Preprints.org

Article

Not peer-reviewed version

Are Food Safety Classifiers Learning Hazards or Memorizing Firms? Entity-Level Leakage in FDA Recall Severity Prediction

Peilun Li [†] and Juk-Sen Tang [*,†]

Posted Date: 4 March 2026

doi: 10.20944/preprints202603.0343.v1

Keywords: food recall severity; entity-level leakage; evaluation bias; predictive regulatory triage; temporal concept drift; explainable AI (XAI); openFDA

Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2026 doi:10.20944/preprints202603.0343.v1

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

Are Food Safety Classifiers Learning Hazards or Memorizing Firms? Entity-Level Leakage in FDA Recall Severity Prediction

Peilun Li [1,†] and **Juk-Sen Tang** [2,*,†]

[1] Department of Food Science and Agricultural Chemistry, MacDonald Campus, McGill University, Montréal, QC H9X 3V9, Canada
[2] Department of Agricultural Economics, MacDonald Campus, McGill University, Montréal, QC H9X 3V9, Canada
* Correspondence: juksen.tang@mail.mcgill.ca
† These authors contributed equally to this work.

Abstract

Machine learning (ML) models for predicting food recall severity could accelerate regulatory triage, yet no systematic benchmark exists on the U.S. Food and Drug Administration (FDA) open-access database. We construct the first comprehensive ML benchmark for FDA food recall severity classification (Class I / II / III) using 28,448 enforcement records spanning 2012–2025. A 1,437-dimensional feature space is engineered from TF-IDF and Sentence-BERT embeddings of recall narratives, structured categorical attributes, and temporal indicators. Five classifiers (Logistic Regression, Random Forest, XGBoost, LightGBM, CatBoost) are trained with Optuna-tuned hyperparameters. Under standard random splitting, XGBoost achieves Macro-F1 = 0.89; however, a multi-layer leakage audit reveals that this figure is inflated by entity-level autocorrelation. When firm-aware group splitting, temporal splitting, or their combination is applied, Macro-F1 drops to approximately 0.57. A firm-mode baseline—assigning each company's historically most frequent severity class—reaches 0.82 under random splitting, demonstrating that 92% of the apparent performance stems from firm-level memorisation. Identity-masking experiments confirm that the leakage is structural rather than attributable to explicit company-name tokens. A 2×2 factorial decomposition shows that firm overlap and temporal continuity are highly collinear; removing either suffices to expose the true generalisation floor. A hazard-type decomposition reveals that pathogen–severity associations transfer across firms, whereas labelling and GMP violations are highly firm-specific, explaining the disproportionate collapse of Class III prediction under group splitting. SHAP analysis, feature ablation, and a nine-year continuous-learning simulation provide additional insights into model behaviour and retraining strategies. We recommend that food-safety ML studies adopt group-aware or temporal evaluation protocols, report entity-overlap statistics, and include entity-prior baselines to prevent overstated conclusions.

Keywords: food recall severity; entity-level leakage; evaluation bias; predictive regulatory triage; temporal concept drift; explainable AI (XAI); openFDA

1. Introduction

The U.S. Food and Drug Administration (FDA) classifies food recalls into three severity tiers: Class I denotes situations where exposure to a violative product carries a reasonable probability of serious adverse health consequences or death; Class II covers cases where exposure may cause temporary or medically reversible health effects; and Class III applies when the product is unlikely to cause any adverse health consequence [1]. This tiered system is central to regulatory decision-making, guiding resource allocation, public notification urgency, and follow-up monitoring intensity. The classification process currently relies on expert evaluation, which, although thorough, faces growing time pressure as the volume of recall events continues to rise. Between 2012 and 2025, the openFDA

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2026 doi:10.20944/preprints202603.0343.v1

Food Enforcement database accumulated over 28,000 recall records spanning a diverse array of product categories, hazard types, and reporting firms [2]. The public health stakes are considerable: the most recent U.S. estimates attribute approximately 9.4 million illnesses, 55,961 hospitalisations, and 1,351 deaths annually to major foodborne pathogens alone [3], imposing an economic burden exceeding $75 billion per year [4]. At the firm level, severe recall events can erode over $100 million in shareholder value within days [5,6], with Class I events in the low-moisture food sector alone associated with median market capitalisation losses of $243 million [7]. Moreover, the economic disruption propagates along the entire supply chain [8], underscoring the need for accurate, automated severity triage. The FDA itself has recognised this imperative in its New Era of Smarter Food Safety blueprint, which identifies predictive analytics as a core element of modernised food safety [9]. This strategic direction builds on the foundation laid by the Food Safety Modernization Act (FSMA) [10], which shifted the U.S. regulatory paradigm from reactive response to preventive control. FSMA grants the FDA mandatory recall authority and requires firms to maintain hazard analysis and preventive control plans, creating a regulatory environment in which predictive severity triage tools could directly support compliance and enforcement workflows.

Machine learning (ML) has emerged as a promising tool for automating risk assessment in food safety [11,12]. A growing body of literature has demonstrated the potential of supervised classifiers on the European Rapid Alert System for Food and Feed (RASFF) database, achieving accuracies ranging from 74% to 97.8% across various formulations [13–15]. Natural language processing (NLP) techniques have further expanded the feature space available for food-safety models, with recent work applying Transformer-based architectures to hazard detection tasks [16]. Importantly, recent large-scale empirical studies have demonstrated that gradient-boosted tree models remain competitive with or superior to deep learning on medium-scale tabular data [17,18], providing a strong empirical basis for the model family employed in this study.

Despite this progress, almost all existing studies target the EU RASFF database or other non-FDA sources. To date, no peer-reviewed study has attempted systematic Class I/II/III severity prediction on openFDA food recall data. Two tangentially related works exist: an unpublished preprint that analyses openFDA recall patterns but predicts termination status rather than severity [19], and a study of FDA drug (not food) recalls limited to 235 records [20]. This leaves a significant gap in the literature for a rigorous ML benchmark on the largest publicly available food recall database in the United States.

Beyond the absence of FDA-specific benchmarks, a more fundamental concern pervades the food-safety ML literature: evaluation methodology. The prevailing practice is to split data randomly into training and test sets, implicitly assuming that test-time instances are drawn from the same distribution as training data. This assumption is violated when the same reporting entity (e.g., a food manufacturer) contributes multiple records, because recalls from the same firm share product types, hazard profiles, and distribution patterns. Recent econometric work has formally demonstrated that naively applying standard cross-validation to panel data—data with repeated observations of the same entities over time—leads to severely inflated out-of-sample performance [21]. Food recall databases, where the same firms contribute multiple records across years, constitute precisely such panel structures. Bouzembrak and Marvin [22] provided early evidence of this issue in their Bayesian network study on RASFF food fraud, observing that accuracy dropped from 80% on previously seen country–product combinations to 52% on unseen ones. However, their observation remained incidental rather than systematic.

In clinical ML, the analogous problem—patient-level data leakage—has received extensive attention. Studies on histopathological image classification [23] and ECG analysis [24] have demonstrated that failing to segregate data by patient identity inflates reported performance substantially. Kapoor and Narayanan [25] further provided a comprehensive taxonomy of leakage scenarios across ML-based science. Yet food-safety research has largely not internalised these lessons: none of the recent high-performing RASFF studies [13,14] report group-aware or temporal evaluation protocols.

This study addresses the above gaps through four objectives. First, we construct the first comprehensive ML benchmark for FDA food recall severity classification, comprising 28,448 records, 1,437

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2026 doi:10.20944/preprints202603.0343.v1

engineered features, five tuned classifiers, and a rule-based baseline. Second, we design a multi-layer leakage audit with four splitting strategies (Random, Group-by-firm, Temporal, Group+Temporal), a firm-mode baseline, and identity-masking experiments to quantify entity-level autocorrelation. Third, we apply a 2×2 factorial design (firm overlap $\times$ time overlap) to decompose the contributions of firm-level memorisation and temporal concept drift. Fourth, based on our findings, we propose concrete evaluation protocol recommendations for food-safety ML research.

The remainder of this paper is organised as follows. Section 2 reviews related work on food-safety ML and data leakage. Section 3 describes the data, feature engineering, models, and evaluation protocols. Section 4 presents results across all evaluation dimensions. Section 5 discusses implications, practical value, and limitations. Section 6 concludes with recommendations.

2. Related Work

2.1. Machine Learning for Food Safety Prediction

We organise the literature by data source, proceeding from the most to least directly relevant to our work.

FDA studies.

Despite the scale and public availability of the openFDA database, ML research on FDA food recalls remains scarce. A ResearchGate preprint [19] applied clustering and basic classification to openFDA records from 2017–2025 but targeted recall termination status rather than severity. Mulla and Patel [20] analysed 235 FDA drug recalls (2018–2023) using text analytics, but the limited sample size and drug-specific scope restrict generalisability to food recalls. To our knowledge, *no peer-reviewed study has addressed Class I/II/III multi-class severity prediction on openFDA food recall data*.

EU RASFF studies.

The European RASFF database has attracted considerably more attention. Nogales et al. [13] compared neural (MLP, 1D-CNN) and non-neural (Random Forest, SVM, Boosting) classifiers, achieving 72–89% accuracy using categorical entity embeddings and standard random splitting. Sari et al. [14] pushed the state of the art to 97.8% accuracy with BERT and RoBERTa models augmented by LIME and SHAP explanations; a companion study [26] provided a systematic comparison across classical and Transformer-based architectures. Papadopoulos et al. [15] reported 74% accuracy with MLP and CNN models on RASFF data. Bouzembrak and Marvin [22] applied Bayesian networks to predict food fraud types in RASFF, achieving 80% accuracy on previously seen country–product combinations but only 52% on unseen combinations—an early, though unsystematised, observation of entity-level generalisation gaps.

A critical commonality across these studies is their reliance on standard random splitting for evaluation. As we demonstrate in Section 4, entity-level autocorrelation—a structural property of regulatory databases where the same entities contribute repeated records—can inflate performance substantially under such protocols.

NLP frontiers in food safety.

Randl et al. [16] organised the SemEval-2025 shared task on food hazard detection, assembling 6,644 multi-national incident reports for hazard and product category extraction. Top-performing systems employed BERT, RoBERTa, and large language models, reaching Macro-F1 = 0.82 on coarse-grained categories. While this work showcases state-of-the-art NLP on food-safety text, it focuses on information extraction rather than severity prediction and does not engage with the regulatory classification logic underpinning systems such as the FDA's three-tier framework. Recent reviews have further highlighted the growing role of ML in food safety risk assessment [27], though standardised evaluation protocols remain an open challenge.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2026 doi:10.20944/preprints202603.0343.v1

2.2. Data Leakage in Machine Learning Evaluation

Data leakage occurs when information from the test distribution inadvertently enters the training process, leading to inflated performance estimates [25]. The most relevant variant for our setting is *entity-level leakage*, whereby multiple records from the same real-world entity (patient, firm, sensor) appear in both training and test sets.

In clinical ML, the consequences of entity-level leakage are well documented. Patient-level data segregation has been shown to substantially affect reported accuracy in histopathological image classification [23]; similarly, in ECG analysis, leave-source-out cross-validation reveals that standard K-fold designs overestimate performance on unseen recording sources [24]. Kapoor and Narayanan [25] provided a broader taxonomy of leakage scenarios, demonstrating that the problem pervades ML-based science across disciplines including political science, materials science, and genomics.

In food safety, however, *entity-level leakage has received almost no systematic attention*. Bouzembrak and Marvin [22] noted the 80%→52% accuracy drop for unseen entity combinations but did not pursue a decomposition of the phenomenon. The high-performing systems cited above [13,14] do not report group-aware or temporal evaluation results. This leaves open the question of how much reported performance in food-safety ML reflects genuine pattern learning versus entity memorisation.

2.3. Summary of Research Gaps

Three gaps motivate the present study. First, no systematic ML benchmark exists for FDA food recall severity (Class I/II/III) prediction. Second, the food-safety ML literature lacks entity-level leakage audits analogous to those now standard in clinical ML. Third, no prior work has quantitatively decomposed evaluation bias into firm-level autocorrelation and temporal concept drift. We address all three gaps with a comprehensive benchmark and a five-layer leakage audit framework.

3. Materials and Methods

3.1. Data Collection and Preprocessing

3.1.1. Data Source

All records were retrieved from the openFDA Food Enforcement API [2], which provides structured metadata for every FDA food recall event since 2012. The dataset spans 20 June 2012 to 25 February 2025 and contains 28,448 unique enforcement records after removing one duplicate. The `status` field was excluded from the feature set to prevent target leakage, as recall status is determined after severity classification.

3.1.2. Class Distribution

The target variable is the three-level severity classification assigned by the FDA: Class I (12,542 records, 44.1%), Class II (14,250, 50.1%), and Class III (1,656, 5.8%). The pronounced imbalance toward the minority Class III motivated the use of balanced sample weights throughout model training.

Table 1. Dataset summary.

Attribute	Value
Source	openFDA Food Enforcement API
Time span	2012-06-20 to 2025-02-25
Total records	28,448
Unique firms	4,197
Class I / II / III	12,542 (44.1%) / 14,250 (50.1%) / 1,656 (5.8%)

3.2. Feature Engineering

A 1,437-dimensional feature vector was constructed for each record, comprising three groups: NLP features, structured features, and temporal features.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2026 doi:10.20944/preprints202603.0343.v1

3.2.1. NLP Features (1,392 dimensions)

- **TF-IDF (1,000 d):** Term Frequency–Inverse Document Frequency vectors were extracted separately from the `reason_for_recall` and `product_description` fields (500 dimensions each), using unigrams and bigrams with sublinear term-frequency scaling.
- **Sentence-BERT (384 d):** The `reason_for_recall` text was encoded using the `all-MiniLM-L6-v2` model [28] to capture semantic similarity.
- **Hazard keywords (5 d):** Binary indicators for five hazard categories (allergen, pathogen, foreign material, labelling defect, chemical contamination), derived from keyword dictionary matching on the recall reason text. These categories reflect the principal hazard classes recognised in food-safety risk assessment [29]: pathogen contamination carries the highest acute health risk and is almost synonymous with Class I severity, while allergen and chemical hazards span a wider severity range depending on the specific agent and exposure level.
- **Text meta-features (3 d):** Character length and word count of the recall reason, and character length of the product description.

3.2.2. Structured Features (40 dimensions)

Structured features include: voluntary/mandated recall status (1 d); distribution scope categories (5 d, one-hot); number of affected states (1 d); state-group indicators (21 d, grouped by Census region); product category indicators (11 d, one-hot); and the reporting firm's historical recall count (1 d), computed using only temporally preceding records to prevent look-ahead leakage. Product category is included because different food matrices carry inherently different risk profiles—ready-to-eat products and those with high water activity are disproportionately associated with pathogen-related Class I recalls. Distribution scope reflects the public health footprint of a recall: nationwide distribution amplifies potential exposure and typically correlates with higher severity classification. The firm's historical recall count captures systematic food-safety management deficiencies: firms with repeated recalls often share persistent facility-level hazards such as environmental pathogen harbourage [30].

3.2.3. Temporal Features (5 dimensions)

Five temporal features were derived from the recall report date: `report_month`, `report_quarter`, `report_year`, `response_lag_days` (days between the report date and the event initiation date), and `rolling_12m_count` (the firm's recall count in the preceding 12 months).

Table 2. Feature group summary.

Group	Dimensions	Description
TF-IDF	1,000	Bag-of-words from reason & product text
Sentence-BERT	384	Semantic embedding of recall reason
Hazard keywords	5	Binary hazard-category indicators
Text meta-features	3	Length and word-count statistics
Structured categorical	39	Recall type, scope, state, product category
Firm prior recalls	1	Historical recall count (temporal-safe)
Temporal	5	Date-derived features
Total	**1,437**	

3.3. Model Selection and Training

Five classifiers were evaluated: Logistic Regression (LR), Random Forest (RF), XGBoost [31], LightGBM [32], and CatBoost [33]. The choice of gradient-boosted tree ensembles as the primary model family is supported by recent benchmarking evidence showing that tree-based methods consistently match or outperform deep learning on medium-scale tabular classification tasks [17,18,34,35], and that XGBoost, LightGBM, and CatBoost achieve near-identical performance when given comparable hyperparameter budgets [36]. Hyperparameters were tuned with Optuna [37] using the Tree-structured

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2026 doi:10.20944/preprints202603.0343.v1

Parzen Estimator (TPE) sampler over 20 trials with 3-fold stratified cross-validation. Class imbalance was addressed through balanced sample weights for all models. GPU acceleration (NVIDIA L4) was used for the three gradient-boosted tree models. A rule-based baseline that assigns severity based on hazard keyword presence was included for reference.

3.4. Evaluation Protocol Design

The evaluation protocol is the methodological core of this study. We designed four splitting strategies, a firm-mode baseline, and an identity-masking experiment to systematically quantify evaluation bias.

3.4.1. Four Splitting Strategies

Table 3. Evaluation splitting strategies.

Strategy	Abbr.	Train set	Test set	Simulates
Random Stratified	R	Random 80%	Random 20%	Standard practice
Group (by firm)	G	Firm set A	Firm set B (disjoint)	Cross-firm generalisation
Temporal	T	$\leq$2023	$\geq$2024	Future prediction
Group + Temporal	GT	$\leq$2023	$\geq$2024, new firms	Strictest scenario

Our multi-split design is informed by the panel-data leakage framework of Cerqua et al. [21], who showed that both cross-sectional and longitudinal dimensions of panel data can independently introduce leakage. The Random (R) split serves as a reproduction of standard literature practice. The Group (G) split ensures that no firm appears in both training and test sets, isolating cross-firm generalisation. The Temporal (T) split uses a fixed cutoff (end of 2023) to simulate prospective deployment. The Group+Temporal (GT) split combines both constraints, representing the strictest evaluation scenario.

3.4.2. Bootstrap Robustness

For the R, G, and GT strategies, experiments were repeated with five random seeds (42, 123, 456, 789, 2024) and results are reported as mean $\pm$ standard deviation. The T strategy uses a fixed temporal cutoff and is therefore deterministic.

3.4.3. Firm-Mode Baseline

A non-ML baseline was constructed by assigning each firm the severity class most frequently observed in its training-set recalls. For firms absent from the training set, the global majority class was assigned. This baseline quantifies the upper bound of performance achievable through entity memorisation alone.

3.4.4. Identity-Masking Experiment

To determine whether leakage operates through explicit company-name tokens in the text features, all occurrences of firm names in `reason_for_recall` and `product_description` were replaced with the placeholder `[FIRM]` using dictionary matching. TF-IDF features were then re-extracted from the masked text, and models were retrained under all four splitting strategies.

3.4.5. 2 $\times$ 2 Factorial Decomposition

The four splits form a 2 $\times$ 2 factorial design crossing firm overlap (same firms vs. new firms in test) with temporal overlap (IID vs. future test data). This design enables the decomposition of evaluation bias into firm-level autocorrelation, temporal concept drift, and their interaction.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2026 doi:10.20944/preprints202603.0343.v1

3.5. Interpretability Analysis

In regulatory decision-making contexts, model interpretability is not merely an academic preference but a practical requirement for building stakeholder trust [38]. Indeed, Rudin [39] has argued that high-stakes domains should prioritise inherently interpretable models; when complex models are necessary, post-hoc explanations become essential. Recent food-safety studies have begun integrating post-hoc explanation methods such as SHAP and LIME to enhance model transparency [14,40], and regulatory agencies including the FDA are increasingly attentive to explainability standards [41].

SHAP (SHapley Additive exPlanations) [42] TreeExplainer was applied to the XGBoost model using a subsample of 500 test instances. Three levels of analysis were conducted: global feature importance (mean absolute SHAP values across all classes), per-class feature importance (disaggregated by severity class), and dependence plots for the three most influential features.

3.6. Supplementary Experiments

3.6.1. Feature Ablation

Six configurations were evaluated under the Random split to assess the contribution of each feature group: Full (all 1,437 features), No NLP (structured + temporal only), NLP Only (1,392 NLP features), No Temporal (all except temporal features), No SBERT (all except Sentence-BERT), and No TF-IDF (all except TF-IDF features).

3.6.2. Continuous Learning Simulation

Temporal concept drift—the phenomenon whereby data distributions shift over time [43]—may degrade models trained on historical data. Three retraining strategies were compared over nine annual evaluation windows (2018–2026): *Static* (train on data up to 2017 only), *Expanding* (train on all data up to year $t-1$), and *Sliding* (train on the most recent three years). This simulation assesses the importance of data recency versus volume for maintaining predictive performance under temporal drift.

4. Results

4.1. Model Performance Under Standard Evaluation

Table 4 presents the performance of all six models under the standard 80/20 random stratified split. XGBoost achieves the highest Macro-F1 of 0.891, closely followed by LightGBM (0.889) and CatBoost (0.888). The three gradient-boosted tree ensembles are virtually indistinguishable, suggesting that model architecture is not the performance bottleneck. Random Forest achieves 0.862, while Logistic Regression (0.483) falls below even the rule-based baseline (0.502), confirming the non-linear nature of the classification task.

Table 4. Model performance under 80/20 random stratified split.

Model	Accuracy	Macro-F1	W-F1	Kappa	AUC-OVR
Rule-Based	0.698	0.502	0.675	0.441	—
LR	0.571	0.483	0.599	0.300	0.735
RF	0.907	0.861	0.905	0.828	0.982
XGBoost	0.931	0.891	0.930	0.873	0.988
LightGBM	0.929	0.889	0.928	0.869	0.988
CatBoost	0.925	0.888	0.925	0.863	0.982

W-F1 = Weighted F1. AUC-OVR = One-vs-Rest Area Under the ROC Curve.

Figure 1 visualises the comparison across Accuracy, Macro-F1, and Cohen's Kappa. The confusion matrix for XGBoost (Figure 2) reveals strong Class I and II discrimination but notable confusion involving Class III, consistent with the severe class imbalance (5.8% of records).

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2026 doi:10.20944/preprints202603.0343.v1

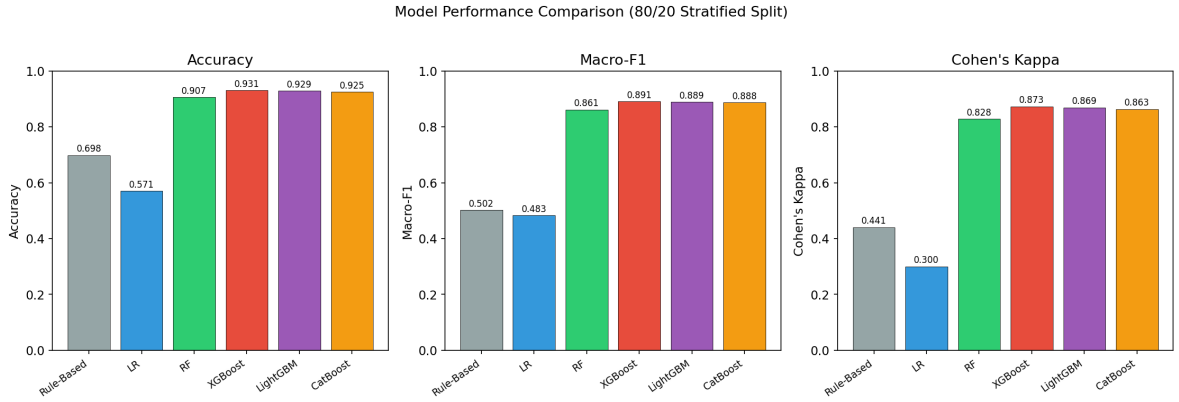

Figure 1. Model comparison under random stratified split: Accuracy, Macro-F1, and Cohen's Kappa for all six models.

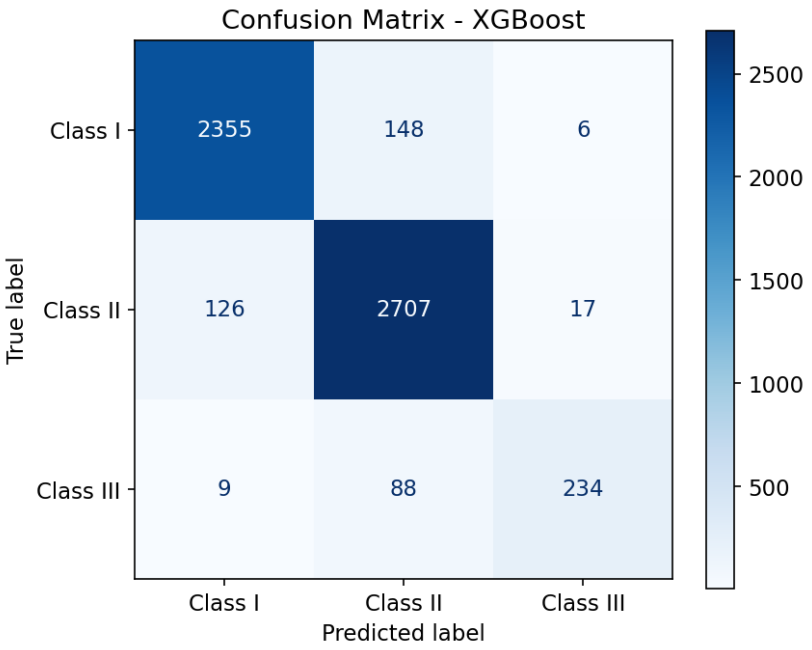

Figure 2. Confusion matrix for XGBoost under the 80/20 random split.

An additional binary formulation (Class I vs. rest) yielded AUC = 0.988, indicating that distinguishing high-severity recalls from lower-severity ones is highly feasible under random evaluation—a point revisited in Section 5. The temporal split result for XGBoost (Macro-F1 = 0.578) revealed a dramatic 0.313-point gap compared to the random split, motivating the systematic leakage audit that follows.

4.2. Leakage Audit: Systematic Comparison Across Four Splits

Table 5 presents XGBoost performance across the four splitting strategies. The Random split yields Macro-F1 = 0.891 ± 0.008, while Group, Temporal, and Group+Temporal splits produce 0.556 ± 0.024, 0.578, and 0.570 ± 0.008, respectively. The $\sim$0.32-point chasm between Random and the three leakage-aware splits is statistically robust: bootstrap confidence intervals do not overlap.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2026 doi:10.20944/preprints202603.0343.v1

Table 5. XGBoost performance across four splitting strategies (mean ± SD over 5 seeds; Temporal is deterministic).

Split	Macro-F1	Balanced Acc.	W-F1	Kappa
Random (R)	0.891 ± 0.008	0.866 ± 0.013	0.928 ± 0.002	0.870 ± 0.004
Group (G)	0.556 ± 0.024	0.551 ± 0.020	0.723 ± 0.032	0.505 ± 0.061
Temporal (T)	0.578	0.560	0.684	0.434
Group+Temporal (GT)	0.570 ± 0.008	0.550 ± 0.006	0.669 ± 0.006	0.414 ± 0.011

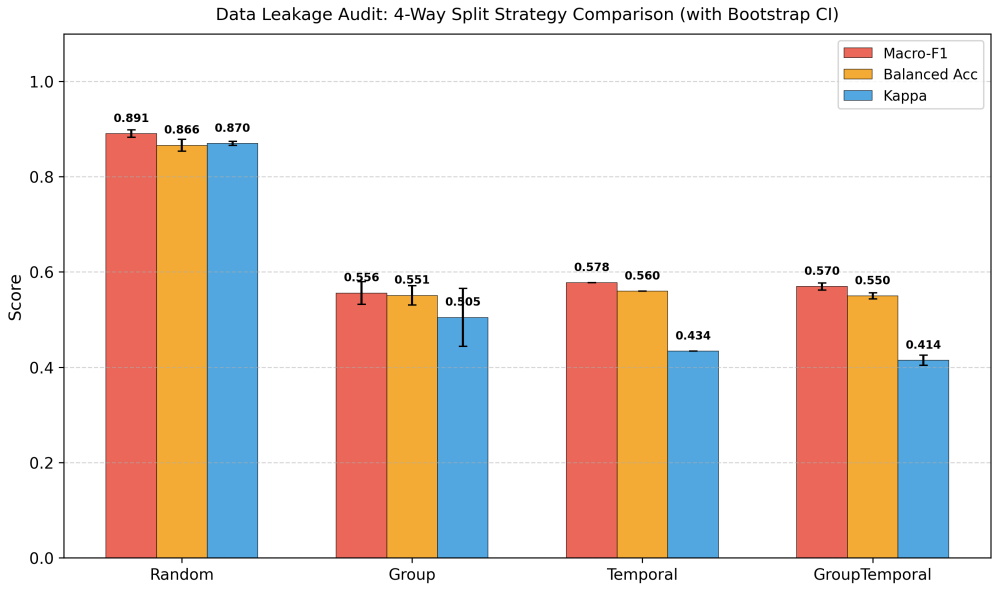

Figure 3. Macro-F1 across four evaluation protocols. Error bars denote ±1 SD over bootstrap seeds. The ∼0.32-point gap between Random and the three leakage-aware splits demonstrates the magnitude of entity-level inflation.

4.2.1. 2 × 2 Factorial Decomposition

The four splits form a natural 2 × 2 design (Table 6). Removing firm overlap alone (R → G) reduces Macro-F1 by 0.335; removing temporal continuity alone (R → T) reduces it by 0.313; removing both (R → GT) reduces it by 0.321. The near-identical floor across G, T, and GT (∼0.57) indicates that the two leakage sources are highly collinear: recalls from the same firm cluster in time, so eliminating either dimension of overlap suffices to expose the true generalisation level.

Table 6. 2 × 2 factorial decomposition of evaluation bias (Macro-F1).

	Same firms (overlap)	New firms (disjoint)
IID (random time)	0.891 (R)	0.556 (G)
Future (temporal)	0.578 (T)	0.570 (GT)

4.2.2. Per-Class Breakdown

Table 7 reveals that Class III suffers the most dramatic degradation: recall plummets from 0.714 under Random to 0.094 under Group splitting—a 87% relative decrease. This is expected given that Class III accounts for only 5.8% of records and is concentrated among a small number of firms, making it highly susceptible to entity memorisation.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2026 doi:10.20944/preprints202603.0343.v1

Table 7. Per-class recall across splitting strategies (XGBoost).

Class	Random	Group	Temporal	Group+Temp.
Class I	0.937	0.797	0.789	0.780
Class II	0.947	0.761	0.682	0.672
Class III	0.714	0.094	0.208	0.198

4.2.3. Hazard Type and Severity Class Association

To ground the preceding statistical findings in food-safety domain knowledge, Table 8 presents the prevalence of each hazard category across the three severity classes.

Table 8. Prevalence of hazard categories by severity class (% of records within each class flagged for the hazard type).

Hazard type	Class I	Class II	Class III
Pathogen	74.9%	17.3%	2.7%
Allergen	28.4%	37.2%	32.6%
Foreign material	0.3%	14.9%	8.2%
Chemical	1.5%	2.9%	6.5%
Labelling	0.9%	3.1%	8.2%

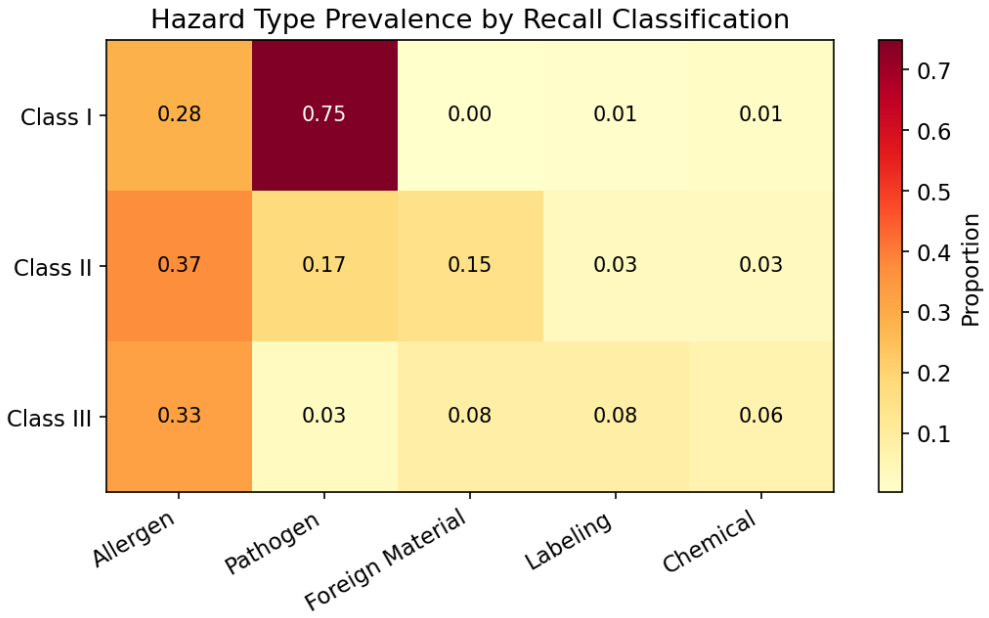

Figure 4. Hazard category prevalence by severity class. Pathogen-related recalls are overwhelmingly Class I, while labelling and chemical hazards are more prevalent in Class III.

Pathogen contamination (Salmonella, Listeria, E. coli) dominates Class I recalls (74.9%), consistent with FDA guidance that pathogen-related hazards pose the most serious health risk [29]. Undeclared allergens are distributed more evenly across classes, reflecting the heterogeneous severity of allergen-related incidents depending on the specific allergen and exposure context [44]. Foreign material, chemical contaminants, and labelling defects are relatively more common in Class II and III, where health consequences are typically less severe. This distribution aligns with the broader patterns documented in recent large-scale analyses of FDA recall records [45,46].

These patterns have direct implications for the leakage audit results: the pathogen → Class I mapping is largely hazard-intrinsic and therefore transferable across firms, whereas Class III recalls—

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2026 doi:10.20944/preprints202603.0343.v1

dominated by labelling and GMP violations—are highly firm-specific, explaining why Class III recall collapses from 0.714 to 0.094 under group splitting (Table 7).

4.3. Firm-Mode Baseline: Quantifying Entity Memorisation

To isolate the contribution of entity identity, we constructed a non-ML firm-mode baseline (Table 9). Under the Random split, this baseline achieves Macro-F1 = 0.818—only 0.075 below XGBoost's 0.893. This implies that 92% (0.818/0.893) of the apparent performance can be attributed to firm-level memorisation. Under the GT split, firm-mode collapses to 0.226, while XGBoost retains 0.564, yielding a genuine model increment of +0.339—far larger than the +0.075 observed under Random evaluation.

Table 9. Firm-mode baseline versus XGBoost across four splits (Macro-F1).

Split	Firm-Mode	XGBoost	Δ (Model increment)
Random (R)	0.818	0.893	+0.075
Group (G)	0.207	0.596	+0.389
Temporal (T)	0.306	0.578	+0.272
Group+Temporal (GT)	0.226	0.564	+0.339

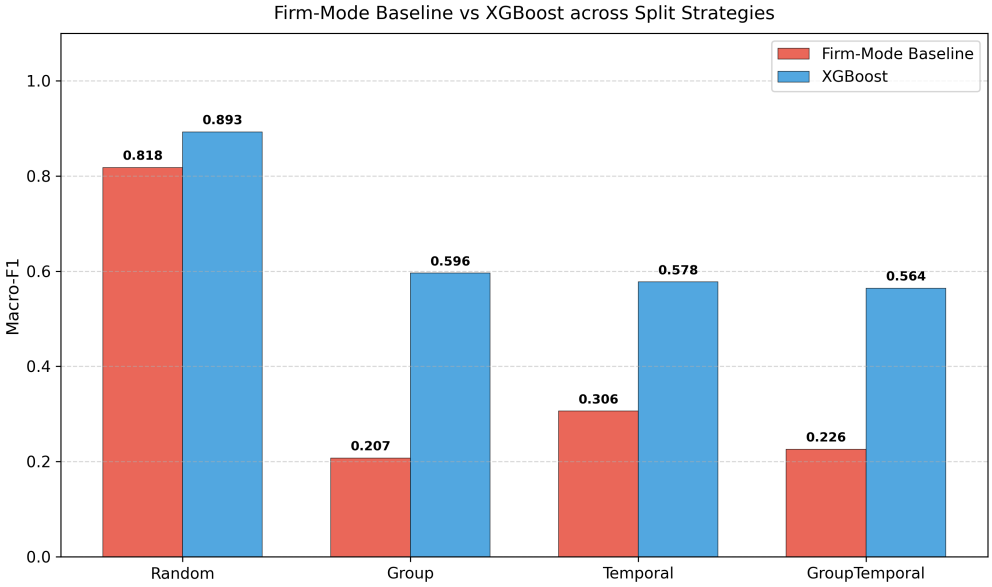

Figure 5. Firm-mode baseline versus XGBoost across four evaluation protocols. The narrowing gap under Random splitting versus the widening gap under stricter protocols reveals the true model increment.

This result has an important implication: *the stricter the evaluation, the more the model's genuine learning capacity is revealed*. Under Random evaluation, the model's true contribution is obscured by the ease of entity memorisation.

4.4. Identity Masking: Verifying the Leakage Pathway

Table 10 shows the effect of replacing firm names in the text with `[FIRM]` tokens before re-extracting TF-IDF features.

Table 10. Effect of identity masking on Macro-F1 (Δ = Masked − Original).

Split	Original	Masked	ΔF1
Random (R)	0.893	0.889	−0.004
Group (G)	0.596	0.596	+0.000
Temporal (T)	0.578	0.593	+0.016
Group+Temporal (GT)	0.564	0.583	+0.019

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2026 doi:10.20944/preprints202603.0343.v1

All deltas are at most 0.019, indicating that explicit company-name tokens are not the primary leakage channel. The leakage is *structural*: firms that repeatedly recall products share product types, hazard profiles, distribution patterns, and regulatory language that collectively create distinguishable signatures in the feature space. Under the GT split, masking slightly improves performance (+0.019), suggesting that company names act as noise for unseen firms.

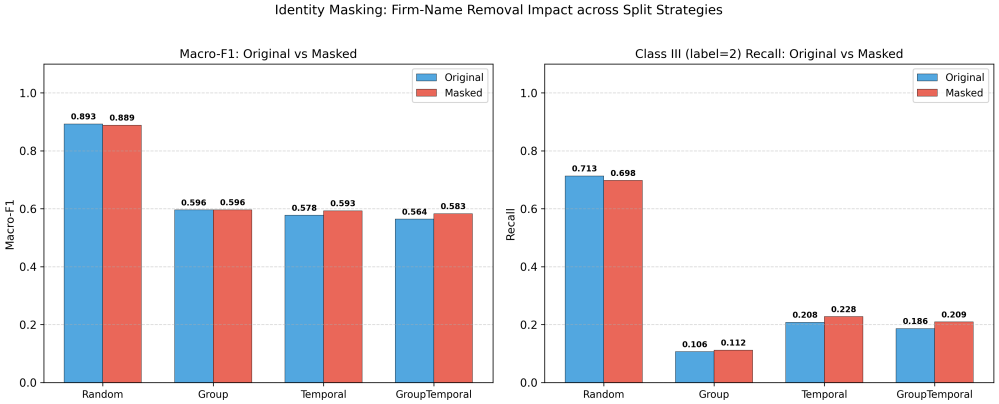

Figure 6. Identity-masking experiment: Macro-F1 comparison between original and masked text features across four splits.

4.5. SHAP Interpretability Analysis

SHAP TreeExplainer was applied to the XGBoost model trained under the Random split. Figure 7 presents the top-20 features by global mean absolute SHAP value. Hazard keywords (pathogen, allergen) and TF-IDF terms related to undeclared allergens and microbial contamination dominate, aligning with the regulatory logic that pathogen-related recalls are predominantly classified as Class I.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2026 doi:10.20944/preprints202603.0343.v1

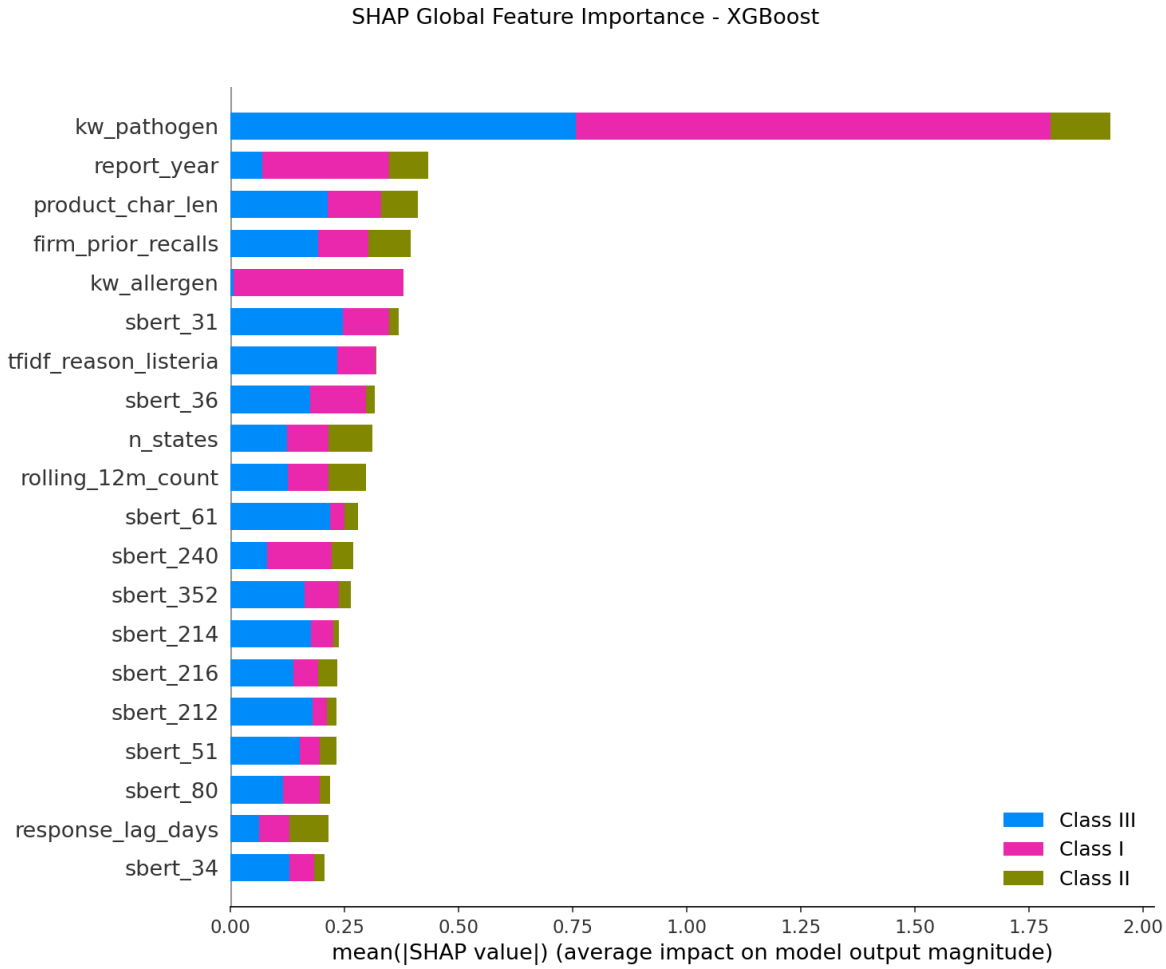

Figure 7. SHAP global feature importance: top 20 features by mean |SHAP| value.

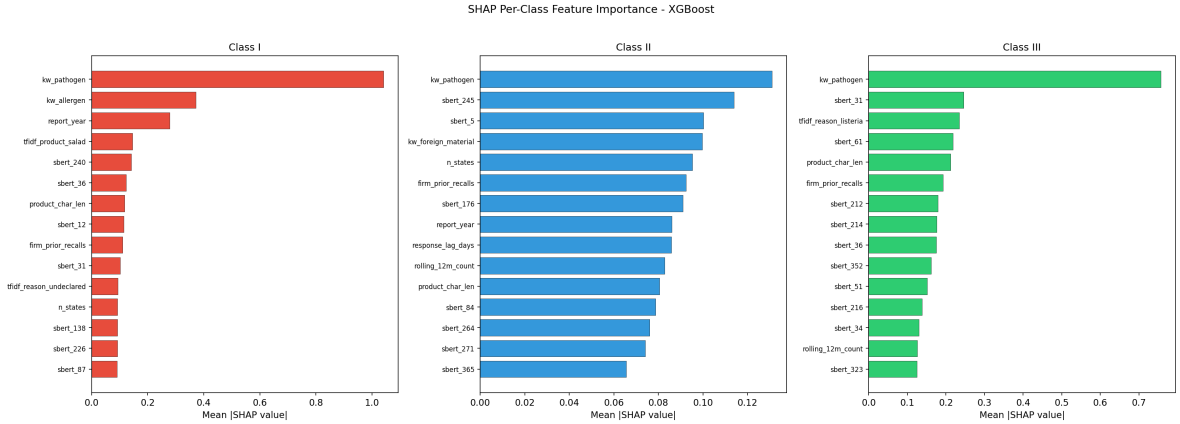

Figure 8. SHAP per-class feature importance for Class I, II, and III.

Per-class analysis (Figure 8) reveals distinct feature profiles: Class I is driven by pathogen and allergen indicators; Class II by labelling and manufacturing-process terms; and Class III shows weaker, more diffuse signals. Notably, `firm_prior_recalls` ranks among the global top-10 features, confirming that the model leverages firm history—a finding consistent with the firm-mode baseline results in Section 4.3.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2026 doi:10.20944/preprints202603.0343.v1

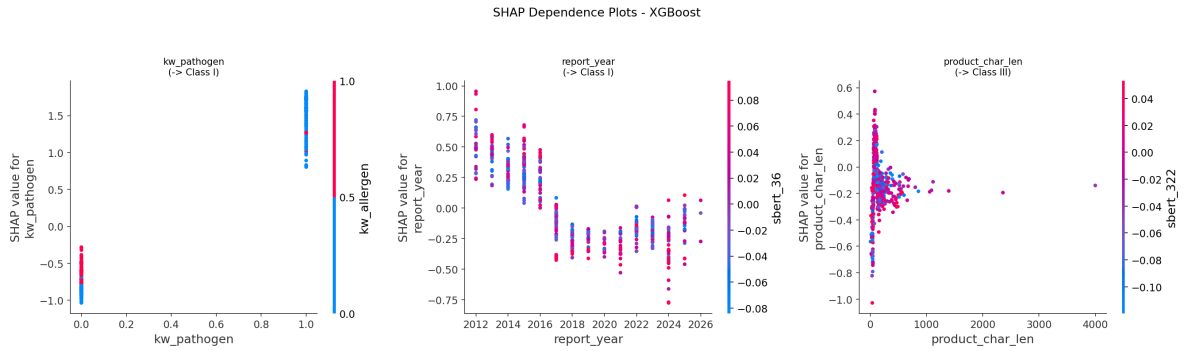

Figure 9. SHAP dependence plots for the top three features.

4.6. Feature Ablation

Table 11 reports ablation results under the Random split. Removing all NLP features reduces Macro-F1 by only 0.067, while NLP-only features achieve 0.890—nearly identical to the full model. Temporal features contribute negligibly ($\Delta = -0.002$). These results should be interpreted cautiously: under Random splitting, the structured features (40 dimensions) suffice to achieve 0.821, largely because they capture firm-level patterns that the random split fails to exclude.

Table 11. Feature ablation results under random split (XGBoost Macro-F1).

Configuration	Macro-F1	ΔF1
Full (1,437 d)	0.888	—
No NLP (45 d)	0.821	−0.067
NLP Only (1,392 d)	0.889	+0.002
No Temporal (1,432 d)	0.886	−0.002
No SBERT (1,053 d)	0.883	−0.005
No TF-IDF (437 d)	0.885	−0.003

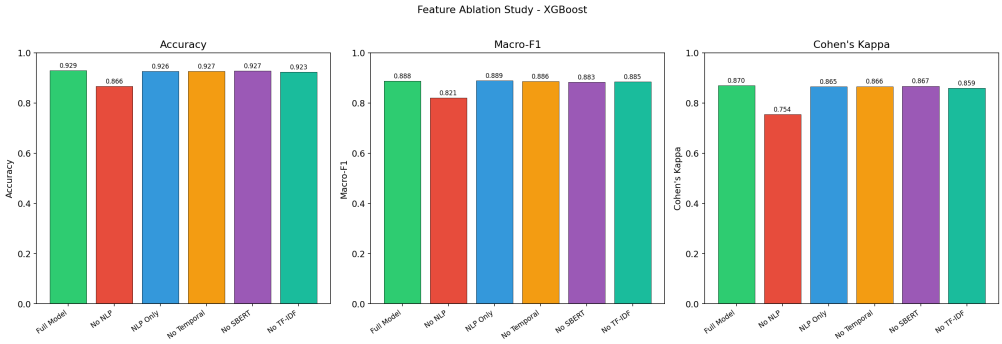

Figure 10. Feature ablation comparison (Macro-F1 under random split).

4.7. Continuous Learning Simulation

Figure 11 presents the Macro-F1 trajectory across nine annual windows for three retraining strategies. Expanding-window training (mean F1 = 0.578) consistently outperforms both static (0.517) and sliding (0.547) approaches. The advantage of expanding over sliding suggests that data volume matters more than temporal recency, likely because the rare Class III requires a large data pool to achieve adequate representation. This is consistent with the leakage audit's finding that temporal concept drift is modest relative to entity-level effects.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2026 doi:10.20944/preprints202603.0343.v1

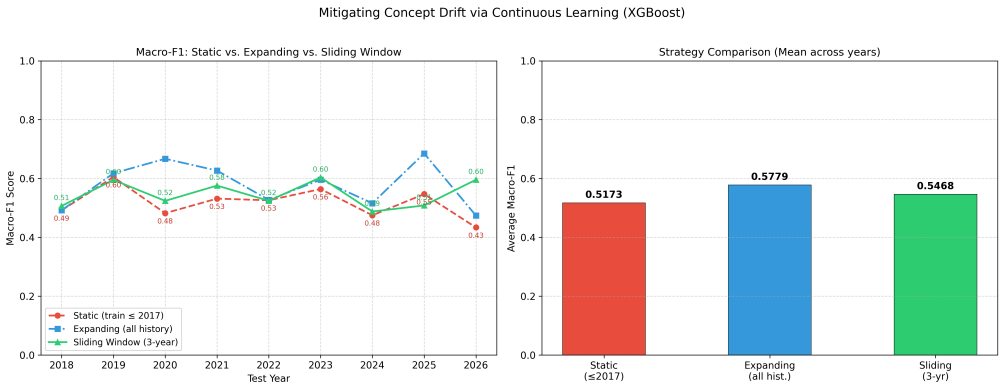

Figure 11. Continuous learning simulation: annual Macro-F1 for three retraining strategies (2018–2026).

A further challenge not captured by this simulation is the emergence of novel hazard types absent from historical training data—such as the muscimol-containing mushroom products and lead-contaminated cinnamon incidents that prompted major recalls in 2024 [46]. Such distributional shifts represent a harder form of concept drift than the gradual temporal trends modelled here, and suggest that operational deployment would benefit from anomaly detection or human-in-the-loop mechanisms for out-of-distribution hazards.

5. Discussion

5.1. Significance of Core Findings

The central finding of this study is that standard random splitting inflates Macro-F1 by approximately 0.32 points on the FDA food recall severity prediction task. This inflation is not a modelling artefact but a structural property of regulatory databases in which repeated entities (firms) contribute clusters of similar records.

Our results extend and systematise the early observation by Bouzembrak and Marvin [22], who reported an accuracy drop from 80% to 52% for unseen country–product combinations in RASFF food fraud prediction. While their finding remained incidental—noted but not decomposed—our study (1) applies the same logic to a different regulatory database (FDA), (2) decomposes the performance inflation into firm-level autocorrelation and temporal concept drift via a 2×2 factorial design, and (3) demonstrates their high collinearity, showing that removing either source of overlap suffices to reveal the true generalisation floor. Our 2×2 factorial decomposition empirically confirms, in a food safety context, the theoretical predictions of Cerqua et al. [21]: neglecting the panel structure of regulatory databases leads to substantial performance inflation.

This finding has direct implications for the interpretation of existing high-performing systems on food-safety databases. Nogales et al. [13] reported 72–89% accuracy on RASFF, and Sari et al. [14] achieved 97.8% accuracy, both using standard random splitting. We do not claim that these results are invalid, but our leakage audit demonstrates that entity-level autocorrelation—a structural property shared across regulatory recall databases—can substantially inflate reported performance. We recommend that these results be re-evaluated under group-aware or temporal protocols to establish reliable generalisation baselines.

The parallel with other domains reinforces this point. Patient-level data segregation is now considered essential in medical image analysis [23], subject-level leakage has been shown to drastically inflate prediction performance in connectome-based neuroimaging [47], and leave-source-out designs reduce apparent performance in ECG classification [24]. Kapoor and Narayanan [25] and Cerqua et al. [21] documented that data leakage affects reproducibility across multiple scientific domains. Our study extends this evidence to food safety—a domain where the issue has received almost no systematic attention.

The practical consequences of evaluation bias extend beyond academic metrics. If a model trained under entity-leaked conditions misclassifies a Class I event as Class III, the resulting delay in public

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2026 doi:10.20944/preprints202603.0343.v1

notification and resource deployment could amplify the economic and health toll. Individual Class I recalls in the low-moisture food sector have been associated with median market capitalisation losses of $243 million [7], and supply-chain disruption from a single contamination event can persist for months [8]. International evidence further confirms that food recall announcements trigger significant investor penalties [48]. Against a national backdrop where foodborne illness costs an estimated $75 billion annually [4] and federal oversight remains fragmented [49], the distinction between genuine and inflated model performance is not merely methodological but consequential for public health resource allocation.

5.2. The True Value of ML Models

The Group+Temporal Macro-F1 of ∼0.57 should not be viewed with undue pessimism. Under this strictest evaluation, XGBoost still exceeds the firm-mode baseline by 0.339 points (Table 9), indicating that the model captures genuine patterns beyond entity identity—including hazard–severity associations, product-category signals, and temporal trends.

The binary formulation (Class I versus rest) achieved AUC = 0.988 under random evaluation, suggesting strong practical utility for a simpler "is this recall severe?" triage decision. However, this binary result requires re-validation under group-aware protocols before operational conclusions can be drawn.

The ablation experiment (Section 4.6) reveals that structured features alone (45 dimensions) achieve Macro-F1 = 0.821 under the Random split, while removing NLP features causes only a 0.067-point reduction. This finding should be interpreted carefully: under random evaluation, structured features largely capture firm-level patterns. The more informative question—how each feature group performs under GT evaluation—is left for future work, but would clarify which features drive genuine cross-firm generalisation versus entity memorisation.

The hazard-type analysis (Table 8) provides a domain-grounded explanation for these patterns. The pathogen → Class I association is a hazard-intrinsic regularity: regardless of the producing firm, Salmonella or Listeria contamination poses a severe health risk and is classified accordingly [29]. This mapping is therefore relatively transferable across firms, explaining why Class I recall degrades least under group splitting (from 0.937 to 0.797). By contrast, undeclared-allergen recalls span all three severity classes depending on the specific allergen, its concentration, and whether it triggered adverse events [44], making allergen severity more context-dependent and harder to predict for unseen firms. Class III recalls—predominantly labelling defects and GMP violations—are the most firm-specific: each company's compliance failures reflect its unique production processes, facility conditions, and quality management systems. This explains the catastrophic Class III recall collapse (0.714 → 0.094) under group splitting, as the model can no longer rely on firm-specific signatures to identify these low-severity patterns.

For practical deployment, our results suggest two actionable insights. First, the Expanding retraining strategy outperforms both Static and Sliding approaches in the continuous learning simulation, indicating that historical data volume matters more than recency for rare-class prediction. Second, at early stages of a recall event—when detailed narrative text may not yet be available—structured features alone provide a viable initial risk signal, though cross-firm generalisation remains the fundamental bottleneck. From a food-science perspective, the firm-level memorisation phenomenon reflects a structural reality of food production: a firm's recall profile is largely determined by its product portfolio, manufacturing processes, facility sanitation regime, and supply-chain structure [30]. A dairy processor with persistent Listeria environmental contamination will generate repeated Class I recalls, while a confectionery firm with labelling compliance gaps will produce recurring Class III events. The model's ability to "memorise" these patterns is therefore not merely a statistical artefact but a reflection of genuine production-system determinism—one that, however, does not generalise to firms outside the training set. These findings align with the FDA's strategic vision of leveraging predictive analytics for faster, more targeted food safety interventions [9].

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2026 doi:10.20944/preprints202603.0343.v1

5.3. Evaluation Protocol Recommendations

Based on our findings, we propose three concrete recommendations for food-safety ML research. First, studies should report results under at least two evaluation protocols: a standard random split (for comparability with existing literature) and a group-aware or temporal split (for generalisation validity); the gap between protocols is itself an informative diagnostic. Second, authors should disclose the degree of entity overlap between training and test sets (e.g., "X out of Y firms appear in both train and test"). This single statistic helps readers assess the risk of entity-level inflation without requiring full replication. Third, a simple non-ML entity-prior baseline—assigning each entity its historically most frequent class—should be included to provide a lower bound for the contribution of entity memorisation. If a complex model does not substantially exceed this baseline under random evaluation, entity-level leakage is likely driving the results. These recommendations are neither expensive nor disruptive: group-aware splitting requires only an entity identifier (universally available in regulatory databases), and the entity-prior baseline can be computed in a few lines of code. We further recommend that models deployed in regulatory settings be accompanied by post-hoc interpretability analyses. As the broader XAI literature emphasises [38,41], transparency in model predictions is essential for regulatory trust—particularly in food safety, where SHAP-based explanations have already been shown to provide actionable insights [14,40].

5.4. Limitations

We acknowledge several limitations. First, feature ablation was conducted solely under the Random protocol; the relative importance of feature groups may differ under group-aware evaluation, where entity-specific signals are unavailable. Second, the `reason_for_recall` field is written after the severity classification is known and may embed classification rationale, making the task partly retrospective. However, removing all NLP features still yields Macro-F1 = 0.821 under Random splitting, suggesting that structured features carry independent predictive value. Third, only XGBoost was used for the multi-split audit. Given that the three boosting models differed by less than 0.003 under Random evaluation, we consider XGBoost representative, but cross-model validation would strengthen the conclusions. Fourth, text masking replaced only firm names; brand names, addresses, and other indirect identifiers were not masked, and more aggressive anonymisation might further reduce residual entity leakage. Fifth, Sentence-BERT embeddings were generated with the general-purpose `all-MiniLM-L6-v2` model without domain-specific fine-tuning, which may underestimate the potential of semantic features for food-safety text.

6. Conclusions

This study constructed the first comprehensive machine learning benchmark for FDA food recall severity classification (Class I/II/III), comprising 28,448 openFDA enforcement records (2012–2025), a 1,437-dimensional feature space combining TF-IDF, Sentence-BERT, structured, and temporal features, and five tuned classifiers evaluated under four distinct splitting protocols.

Our multi-layer leakage audit reveals that the standard random-split Macro-F1 of 0.89 is inflated by approximately 0.32 points due to entity-level autocorrelation. Under group-aware, temporal, and combined splitting, true generalisation performance converges to approximately 0.57. A firm-mode baseline achieves 0.82 under random evaluation, demonstrating that 92% of apparent performance stems from entity memorisation. A 2×2 factorial decomposition shows that firm overlap and temporal continuity are highly collinear—removing either suffices to expose the true performance floor. Identity-masking experiments confirm that the leakage is structural rather than attributable to explicit name tokens. Despite these sobering findings, XGBoost still exceeds the firm-mode baseline by 0.339 under the strictest evaluation, indicating genuine learned patterns beyond entity identity.

We recommend that food-safety ML studies (1) report results under both random and group-aware or temporal evaluation protocols, (2) disclose entity-overlap statistics between training and test sets, and (3) include entity-prior baselines to contextualise model performance. These measures

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2026 doi:10.20944/preprints202603.0343.v1

are computationally inexpensive and can be adopted with minimal disruption to existing research workflows. Future work should extend the leakage audit to other food-safety databases (e.g., EU RASFF), conduct feature ablation under group-aware splits, and explore domain-adapted language models for food-safety text classification.

Author Contributions: Conceptualisation, P.L. and J.-S.T.; methodology, P.L. and J.-S.T.; software, P.L. and J.-S.T.; validation, P.L. and J.-S.T.; formal analysis, P.L. and J.-S.T.; investigation, P.L. and J.-S.T.; data curation, P.L. and J.-S.T.; writing—original draft preparation, P.L. and J.-S.T.; writing—review and editing, P.L. and J.-S.T.; visualisation, P.L. and J.-S.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data used in this study were retrieved from the openFDA Food Enforcement API (https://open.fda.gov/apis/food/enforcement/), which is publicly accessible. The analysis code is available from the corresponding author upon reasonable request.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

FDA	U.S. Food and Drug Administration
ML	Machine Learning
NLP	Natural Language Processing
SHAP	SHapley Additive exPlanations
SBERT	Sentence-BERT
TF-IDF	Term Frequency–Inverse Document Frequency
RASFF	Rapid Alert System for Food and Feed
AUC	Area Under the Curve
ROC	Receiver Operating Characteristic
TPE	Tree-structured Parzen Estimator

References

1. U.S. Food and Drug Administration. Regulatory Procedures Manual, Chapter 7: Recall Procedures. https://www.fda.gov/media/71776/download, 2024.
2. U.S. Food and Drug Administration. openFDA Food Enforcement API. https://open.fda.gov/apis/food/enforcement/, 2025. Accessed: 2025-02-25.
3. Scallan Walter, E.J.; et al. Foodborne Illness Acquired in the United States—Major Pathogens, 2019. *Emerging Infectious Diseases* **2025**, *31*. https://doi.org/10.3201/eid3104.240913.
4. Hoffmann, S.; et al. Economic Burden of Foodborne Illnesses Acquired in the United States. *Foodborne Pathogens and Disease* **2025**, *22*, 4–14. https://doi.org/10.1089/fpd.2023.0157.
5. Thomsen, M.R.; McKenzie, A.M. Market Incentives for Safe Foods: An Examination of Shareholder Losses from Meat and Poultry Recalls. *American Journal of Agricultural Economics* **2001**, *83*, 526–538. https://doi.org/10.1111/0002-9092.00175.
6. Pozo, V.F.; Schroeder, T.C. Evaluating the costs of meat and poultry recalls to food firms using stock returns. *Food Policy* **2016**, *59*, 66–77. https://doi.org/10.1016/j.foodpol.2015.12.007.
7. Sahin, O.; Coronado, O.M.; Bao, Y. Monetizing the Impact of Food Safety Recalls on the Low-Moisture Food Industry. *Journal of Food Protection* **2020**, *83*, 1130–1137. https://doi.org/10.4315/JFP-19-485.
8. Spalding, A.; et al. Economic Impacts of Food Safety Incidents in a Modern Supply Chain: E. coli in the Romaine Lettuce Industry. *American Journal of Agricultural Economics* **2023**, *105*, 1087–1112. https://doi.org/10.1111/ajae.12341.
9. U.S. Food and Drug Administration. New Era of Smarter Food Safety Blueprint. https://www.fda.gov/food/new-era-smarter-food-safety/new-era-smarter-food-safety-blueprint, 2020.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2026 doi:10.20944/preprints202603.0343.v1

10. U.S. Food and Drug Administration. Current Good Manufacturing Practice, Hazard Analysis, and Risk-Based Preventive Controls for Human Food. 21 CFR Part 117, 2015. FSMA Final Rule.
11. Deng, X.; Cao, S.; Horn, A.L. Emerging Applications of Machine Learning in Food Safety. *Annual Review of Food Science and Technology* **2021**, *12*, 513–538. https://doi.org/10.1146/annurev-food-071720-024112.
12. Wang, X.; Bouzembrak, Y.; Lansink, A.G.J.M.O.; van der Fels-Klerx, H.J. Application of machine learning to the monitoring and prediction of food safety: A review. *Comprehensive Reviews in Food Science and Food Safety* **2022**, *21*, 416–434. https://doi.org/10.1111/1541-4337.12868.
13. Nogales, A.; Daz-Morón, R.; García-Tejedor, Á.J. A comparison of neural and non-neural machine learning models for food safety risk prediction with European RASFF data. *Food Control* **2022**, *134*, 108697. https://doi.org/10.1016/j.foodcont.2021.108697.
14. Sari, Y.; et al. AI-driven risk assessment in food safety: a Transformer-based approach with explainable AI using the EU RASFF database. *Food and Bioprocess Technology* **2025**. https://doi.org/10.1007/s11947-025-03819-4.
15. Papadopoulos, T.; et al. A Machine Learning Approach for Food Safety Risk Prediction Using RASFF Data, 2020. arXiv:2009.06704.
16. Randl, K.R.; et al. SemEval-2025 Task 9: The Food Hazard Detection Challenge. In Proceedings of the Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025), 2025. arXiv:2503.19800.
17. Grinsztajn, L.; Oyallon, E.; Varoquaux, G. Why do tree-based models still outperform deep learning on typical tabular data? In Proceedings of the Advances in Neural Information Processing Systems, 2022, Vol. 35. Datasets and Benchmarks Track.
18. Shwartz-Ziv, R.; Armon, A. Tabular data: Deep learning is not all you need. *Information Fusion* **2022**, *81*, 84–90. https://doi.org/10.1016/j.inffus.2021.11.011.
19. Anonymous. Analyzing FDA Food Recall Patterns Using Machine Learning. ResearchGate preprint, 2024. Available at: https://www.researchgate.net/publication/391367432.
20. Mulla, S.; Patel, N. A comprehensive analysis of FDA drug recall data using text analytics and machine learning. *AI & Society* **2025**. https://doi.org/10.1007/s00146-025-02598-y.
21. Cerqua, A.; Ferrante, M.; Letta, M.; Ferrante, L.; Pinto, G. Panel Data and Cross-Validation: The Pitfalls of Standard ML Methods. *Journal of Econometrics* **2025**. Forthcoming.
22. Bouzembrak, Y.; Marvin, H.J.P. Prediction of food fraud type using data from the 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.
23. Bussola, N.; Marini, S.; Maggio, V.; Jurman, G.; Furlanello, C. A weakly supervised approach for estimating spatial resolution in whole-slide histopathological images. *Scientific Reports* **2021**. Patient-level data segregation in medical ML.
24. Strodthoff, N.; Wagner, P.; Schaeffter, T.; Samek, W. Deep learning for ECG analysis: benchmarks and insights from PTB-XL. *IEEE Journal of Biomedical and Health Informatics* **2021**, *25*, 1519–1528. https://doi.org/10.1109/JBHI.2020.3022989.
25. Kapoor, S.; Narayanan, A. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns* **2023**, *4*, 100804. https://doi.org/10.1016/j.patter.2023.100804.
26. Sari, Y.; et al. A systematic comparison of machine learning approaches for food safety risk classification using RASFF data. *British Food Journal* **2025**.
27. Wu, Y.; Liu, X.; Li, A.; Shi, H. Application of Machine Learning in Food Safety Risk Assessment: A Review. *Foods* **2024**, *13*, 4144. https://doi.org/10.3390/foods13244144.
28. Reimers, N.; Gurevych, I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In Proceedings of the Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2019. https://doi.org/10.18653/v1/D19-1410.
29. Batz, M.B.; Hoffmann, S.; Morris, Jr., J.G. Ranking the Disease Burden of 14 Pathogens in Food Sources in the United States Using Attribution Data from Outbreak Investigations and Expert Elicitation. *Journal of Food Protection* **2012**, *75*, 1278–1291. https://doi.org/10.4315/0362-028X.JFP-11-418.
30. Ivanek, R.; Gröhn, Y.T.; Wiedmann, M. Listeria monocytogenes in Multiple Habitats and Host Populations: Review of Available Data for Mathematical Modeling. *Foodborne Pathogens and Disease* **2006**, *3*, 319–336. https://doi.org/10.1089/fpd.2006.3.319.
31. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 785–794. https://doi.org/10.1145/2939672.2939785.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2026 doi:10.20944/preprints202603.0343.v1

32. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.Y. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In Proceedings of the Advances in Neural Information Processing Systems, 2017, Vol. 30.
33. Prokhorenkova, L.; Gusev, G.; Vorobev, A.; Dorogush, A.V.; Gulin, A. CatBoost: unbiased boosting with categorical features. *Advances in Neural Information Processing Systems* **2018**, *31*.
34. Borisov, V.; Leemann, T.; Seßler, K.; Haug, J.; Pawelczyk, M.; Kasneci, G. Deep Neural Networks and Tabular Data: A Survey. *IEEE Transactions on Neural Networks and Learning Systems* **2022**, *35*, 7499–7519. https://doi.org/10.1109/TNNLS.2022.3229161.
35. McElfresh, D.; et al. When Do Neural Nets Outperform Boosted Trees on Tabular Data? In Proceedings of the Advances in Neural Information Processing Systems, 2024, Vol. 37.
36. Bentéjac, C.; Császárik, A.; Berta, M. Benchmarking state-of-the-art gradient boosting algorithms for classification, 2023. arXiv:2305.17094.
37. Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; Koyama, M. Optuna: A Next-generation Hyperparameter Optimization Framework. In Proceedings of the Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2019, pp. 2623–2631. https://doi.org/10.1145/3292500.3330701.
38. Barredo Arrieta, A.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; Garcia, S.; Gil-Lopez, S.; Molina, D.; Benjamins, R.; 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.
39. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* **2019**, *1*, 206–215. https://doi.org/10.1038/s42256-019-0048-x.
40. Büyüktepe, O.; et al. Food fraud detection using explainable artificial intelligence. *Expert Systems* **2025**, p. e13387. https://doi.org/10.1111/exsy.13387.
41. Noor, F.; et al. Unveiling Explainable AI in Healthcare: Current Trends, Challenges, and Future Directions. *WIREs Data Mining and Knowledge Discovery* **2025**, p. e70018. https://doi.org/10.1002/widm.70018.
42. Lundberg, S.M.; Lee, S.I. A Unified Approach to Interpreting Model Predictions. In Proceedings of the Advances in Neural Information Processing Systems, 2017, Vol. 30.
43. Lu, J.; Liu, A.; Dong, F.; Gu, F.; Gama, J.; Zhang, G. Learning under Concept Drift: A Review. *IEEE Transactions on Knowledge and Data Engineering* **2019**, *31*, 2346–2363. https://doi.org/10.1109/TKDE.2018.2876857.
44. Gendel, S.M. Comparison of international food allergen labeling regulations. *Food Additives & Contaminants: Part A* **2012**, *29*, 95–103. https://doi.org/10.1080/19440049.2011.620198.
45. Kendall, H.; et al. An Analysis of Food Recalls in the United States, 2002–2023. *Journal of Food Protection* **2024**. https://doi.org/10.1016/j.jfp.2024.100404.
46. U.S. PIRG Education Fund. Food for Thought 2025: Lessons from Food Recalls in 2024. https://pirg.org/resources/food-for-thought-2025/, 2025.
47. Rosenblatt, M.; Tejavibulya, L.; Jiang, R.; Noble, S.; Scheinost, D. Data leakage inflates prediction performance in connectome-based machine learning models. *Nature Communications* **2024**, *15*, 1829. https://doi.org/10.1038/s41467-024-46150-w.
48. Zhao, X.; Li, Y.; Flynn, B. The financial impact of product recall announcements in China. *International Journal of Production Economics* **2013**, *142*, 115–123. https://doi.org/10.1016/j.ijpe.2012.10.007.
49. U.S. Government Accountability Office. Food Safety: Status of Foodborne Illness in the U.S. GAO-25-107606, 2025. Available at: https://www.gao.gov/products/gao-25-107606.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.