dataset, SHAP local explanations reveal how specific features influence individual predictions. The study analyzed four non-conforming seafood import samples correctly identified by our ensemble model to demonstrate the model's decision-making process at the instance level.

Fig. 6A shows feature contributions for Sample 1, white leg shrimp imported from Vietnam with a prediction probability of 0.7865 for non-conformity. The exporting company ratio (0.5) strongly pushed toward non-conformity, followed by total net weight (1321.6 kg) and importer ratio (0.0083). Conversely, product name ("White Leg Shrimp") and temporal features (Year and Month = 2022-05) contributed significantly toward conformity classification.

Fig. 6B presents Sample 2, short-neck clams from China with the highest non-conformity prediction probability (0.9520) among the analyzed samples. For this case, document inspection was the dominant factor pushing toward non-conformity, along with quantity (233) and net weight (4660.0 kg). Unlike Sample 1, almost all features consistently contributed toward non-conformity, with only the exporting company ratio showing a negative influence.

Fig. 6C shows Sample 3, an imported octopus from China with a prediction probability of 0.8687. This sample demonstrates a distinctive pattern where the exporting company ratio (1.0) exerted the strongest influence toward non-conformity. Interestingly, both the name of the product ("Octopus") and the total quantity (2420) showed strong negative contributions, indicating that these features typically suggest conformity. However, these were outweighed by positive contributions from overseas manufacturer ratio (0.0195) and year-related indicators.

Fig. 6D presents Sample 4, red sea bream with a prediction probability of 0.8948. Similarly to other samples, the total net weight (4000.0 kg) strongly influenced the non-conformity prediction, along with temporal characteristics (year and week = 2019-28). In particular, the total quantity (1) differed significantly from other samples, pushing toward conformity, while the exporting company ratio similarly contributed to conformity.

This comparative analysis reveals several important patterns across diverse seafood products. Weight-related metrics consistently influence non-conformity predictions, with higher weights generally increasing risk. Ratio metrics derived from historical records (exporting company, manufacturer, importer) play significant but context-dependent roles. The same features can have opposite effects in different contexts, highlighting the model's ability to capture complex patterns. Temporal features contribute meaningful information across all samples, supporting the importance of seasonal variations identified in the global feature importance analysis. These SHAP explanations demonstrate the ensemble model's nuanced decision-making process and provide valuable insights for regulatory authorities to develop more targeted inspection strategies based on product-specific risk factors.

5. Discussion

This study addressed the significant challenge of predicting non-conformity in imported seafood products—a critical task in food safety management characterized by severe class imbalance. The methodological framework integrated ensemble learning techniques with explainable AI approaches to develop robust prediction models while maintaining interpretability. The empirical findings warrant comprehensive discussion across several dimensions.

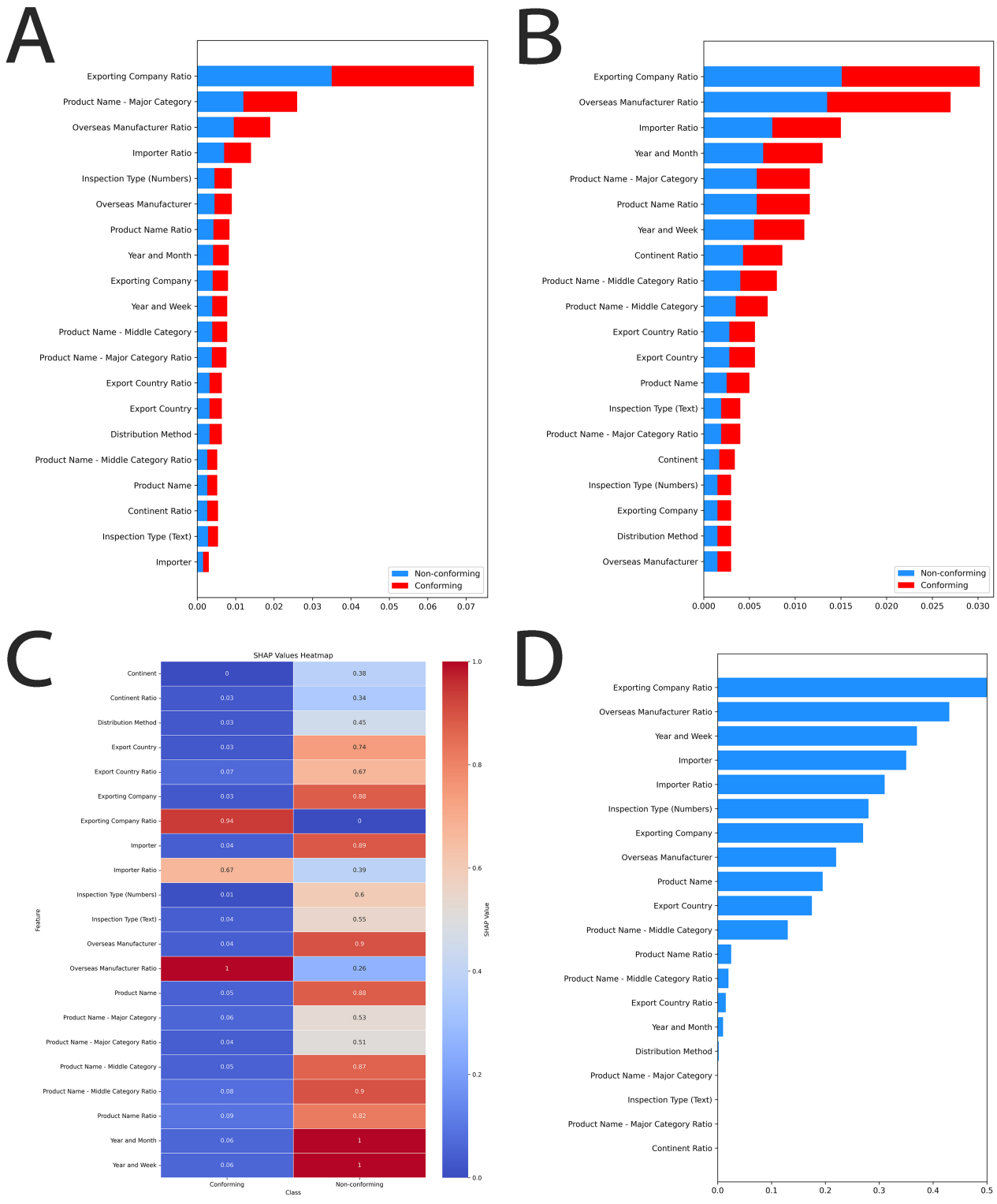

Fig. 5. SHAP values of the various models' importance feature scores. (A) DT model's important features. (B) RF model's important features. (C) Heatmap of normalized log-likelihood contributions in the NB model, with darker colors showing stronger influence on non-conformity classification. (D) LR model's important features.

Four different models were used, namely the NB, DT, RF and LR models, which were stacked together using class-weight cost-sensitive learning to create both hard-voting and soft-voting ensemble models. These models were applied to forecast nonconformance after training, and their performances were evaluated by using the test dataset.

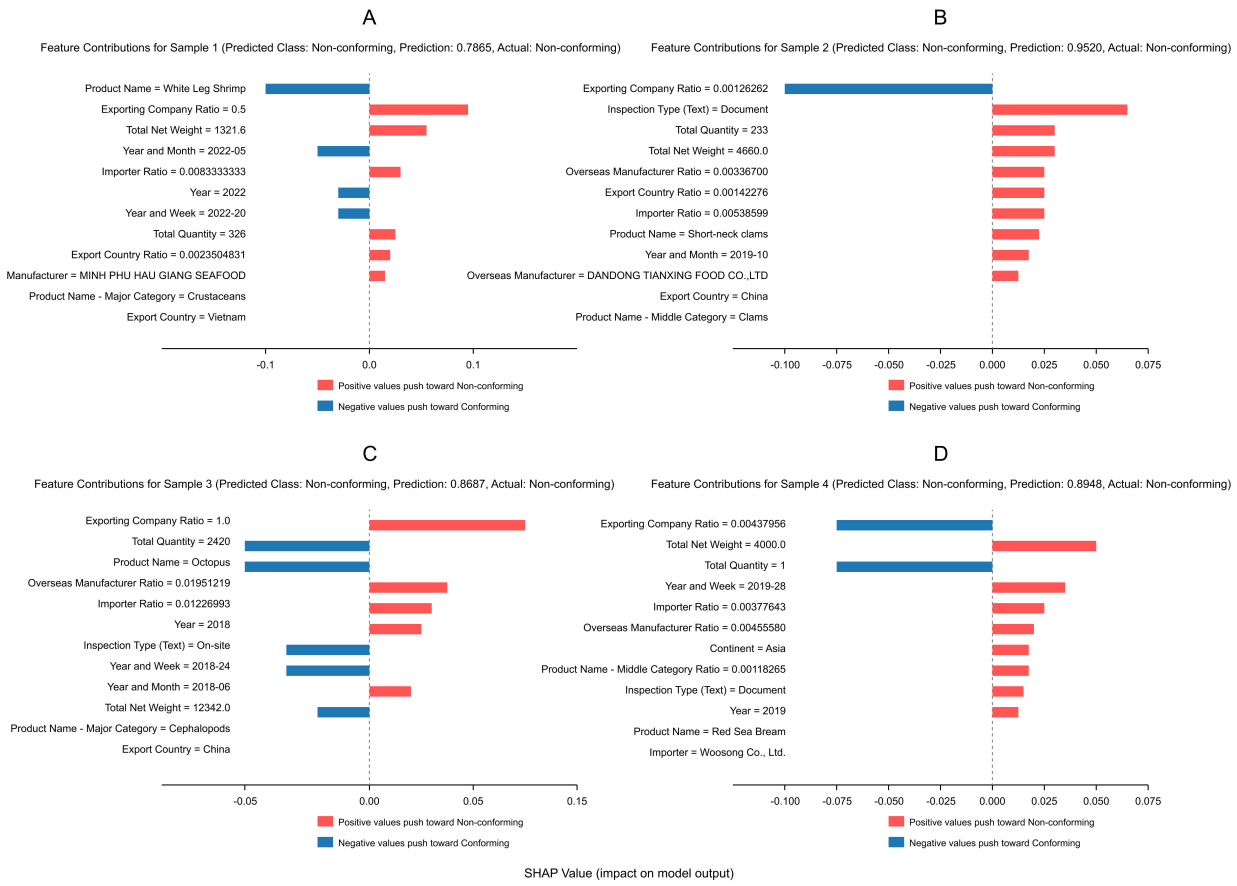

Fig. 6. SHAP local explanations using waterfall plot for individual seafood samples. (A) Sample 1: White leg shrimp (prediction: 0.7865). (B) Sample 4: Red sea bream (prediction: 0.8948). (C) Sample 3: Octopus (prediction: 0.8887). (D) Sample 2: Short-neck clams (prediction: 0.9520).

Table 3 summarizes the performance metrics of the models. Among the two ensemble methods, soft voting outperformed hard voting in terms of the recall score, receiving 75.57% of votes compared to the hard voting method's 44.32%. Additionally, the soft voting method achieved a higher AUC score of 87.49%, whereas the hard voting method obtained a score of 72.07%. However, hard voting achieved better results in terms of the ACR (99.69%, c.f., 99.35% for soft voting). Similarly, hard voting outperformed soft voting in the PPV and F1 scores, with scores of 35.62 and 39.49%, respectively; soft voting obtained scores of 22.32 and 34.46%, respectively. These findings suggest that the soft voting method performed better than the hard voting method in terms of the recall and AUC scores, thereby indicating that the soft voting ensemble is more effective in correctly identifying non-conformity data and predicting inspection outcomes. Thus, the soft voting approach may be more suitable in cases where the recall and AUC scores are more critical than the PPV and ACR scores. This finding aligns with established theoretical frameworks regarding ensemble techniques for imbalanced classification problems while extending empirical validation specifically to the domain of food safety inspection by Douzas *et al.*, and Wu & Weng [9,29,30].

To gain insight into the decision-making process of each model in the ensemble, SHAP values were calculated for all models. Additionally, local-level feature contributions to the predictions of four representative samples were analyzed to obtain deeper insights. Global feature importance analysis revealed the predominant influence of historical performance indicators, particularly the exporting country ratio, major product category, overseas manufacturer ratio, and importer ratio. These findings validate existing risk-based approaches to food safety management that prioritize historical compliance patterns as predictors of future regulatory adherence, as demonstrated by Djekic & Jankovic [31] in their analysis of food safety notifications in the European Union. More importantly, these findings resonate with existing risk-based inspection frameworks, such as the system developed by the Korean Ministry of Food and Drug Safety, outlined in patent KR20140077006A [8].

The exporting country ratio emerged as one of the most influential features across all models. This metric captures historical non-conformity rates by country, reflect-

ing differences in regulatory standards and supply chain controls. This finding aligns with Fanani and Yuadi [32], who identified Russia, Mauritania, Papua New Guinea, and Solomon Islands as high-risk exporters to the U.S. Our SHAP analysis confirms that country-specific factors strongly influence non-conformity predictions, indicating that regional disparities in fishing practices and compliance create identifiable risk patterns that can inform targeted inspection protocols.

The overseas manufacturer ratio significantly influences food safety outcomes, highlighting the importance of producer-specific factors in risk assessment. This reflects variations in quality control systems, manufacturing practices, and compliance histories at the facility level. Our local SHAP analysis Fig. 6B of Sample 2 revealed that even products from low-risk countries may present elevated risk when produced by manufacturers with higher non-conformity ratios. This finding supports inspection approaches focused on manufacturer-specific histories rather than relying solely on country-level assessments, consistent with Riviere & Buckley's [33] research on quality control systems and Suanin [34] findings on the importance of compliance histories in seafood product safety.

The importer ratio demonstrated substantial predictive power, reflecting the critical role of importer compliance history in forecasting inspection outcomes. As shown in Fig. 6A,C, varying importer ratios significantly impacted non-conformity predictions, with higher historical non-conformity rates strongly correlating with increased risk assessment. This pattern indicates that importer-specific factors, such as supplier selection practices, quality management systems, and regulatory compliance commitment, create persistent risk patterns that can be effectively captured through historical performance metrics. The model's ability to identify these patterns enables regulatory authorities to implement targeted verification protocols for importers with problematic compliance histories while potentially reducing inspection burdens on consistently high-performing entities. This risk-based approach to importer assessment aligns with Bouzembrak & Marvin's [35] framework for prioritizing inspection resources based on historical compliance data and supply chain characteristics.

The importance of temporal features—such as year, month, and week—in predicting seafood inspection outcomes highlights the presence of seasonal patterns in food safety risks. Seasonal variations, particularly fluctuations in water temperature, can significantly influence microbial activity in seafood. Warmer temperatures during summer months are associated with increased prevalence of pathogens like Vibrio species, a trend supported by Zhang *et al*. [36]. This seasonal impact is further illustrated in Fig. 6D, where a notable increase in non-conformity predictions is observed around the 28th week, corresponding to the summer season. Besides, seasonal harvesting patterns create periodic surges in processing volumes that can strain quality control systems. Additionally, rainfall variations affect coastal water quality through agricultural runoff, potentially introducing contaminants into harvesting areas during specific seasons [37]. In the local explanation Fig. 6C also identifies June, which is a rainy month, as the influence factor on the non-conformity result.

The major product category emerged as a significant predictor across all models, highlighting the variability in risk profiles among different seafood types. This finding aligns with the established understanding that certain seafood categories inherently pose higher contamination risks due to their biological characteristics and supply chain complexities. For example, filter-feeding bivalves like short-neck clams (Sample 2 in Fig. 6B) are particularly susceptible to environmental contaminants, while products requiring extensive processing may experience quality control challenges. The pronounced influence of product categories on model predictions was consistently observed in the SHAP analysis, with varied effects across different seafood types. This suggests that inspection resources should be strategically allocated based on product-specific risk profiles rather than applying uniform protocols across all seafood imports, an approach supported by FAO [38] research on compliance patterns in processed food exports.

6. Conclusion

This study addressed the significant challenge of developing a generalizable machine learning model for predicting non-conformity in seafood product importation inspections. By implementing an ensemble approach that combines Decision Trees, Random Forests, Logistic Regression, and Naive Bayes models, the study handled the severe class imbalance (0.2% non-conformity rate) inherent in food safety inspection data.

The soft voting ensemble technique demonstrated better performance in identifying non-conforming seafood products, achieving a recall of 75.57% and an AUC of 87.49%, significantly outperforming the hard voting method (44.32% recall, 72.07% AUC). Through SHAP analysis, the study identified key factors influencing inspection outcomes, including exporting country ratio, major product category, overseas manufacturer ratio, importer ratio, and seasonal variations.

The study findings suggest that historical compliance patterns serve as strong predictors of future regulatory adherence, with particular importance placed on country-specific factors and producer-specific compliance histories. The identification of temporal patterns indicates seasonal variations in food safety risks, likely influenced by environmental factors such as water temperature and rainfall patterns.

Importantly, while this research focused specifically on seafood products, the methodological framework developed here can be readily applied to other product categories within the food sector and potentially beyond. The

ensemble learning approach combined with explainable AI techniques provides a good template that can be adapted to predict inspection outcomes for diverse imported goods such as meat products, processed foods, and even non-food consumer goods subject to regulatory oversight. The core principles of leveraging historical compliance data, entity-specific risk factors, and temporal patterns remain applicable across various inspection domains facing similar class imbalance challenges.

The methodology developed in this study offers a practical framework for regulatory authorities to implement risk-based inspection protocols, enhancing both efficiency and effectiveness of limited inspection resources. By targeting high-risk imports based on multiple risk dimensions, authorities can simultaneously improve food safety and reduce unnecessary inspection burdens on consistently compliant entities.

Nevertheless, this study has the following limitations. Firstly, there are limitations in refining potential errors or unstructured items in the data, and some important variables (e.g., details of actual inspection criteria) were not included. Secondly, the ensemble model configuration used only four traditional classifiers, and no comparison was made with recently developed deep learning-based models. Thirdly, while SHAP analysis provided valuable insights into feature importance and local explanations, our implementation did not fully explore advanced SHAP capabilities such as interaction effects and dependence plots that could have offered deeper interpretability of the model's decision-making process. Lastly, there is a challenge in handling new entities, such as first-time importers, manufacturers, or products without historical data in the predictive framework.

Future research may consider the following directions: (1) linkage of in-depth feature engineering and external data (e.g., climate, supply chain risks, etc.), (2) multi-layer ensemble or hybrid model configuration and performance optimization based on AutoML, (3) empirical evaluation such as policy field application experiments and cost-effectiveness analysis. In particular, follow-up research is needed to maximize policy application performance through integrated scenario design with field inspection systems and linkage with real-time warning systems. (4) Development of transfer learning methods that apply knowledge from similar established entities to new ones. (5) Further exploration of advanced XAI techniques, including more sophisticated applications of SHAP analysis such as interaction values, dependence plots, and distribution analysis across product categories, which could provide deeper insights into model decision-making processes and enhance the interpretability of predictions for regulatory stakeholders.

Availability of Data and Materials

The data presented in this study are available on request from the corresponding author. The data are not publicly available due to concerns related to the protection of participant privacy and confidentiality.

Author Contributions

SK contributed substantially to the conceptualization and design of the study, performed formal analysis, developed the software, prepared the original draft, and created visualizations. KL contributed to the conceptualization of the study and managed the project. WC supervised the research and contributed to critical revision of the manuscript. All authors contributed to manuscript editing, read and approved the final manuscript, and agree to be accountable for all aspects of the work.

Ethics Approval and Consent to Participate

Not applicable.

Acknowledgment

We would like to thank Editage (https://www.editage.co.kr/) for English language editing. Additional English language refinement was performed using Claude by Anthropic. All the refined texts were reviewed and verified by the authors.

Funding

This research was supported by a grant (21163MFDS517-1) from Ministry of Food and Drug Safety.

Conflict of Interest

Wan-Sup Cho is the owner of BigDataLabs Co., Ltd., and Saksonita Khoeurn was employed by BigDataLabs Co., Ltd. The company develops and markets products related to the topic of this manuscript. However, the company had no role in the design, analysis, or interpretation of the study. The authors declare that there are no conflicts of interest.

Supplementary Material

Supplementary material associated with this article can be found, in the online version, at https://doi.org/10.31083/JFSFQ39242.

References

[1] FAO. ICES–FAO Working Group on Fishing Technology and Fish Behaviour - Report of the 2023 Symposium on Innovations in Fishing Technologies for Sustainable and Resilient Fisheries, 13-17 February 2023, Kochi, India. FAOFisheries and Aquaculture Report, No. 1432. Rome. 2024. https://doi.org/10.4060/cd0312en.

[2] Salama Y, Chennaoui M. Microbial spoilage organisms in seafood products: pathogens and quality control. European Jour-

nal of Microbiology and Infectious Diseases. 2024; 1: 66–89. https://doi.org/10.5455/EJMID.20240518114533.
[3] Ministry of Food and Drug Safety. Imported Food Information Maru. 2018. Available at: https://impfood.mfds.go.kr/ (Accessed: 26 July 2023).
[4] Visciano P. Environmental Contaminants in Fish Products: Food Safety Issues and Remediation Strategies. Foods. 2024; 13: 3511. https://doi.org/10.3390/foods13213511.
[5] Marvin HJP, Bouzembrak Y. A system approach towards prediction of food safety hazards: Impact of climate and agrichemical use on the occurrence of food safety hazards. Agricultural Systems. 2020; 178: 102760. https://doi.org/10.1016/j.agsy.2019.102760.
[6] Paolacci S, Mendes R, Klapper R, Velasco A, Ramilo-Fernandez G, Muñoz-Colmenero M, *et al*. Labels on seafood products in different European countries and their compliance to EU legislation. Marine Policy. 2021; 134: 104810. https://doi.org/10.1016/j.marpol.2021.104810.
[7] Gómez-Andújar NX, Castillo LS, Rivera-Hechem MI, Gaines S, Quintana ACE. Moving beyond binary metrics of compliance in small-scale fisheries. Ecology and Society. 2024; 29: 39. https://doi.org/10.5751/ES-15688-290439.
[8] Hong HW, Shin HS, Dong H, inventor; Ministry of Food and Drug Safety, Republic of Korea, assignee. Risk Prediction-Based Imported Food Inspection System and Method. Republic of Korea, KR20140077006A. 23 June 2014. https://patents.google.com/patent/KR20140077006A/ko.
[9] Wu LY, Weng SS. Ensemble Learning Models for Food Safety Risk Prediction. Sustainability. 2021; 13: 12291. https://doi.org/10.3390/su132112291.
[10] Park J, Lee WH, Kim KT, Park CY, Lee S, Heo TY. Interpretation of ensemble learning to predict water quality using explainable artificial intelligence. The Science of the Total Environment. 2022; 832: 155070. https://doi.org/10.1016/j.scitotenv.2022.155070.
[11] Hellen N, Sabuj HH, Ashraful Alam M. Explainable AI and Ensemble Learning for Water Quality Prediction. In Proceedings of International Conference on Information and Communication Technology for Development: ICICTD 2022 (pp. 235–250). Singapore: Springer Nature Singapore. 2023, January. https://doi.org/10.1007/978-981-19-7528-8_19.
[12] Gong H, Wang M, Zhang H, Elahe MF, Jin M. An Explainable AI Approach for the Rapid Diagnosis of COVID-19 Using Ensemble Learning Algorithms. Frontiers in Public Health. 2022; 10: 874455. https://doi.org/10.3389/fpubh.2022.874455.
[13] Saraswat D, Bhattacharya P, Verma A, Prasad VK, Tanwar S, Sharma G, *et al*. Explainable AI for healthcare 5.0: opportunities and challenges. IEEE Access. 2022; 10: 84486–84517. https://doi.org/10.1109/ACCESS.2022.3197671.
[14] Lundberg SM, Lee SI. A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems 30 (NIPS 2017). Long Beach, California. Curran Associates, Inc. 2017.
[15] Kotsiantis S, Kanellopoulos D, Pintelas P. Handling imbalanced datasets: A review. GESTS International Transactions on Computer Science and Engineering. 2006; 30: 25–36.
[16] Chawla NV, Bowyer KW, Hall LO, KegelmeyerWP. SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research. 2002; 16: 321–357. https://doi.org/10.1613/jair.953.
[17] Han H, Wang WY, Mao BH. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. In International conference on intelligent computing (pp. 878–887). Berlin, Heidelberg: Springer Berlin Heidelberg. 2005, August. https://doi.org/10.1007/11538059_91.
[18] Nguyen HM, Cooper EW, Kamei K. Borderline over-sampling for imbalanced data classification. International Journal of Knowledge Engineering and Soft Data Paradigms. 2011; 3: 4–21. https://doi.org/10.1504/IJKESDP.2011.039875.
[19] Bamhdi AM, Abrar I, Masoodi F. An ensemble based approach for effective intrusion detection using majority voting. TELKOMNIKA (Telecommunication Computing Electronics and Control). 2021; 19: 664–671. http://doi.org/10.12928/telkomnika.v19i2.18325.
[20] Brownlee J. Ensemble Learning Methods for Deep Learning Neural Networks. 2019. Available at: https://machinelearningmastery.com/ensemble-methods-for-deep-learning-neural-networks/ (Accessed: 26 July 2023).
[21] Arrieta AB, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, *et al*. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion. 2020; 58: 82–115. https://doi.org/10.1016/j.inffus.2019.12.012.
[22] Sinharay S. An Overview of Statistics in Education. In Peterson P, Baker E, McGaw B (eds.) International Encyclopedia of Education (pp. 1–11). 3rd edn. Elsevier: Amsterdam, Netherlands. 2010. https://doi.org/10.1016/B978-0-08-044894-7.01719-X.
[23] Myles AJ, Feudale RN, Liu Y, Woody NA, Brown SD. An Introduction to Decision Tree Modeling. Journal of Chemometrics. 2004; 18: 275–285. https://doi.org/10.1002/cem.873.
[24] Rigatti SJ. Random Forest. Journal of Insurance Medicine. 2017; 47: 31–39. https://doi.org/10.17849/insm-47-01-31-39.1.
[25] LaValley MP. Logistic regression. Circulation. 2008; 117: 2395–2399. https://doi.org/10.1161/CIRCULATIONAHA.106.682658.
[26] Sammut C, Webb GI. Encyclopedia of machine learning. Springer Science & Business Media: Berlin, Germany. 2011.
[27] Ling CX, Sheng VS. Cost-sensitive learning and the class imbalance problem. Encyclopedia of Machine Learning. 2008; 2011: 231–235.
[28] Rish I. An empirical study of the naive Bayes classifier. In IJCAI 2001 workshop on empirical methods in artificial intelligence (Vol. 3, No. 22, pp. 41–46). 2001, August.
[29] Sunny S, Pinky S, Jalal S, Kayser M, Wadud M, Mansoor N. Soft Voting Ensemble-Based Approach for Diagnosing Diabetes Mellitus. In 2024 International Conference on Advances in Computing, Communication, Electrical, and Smart Systems (iCACCESS) (pp. 01–06). IEEE. 2024, March. https://doi.org/10.1109/iCACCESS61735.2024.10499577.
[30] Douzas G, Bacao F, Last F. Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE. Information Sciences. 2018; 465: 1–20. https://doi.org/10.1016/j.ins.2018.06.056.
[31] Djekic I, Jankovic D, Rajkovic A. Analysis of foreign bodies present in European food using data from Rapid Alert System for Food and Feed (RASFF). Food Control. 2017; 79: 143–149. https://doi.org/10.1016/j.foodcont.2017.03.047.
[32] Fanani MI, Yuadi I. Predictive Comparative Analysis for Fundamental Risk of US Import and Economy Seafood Market. In 2023 Sixth International Conference on Vocational Education and Electrical Engineering (ICVEE) (pp. 35–40). IEEE. 2023, October. https://doi.org/10.1109/ICVEE59738.2023.10348296.
[33] Committee on Strengthening Core Elements of Regulatory Systems in Developing Countries, Board on Global Health, Board on Health Sciences Policy, Institute of Medicine. Ensuring Safe Foods and Medical Products Through Stronger Regulatory Systems Abroad. In Riviere JE, Buckley GJ (eds.). National Academies Press (US): Washington (DC). 2012. https://doi.org/10.17226/13296.
[34] Suanin W. Processed food exports from developing countries: the effect of food safety compliance. European Review of Agricultural Economics. 2023; 50: 743–770. https://doi.org/10.

1093/erae/jbac030.
[35] Bouzembrak Y, Marvin HJP. Prediction of food fraud type using data from Rapid Alert System for Food and Feed (RASFF) and Bayesian network modelling. Food Control. 2016; 61: 180–187. https://doi.org/10.1016/j.foodcont.2015.09.026.
[36] Zhang Z, Wang F, Lei L, Zheng N, Shen Z, Mu J. Spatio-temporal dynamics of the carbonate system during macroalgae farming season in a semi-closed bay in southeast China. Frontiers in Marine Science. 2024; 11: 1375839. https://doi.org/10.3389/fmars.2024.1375839.
[37] Tian B, Chang S, Ye S, Zhang Y, Wang Y, Wang S, *et al*. Spatio-Temporal Dynamics of Fish Community and Influencing Factors in an Urban River (Haihe River), China. Sustainability. 2025; 17: 231. https://doi.org/10.3390/su17010231.
[38] Food and Agriculture Organization of the United Nations. Guidelines for risk-based fish inspection. FAO Food and Nutrition Paper. 2009; 90: 3–89. https://www.fao.org/4/i0468e/i0468e00.htm.