Early Detection of High-Risk Product Recalls:

A Comparative Study of Multi-Class

Classification Approaches

by Lorena Dorado and Parisa Kamizi

Abstract

Timely identification and classification of product recalls are essential to safeguarding public health. This study explores the application of machine learning and natural language processing (NLP) techniques to predict the severity of product recalls issued by the U.S. Food and Drug Administration (FDA). Using a dataset of over 95,000 FDA recall records, the study developed a multi-class classification system that categorizes recalls into Class I, II, or III based on structured features and textual recall descriptions. Feature engineering incorporated temporal patterns, categorical variables, and text-based features such as TF-IDF and word counts.

Several classification models—including Random Forest, XGBoost, Decision Tree, Multilayer Perceptron, and Logistic Regression—were evaluated using precision, recall, and F1-score metrics. The Random Forest model achieved the best overall performance with an F1-score above 0.93. While the model effectively distinguished Class I and II recalls, Class III predictions proved more complex due to overlapping features. A Streamlit dashboard was deployed to demonstrate real-time classification capability. The findings highlight the potential for AI-driven tools to enhance regulatory decision-making, improve recall timeliness, and strengthen consumer protection.

Table of Contents

Business Background

Product recalls, essential for consumer safety, are managed by the FDA across sectors like food, drugs, and medical devices, classified by risk severity:

- Class I (life-threatening)
- Class II (moderate)
- Class III (low risk)

The recall process begins when a manufacturer reports a defect, and the FDA assesses the risk, determines the recall scope, and monitors its effectiveness. However, delays can occur as manufacturers voluntarily notify the FDA. Despite varied supply chains, FDA records provide opportunities for developing predictive models that apply across industries.

With growing recall complexity and increasing regulatory demands, the FDA and manufacturers need smarter tools to prioritize interventions and improve public safety.

Problem Statement

The current recall classification process is manual, potentially inconsistent, and often delayed. This increases public health risks and limits proactive decision-making. There is a need for a scalable, data-driven system that can:

- Classify recalls based on severity before official designation

- Support regulators with timely, objective insights

- Help manufacturers identify emerging risks early

Summary of The Findings

The FDA recall dataset analysis showed valuable insights for changing from a manual and slower classification process to a faster, data-driven system. Among the 95,082 records across 21 variables, a notable class imbalance was observed, which was handled with resampling techniques. Recall volumes stabilized after 2020, providing a reliable foundation for predictive modeling. Product type and geographic distribution were strong predictors. Visualizations like histograms and bar plots helped show the distribution of categorical variables such as "Product Type" and "Distribution Pattern." The "Product Type" variable showed that devices represented 37.3% of recalls, followed by food/cosmetics at 28.9%, and drugs at 17.6%.

Comparative model evaluations showed that ensemble methods like random forest outperformed logistic regression in identifying high-risk recalls. Unlike competitor models, which are limited to specific industries such as food or medical devices, our approach looks across different types of products, helping discover broader patterns. The results show that using machine learning could make recall decisions faster and more accurately, helping protect people from potentially dangerous products.

Figure 1

Distribution of Product Types

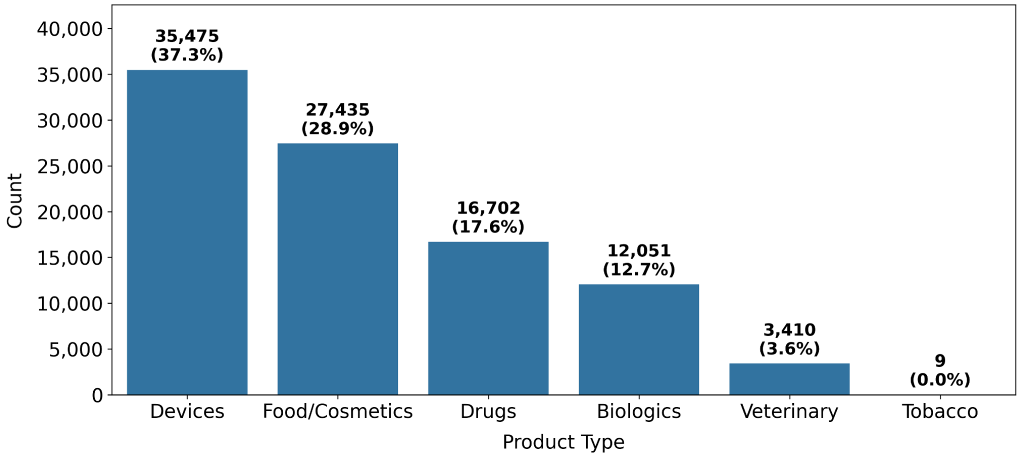
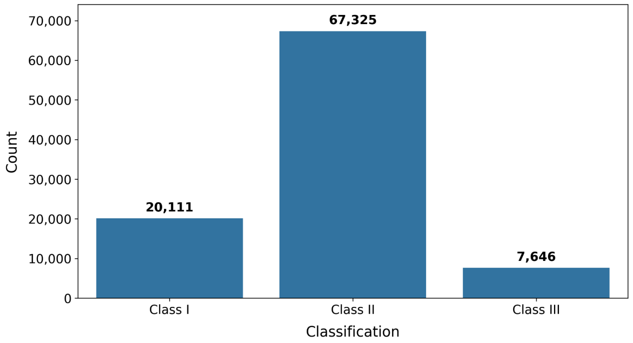


Figure 2

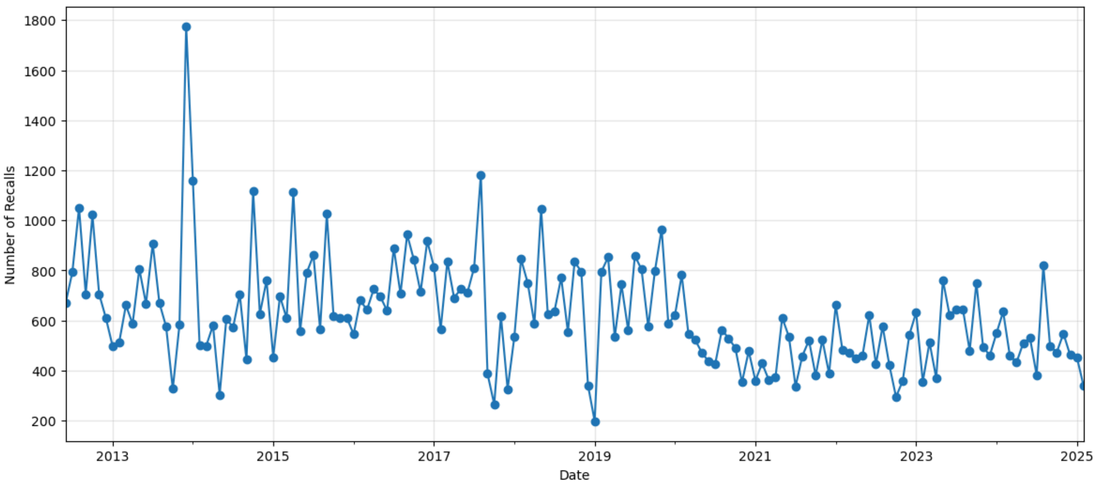
Distribution of Event Classification



Note. This figure shows the relative frequency of Class I, II, and III recalls.

Figure 3

Recall Trend Over Time



Business Questions

Can we reliably predict the severity of a recall event before FDA assignment?
Which features (e.g., product type, recall reason, distribution pattern) best predict high-risk recalls?
Can such a system enhance operational decision-making for manufacturers and regulators?
How might this reduce time-to-classification and public exposure?

Scope of Analysis

This analysis focuses on developing a multi-class classification and risk prediction system for FDA product recalls using recall enforcement data from 2020–2024, covering approximately 95,082 records. The model will use key features such as product type, recall reason (text), distribution, geography, and dates to classify recalls into Class I, II, or III categories. Text-based features will be processed using natural language processing techniques to extract patterns from recall reasons and product descriptions, while categorical data will help support classification accuracy.

The analysis does not include real-time recall updates, internal manufacturer data, or post-recall outcomes due to data access limitations. Instead, the project centers on historical records and publicly available information to evaluate model performance and feature impact. By combining structured and unstructured data, this approach seeks to enhance the accuracy and efficiency of recall classification and contribute insights into seasonal trends, risk scoring, and manufacturer-level risk profiling.

Approach

Data from the U.S. Food and Drug Administration (FDA) recall database was consolidated and preprocessed for classification analysis. The primary dataset contained over 95,000 product recall records, encompassing structured attributes and free-text descriptions. Structured features included categorical variables and temporal patterns, while unstructured text data was processed using NLP techniques such as TF-IDF vectorization and word count extraction. The goal was to build a multi-class classification model to predict recall severity levels (Class I, II, or III).

To prepare the dataset, missing values were addressed, relevant features were engineered, and textual descriptions were transformed into numerical features for model input. A 75%-25% train-test split was applied, and standardization was conducted where appropriate to ensure model compatibility. Multiple classification models were evaluated using five-fold cross-validation, including Random Forest, Decision Tree, XGBoost, Multilayer Perceptron (MLP), and Logistic Regression. Hyperparameters for each model were tuned using grid search optimization based on F1-weighted scores.

Table 1

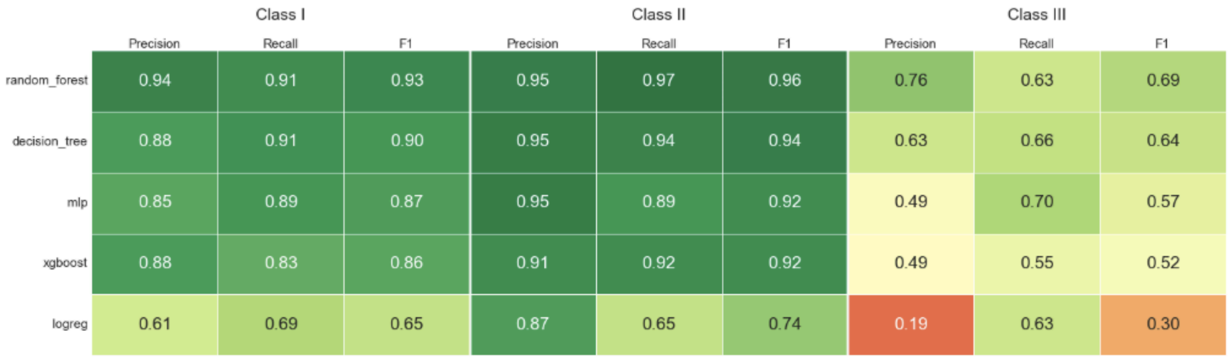
Hyperparameters Tuned for Each Classification Model

Model	Hyperparameter	Values Tested
Logistic Regression	C	0.01, 0.1, 1, 10, 100
	penalty	L1, L2
Decision Tree	max_depth	5, 10, 15, 20, None
	min_samples_split	2, 5, 10
	min_samples_leaf	1, 2, 4
Random Forest	n_estimators	100, 200
	max_depth	10, 20, None
	min_samples_split	2, 5
	min_samples_leaf	1, 2
XGBoost	n_estimators	100, 200
	learning_rate	0.01, 0.1
	max_depth	3, 6
	subsample	0.8, 1.0
	colsample_bytree	0.8, 1.0
MLPClassifier	hidden_layer_sizes	(50,), (100,), (50, 50)
	activation	relu, tanh
	alpha	0.0001, 0.001
	learning_rate	constant, adaptive

Note. All models were evaluated using 5-fold cross-validation with stratified sampling to maintain class distribution across folds. These hyperparameter configurations were optimized for each classification model using grid search, ensuring enhanced performance through cross-validation.

Figure 5

Heatmap of All Metrics Across All Classes and Models During Validation



	Class I			Class II			Class III		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
random_forest	0.94	0.91	0.93	0.95	0.97	0.96	0.76	0.63	0.69
decision_tree	0.88	0.91	0.90	0.95	0.94	0.94	0.63	0.66	0.64
mlp	0.85	0.89	0.87	0.95	0.89	0.92	0.49	0.70	0.57
xgboost	0.88	0.83	0.86	0.91	0.92	0.92	0.49	0.55	0.52
logreg	0.61	0.69	0.65	0.87	0.65	0.74	0.19	0.63	0.30

Note. Performance metrics for five machine learning models across three classes show random forest achieving the highest overall scores while logistic regression demonstrates the poorest performance, particularly for Class III.

Limitations

Computational limitations constrained this study during model training. Although we mitigated this by saving trained models for quick loading in the Streamlit dashboard, the initial model development process was time-consuming. These limitations stem from using local computing resources, and future work could benefit from enhanced hardware capabilities or more efficient modeling pipelines to reduce training time.

Another key limitation was the challenge in automating the extraction of data from the FDA website. Due to restrictions on high-volume web requests, we were unable to scrape thousands of FDA links to determine whether recalls were initiated by manufacturers or resulted from inspections. This limitation restricted our ability to fully capture the timeline of recall detection and understand the duration that products remained in the market, potentially affecting the accuracy of recall classification.

Despite these constraints, this study highlights the value of using multiple performance metrics to evaluate classification models. The joint use of precision and recall is particularly important in high-stakes domains like FDA recall analysis, where accurate detection of risks and maintaining regulatory credibility are both essential. Future work may involve patient-level or case-level analysis, provided more granular data becomes available.

Solution Details

The Random Forest model achieved the best performance overall, with a cross-validated F1 score of 0.9215 and a test set F1 score of 0.9308. It consistently outperformed other models across key evaluation metrics, including precision and recall. Decision Tree and MLP models followed closely, while Logistic Regression demonstrated significantly lower performance, especially in recall. Performance variation by class showed that Random Forest excelled at identifying both Class I and II recalls, while Class III proved more challenging due to overlapping feature characteristics.

A Streamlit dashboard was deployed to visualize model predictions in real time. Overall, the results demonstrate the value of tree-based machine learning models and NLP in accurately classifying recall severity, with practical implications for enhancing public health response and regulatory oversight.

Deployment: Streamlit Dashboard

Table 2

Class-Specific Performance Metrics for the Top-Ranked Model (Random Forest)

Class	Precision	Recall	F1 Score	Support
Class I	0.9368	0.9144	0.9255	1,671
Class II	0.9450	0.9671	0.9560	5,655
Class III	0.7566	0.6296	0.6873	548
Weighted Avg	0.9302	0.9324	0.9308	7,874

Note. Support indicates the number of test samples in each class.

Concluding Summary

This study evaluated multiple classification models across three target classes using precision, recall, and F1-score. Random Forest consistently outperformed others, achieving metrics around 0.97 for Class I, 0.98 for Class II, and 0.75 for the more complex Class III, highlighting its strength in multi-class classification. The consistent drop in performance for Class III across all models suggests greater overlap or variability in this class, though Random Forest remained the most resilient.

Out-of-Bag (OOB) evaluation confirmed the Random Forest model's robustness without requiring cross-validation. Feature importance analysis revealed that both textual features, like Reason_Word_Count, and time-related variables, such as Month_sin and Month_cos, were key predictors. These findings support model interpretability and offer practical direction for refining future data pipelines.

Call to Action (CTA)

Future work should explore more sophisticated modeling of text-based variables, which were limited in this study. While simple representations like bag-of-words or TF-IDF provided a baseline, advanced transformer models such as LLaMA can capture deeper semantic meaning and context. These models can generate rich embeddings that better reflect the nuances of user-generated text, reviews, or descriptions, potentially improving predictive performance across all classes.

Given more time and resources, future research should implement LLaMA for feature extraction or fine-tuning directly on the task. This could be combined with existing structured data to create a more robust, multi-modal modeling framework. The limited seven-week timeframe of this project restricted the ability to fully explore such methods, but longer-term efforts could allow for expanded experimentation, better model tuning, and integration of real-time or external data sources to enhance model generalizability.

Reference

An, J. (2024). Structural topic modeling for corporate social responsibility of food supply chain management: Evidence from FDA recalls on plant-based food products. *Social Responsibility Journal, 20*(6), 1089–1100. https://doi.org/10.1108/SRJ-07-2023-0412

An, Y. (2024). The impact of supply chain analysis on recall effectiveness. *Journal of Supply Chain Management, 60*(1), 12–23. https://doi.org/10.1111/jscm.12306

Analytics Vidhya. (2017, August 8). *Understanding CatBoost – The fastest implementation of gradient boosting*. https://www.analyticsvidhya.com/blog/2017/08/catboost-automated-categorical-data/

Analytics Vidhya. (2021, June 25). *Understanding Random Forest*. https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/

Anthropic. (2025). Claude 3.7 Sonnet (February 2025 version) [Large language model]. https://claude.ai

Arshad, M. A., Shahriar, S., & Anjum, K. (2023). The power of simplicity: Why simple linear models outperform complex machine learning techniques—Case of breast cancer diagnosis. *arXiv preprint arXiv:2306.02449*. https://arxiv.org/abs/2306.02449

Barbosa-Slivinskis, V., Agi-Maluli, I., & Seth-Broder, J. (2024). A machine learning algorithm to predict medical device recall by the Food and Drug Administration. *The Western Journal of Emergency Medicine, 26*(1), 161–170. https://doi.org/10.5811/westjem.21238

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://dl.acm.org/doi/10.1145/2939672.2939785

Darby, J. L., Ketchen, D. J., Jr., Ball, G. P., & Mukherjee, U. (2023). CEO stock ownership, recall timing, and stock market penalties. *Manufacturing and Service Operations Management, 25*(5), 1909–1930. https://doi.org/10.1287/msom.2021.0175

Datacamp. (n.d.). *Random Forests in Python*. https://www.datacamp.com/tutorial/random-forests-classifier-python

Dubin, J. R., Simon, S. D., Norrell, K., Perera, J., Gowen, J., & Cil, A. (2021). Risk of recall among medical devices undergoing US Food and Drug Administration 510(k) clearance and premarket approval, 2008–2017. *JAMA Network Open, 4*(5), Article e217274. https://doi.org/10.1001/jamanetworkopen.2021.7274

Hastie, T., Tibshirani, R., & Friedman, J. (2020). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.

IBM. (n.d.). *What is Random Forest?* https://www.ibm.com/think/topics/random-forest

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning: With applications in R* (2nd ed.). Springer.

Kaggle. (n.d.). *Random Forest Classifier Tutorial*. https://www.kaggle.com/code/prashant111/random-forest-classifier-tutorial

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. Springer.

Lewinson, E. (2022, February 17). *Three approaches to encoding time information as features for ML models.* NVIDIA Developer Blog. https://developer.nvidia.com/blog/three-approaches-to-encoding-time-information-as-features-for-ml-models/

MarkovML. (2023). *Why establish baseline models: A detailed guide for you.* https://www.markovml.com/blog/baseline-models

Molnar, C. (2022). *Interpretable machine learning* (2nd ed.). https://christophm.github.io/interpretable-ml-book/

Mooghali, M., Ross, J. S., Kadakia, K. T., & Dhruva, S. S. (2023). Characterization of US Food and Drug Administration class I recalls from 2018 to 2022 for moderate- and high-risk medical devices: A cross-sectional study. *Medical Devices (Auckland, N.Z.), 16*, Article 111. https://doi.org/10.2147/MDER.S412802

Lee, S., Tseng, C., Lin, G., Yang, Y., Yang, P., Muhammad, K., & Pandey, H. M. (2020). A dimension-reduction based multilayer perception method for supporting the medical decision making. *Pattern Recognition Letters*, *131*, 15-22. https://doi.org/10.1016/j.patrec.2019.11.026

OpenAI. (2025). *ChatGPT (GPT-4 version)*. OpenAI. https://chat.openai.com

Pagliarini, G., & Sciavicco, G. (2021). Decision tree learning with spatial modal logics. *arXiv preprint arXiv:2109.08325*. https://arxiv.org/abs/2109.08325

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830. https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf

Polisena, J., Jutai, J., & Chreyh, R. (2014). A proposed framework to improve the safety of medical devices in a Canadian hospital context. *Medical Devices: Evidence and Research, 7*, 139–147. https://doi.org/10.2147/MDER.S62796

Qiao, P. (2024). Bayesian algorithm for the construction of logistics node delay model and its impact on subsequent nodes in supply chain. *Journal of Electrical Systems, 20*(3), 386–394. https://doi.org/10.52783/jes.2861

Raschka, S., Mirjalili, V., & Raschka, J. (2022). *Machine learning with PyTorch and Scikit-Learn: Develop machine learning and deep learning models with Python*. Packt Publishing.

Shwartz-Ziv, R., & Armon, A. (2021). Tabular data: Deep learning is not all you need. *arXiv preprint arXiv:2106.03253*. https://arxiv.org/abs/2106.03253

Tharwat, A. (2021). Classification assessment methods. *Applied Computing and Informatics, 17*(1), 168–192. https://doi.org/10.1016/j.aci.2018.08.003

Torrence, M. E. (2002). Data sources: Use in the epidemiologic study of medical devices. *Epidemiology, 13*(3), S10–S14. https://doi.org/10.1097/00001648-200205001-00002

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). LLaMA: Open and Efficient Foundation Language Models. *ArXiv*. https://arxiv.org/abs/2302.13971

U.S. Food and Drug Administration. (2024). *Enforcement reports*. https://www.fda.gov/safety/recalls-market-withdrawals-safety-alerts/enforcement-reports

U.S. Food and Drug Administration. (n.d.). *FDA Dashboards - Recalls*. https://datadashboard.fda.gov/ora/cd/recalls.htm

Wang, C., Hefflin, B., Cope, J. U., Gross, T. P., Ritchie, M. B., Qi, Y., & Chu, J. (2010). Emergency department visits for medical device-associated adverse events among children. *Pediatrics, 126*(2), 247–259. https://doi.org/10.1542/peds.2010-0528

Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2021). Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM, 64*(3), 107–115. https://doi.org/10.1145/3446776

Zhou, Y. (2023). The effects of lobbying on the FDA's recall classification. *BMC Medical Ethics, 24*(1), Article 41. https://doi.org/10.1186/s12910-023-00921-0

Zhou, Z.-H. (2012). *Ensemble methods: Foundations and algorithms*. CRC Press. https://doi.org/10.1201/b12207