Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

From Weak Signals to Supply Chain Reliability Risk: A Feedback-Driven Risk Accumulation Framework for Late-Delivery Prediction, Early Warning, and Risk Prioritization

Yan Lyu

lyu.yan@northeastern.edu

Independent Researcher https://orcid.org/0009-0002-2219-2130

Yihua Xu

Independent Researcher

Jiani He

Independent Researcher

Jize Li

Independent Researcher

Rubing Qiang

Independent Researcher

Yidan Hu

Independent Researcher

Method Article

Keywords: Late-delivery risk prediction, supply chain reliability, weak signals, risk accumulation, early-warning system

Posted Date: November 5th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11614868/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

From Weak Signals to Supply Chain Reliability Risk: A Feedback-Driven Risk Accumulation Framework for Late-Delivery Prediction, Early Warning, and Risk Prioritization

Yan Lyu
Independent Researcher, Boston, Massachusetts, USA — lyu.yan@northeastern.edu
Yihua Xu
Independent Researcher, San Jose, California, USA — yxu442@gmail.com
Jiani He
Independent Researcher, Seattle, Washington, USA — jianihe@alum.mit.edu
Jize Li
Independent Researcher, Boston, Massachusetts, USA — jizel@bu.edu
Rubing Qiang
Independent Researcher, Austin, Texas, USA — rubing.qiang@gmail.com
Yidan Hu
Independent Researcher, New York, New York, USA — lenahu823@gmail.com

Abstract

Late-delivery risk can develop through accumulated operational signals rather than a single disruption. This study proposes a feedback-driven risk accumulation framework that converts matured historical outcomes and observable shipment information into leakage-controlled rolling memory, interpretable escalation features, and decision-oriented risk scores. The primary evaluation uses 180,519 DataCo order-line records partitioned chronologically at the Order Id level to prevent entity leakage. Relative to a gradient-boosting baseline, the feedback model improves ROC-AUC from 0.7674 to 0.8013 and F1 from 0.7194 to 0.7527. Ablation results show that rolling matured-outcome memory accounts for most of the predictive improvement, while persistence, intensity, cross-dimensional breadth, and amplification primarily add diagnostic structure. Under a matched FN5 threshold policy, the feedback model reduces illustrative scenario loss from 5,560 to 5,401 under the assumed 5:1 false-negative-to-false-positive penalty. A secondary SMT2020 fab-lot simulation shows that the same outcome-availability-safe logic can be instantiated for semiconductor due-date-miss prediction. Overall, the results support a retrospective early-warning and triage-readiness evaluation that links temporal risk memory with leakage-controlled prediction, matched review-capacity analysis, and scenario-based threshold policies.

Keywords

Late-delivery risk prediction, supply chain reliability, weak signals, risk accumulation, early-warning system

1. Introduction

Modern supply chains operate across globally distributed production, fulfillment, and transportation networks, which makes shipment reliability a critical issue for operational performance and customer commitment. Prior research on supply chain resilience and supply chain risk management shows that firms must identify vulnerabilities, absorb disruptions, and recover performance under uncertainty (Christopher and Peck 2004; Sheffi and Rice 2005; Ho et al. 2015). In digitally connected environments, artificial intelligence and digital twins can further support monitoring and predictive analysis, but shipment-level reliability still requires models that can convert operational signals into actionable early warnings.

The problem addressed in this study is that many late-delivery risk prediction approaches remain largely static. They often classify each shipment based on current-record attributes such as shipping mode, customer location, product category, order quantity, destination region, and scheduled shipping time. These variables are useful, but they do not fully explain how risk develops through repeated weak signals, recent operational memory, and interaction effects. As a result, managers may receive a probability score without a clear data-to-decision structure for exception triage or cost-aware intervention.

However, prediction accuracy alone is not sufficient for operational use. A late-delivery risk model must also explain how risk develops, rank shipments for limited review capacity, and support different alert thresholds when missed late deliveries and false alerts have unequal business costs. This study therefore treats late-delivery risk prediction as a data-to-decision problem rather than only a binary classification task.

This research is needed because late-delivery risks may emerge gradually before they become visible as final outcomes. The issue is especially relevant for semiconductor-inspired supply chain operations, where global dependencies, long lead times, capacity constraints, specialized suppliers, and high disruption costs make reliability monitoring important (Xiong et al. 2025). Although this study does not use proprietary semiconductor shipment data, it develops and evaluates a semiconductor-inspired feedback-driven risk accumulation logic using a public general supply chain dataset as a proxy empirical testbed. The paper also includes a supplementary SMT2020 fab simulation study to examine how the same outcome-availability-safe risk-accumulation logic can be instantiated at the semiconductor lot level. The main paper therefore focuses on the modeling problem of how weak operational signals can be transformed into risk scores, predictive rankings, and cost-sensitive early-warning decisions.

1.1 Objectives

This study examines three related issues: whether leakage-controlled rolling feedback memory improves late-delivery risk prediction beyond static baseline features; which feature groups drive the improvement; and whether the resulting risk scores support top-k exception triage and cost-sensitive early-warning thresholds.

The study aims to reframe late-delivery risk prediction as a dynamic risk-accumulation problem, operationalize leakage-controlled feedback memory through historical rates, rolling-window signals, EWMA memory, and interpretable escalation mechanisms, and link predictive performance to operational use through ranking and cost-sensitive threshold analysis. A supplementary SMT2020 lot-level experiment is included to examine whether the same risk-accumulation logic can be instantiated in a semiconductor fab simulation setting.

1.2 Contributions

This paper makes three contributions. First, it demonstrates that leakage-controlled rolling operational memory improves late-delivery risk prediction beyond static shipment context and simple historical averages. Second, it organizes temporal risk signals into interpretable escalation dimensions—persistence, intensity, cross-dimensional breadth, and amplification—that support managerial diagnosis. Third, it connects the resulting risk scores to retrospective triage-readiness evaluation through top-k prioritization and cost-sensitive threshold policies.

2. Literature Review

This literature review is organized around four themes that directly support the proposed framework: supply chain resilience and digital monitoring, machine learning for logistics delay prediction, weak-signal and cost-sensitive decision analytics, and the resulting research gap. The goal is not to repeat the motivation from the Introduction, but to position the study within prior work and clarify how this paper differs from existing prediction-oriented approaches.

2.1 Supply Chain Resilience and Digital Monitoring

Supply chain resilience research provides the foundation for studying shipment reliability as more than an isolated delivery outcome. Christopher and Peck (2004) emphasized the need to understand supply chain vulnerability, while Sheffi and Rice (2005) described resilience as the ability to absorb disruption and recover performance. Ho et al. (2015) further organized supply chain risk management around risk identification, assessment, mitigation, and monitoring. Recent digital supply chain studies extend this view by showing that artificial intelligence and digital twins can support real-time monitoring, predictive analysis, collaboration, and disruption scenario simulation (Pan et al. 2025; Roman et al. 2025). This stream establishes why proactive monitoring is important, but it does not by itself specify how shipment-level weak signals should be converted into predictive features and operational actions.

2.2 Machine Learning in Logistics Delay Prediction

Machine learning research shows that operational data can support logistics delay and supply chain risk prediction. Baryannis et al. (2019) proposed and evaluated a data-driven machine-learning framework for supply chain risk prediction and highlighted the trade-off between predictive performance and interpretability. Ülkü Küp et al. (2024) examined real-time delivery delay prediction using models such as Logistic Regression, XGBoost, CatBoost, and Random Forest, and Sultana et al. (2025) studied online order fulfillment and late-delivery risk prediction. These studies demonstrate the value of predictive modeling, but many approaches still emphasize current shipment attributes and model accuracy. The present study extends this stream by adding rolling feedback memory, risk accumulation, and ranking-based decision support. Recent operations-oriented studies also show that data-driven methods can support production planning under demand uncertainty and transport-throughput forecasting for logistics planning (Su et al. 2025; Shao et al. 2025).

2.3 Weak Signals, Cost-Sensitive Learning, and Decision-Oriented Analytics

The weak-signal perspective suggests that early and partial indicators can reveal emerging risks before disruptions become fully observable (Ansoff 1975). This idea supports the paper's use of persistence, intensity, spread, and amplification as interpretable mechanisms for turning weak operational signals into accumulated risk. Cost-sensitive learning further supports the decision layer of the framework because different classification errors may carry different operational penalties (Elkan 2001). In shipment monitoring, a missed late delivery can be more costly than a false alert, so the same predicted probabilities may require different thresholds under different business priorities. Decision-oriented risk analytics also argues that prediction should support intervention planning rather than only correlation discovery (Wyrembek et al. 2025).

2.4 Literature Gap and Study Positioning

Taken together, prior studies provide strong foundations for resilience, digital monitoring, delay prediction, weak-signal detection, and cost-sensitive analytics. However, they leave a gap at the shipment-record level: fewer studies integrate rolling feedback memory, interpretable risk escalation, nonlinear prediction, ranking evaluation, and cost-sensitive early-warning thresholds into one data-to-decision framework. This paper addresses that gap by developing a feedback-driven risk accumulation framework and testing it on a public supply chain fulfillment dataset. The study is semiconductor-inspired because semiconductor supply chains face global dependencies, long lead times, specialized suppliers, and high disruption costs (Xiong et al. 2025), but the empirical results are interpreted as proxy evidence rather than direct semiconductor validation. The supplementary SMT2020 experiment adds fab-level simulation evidence by replacing shipment-level late-delivery risk with lot tardiness while keeping the same feedback-risk accumulation pipeline.

Existing delivery-risk models often rely on current shipment attributes or long-run historical aggregates. Less is known about whether short-horizon, leakage-controlled rolling operational memory improves chronological late-delivery risk prediction and supports capacity-constrained exception triage. This study addresses that gap through rolling feedback features, ablation analysis, top-k prioritization, and scenario-based threshold evaluation.

3. Methods

3.1 Conceptual Framework

This study proposes a four-layer feedback-driven risk accumulation framework for late-delivery risk prediction, as shown in Figure 1. The framework explains how late-delivery risk emerges from static operational context, dynamic operational anomalies, temporal feedback memory, nonlinear risk escalation, and cost-sensitive decision support. Different from a conventional static classification model, the proposed framework treats late-delivery risk as a dynamic process in which weak operational signals are captured, transformed into accumulated risk, used for predictive analytics, and then converted into early-warning actions.

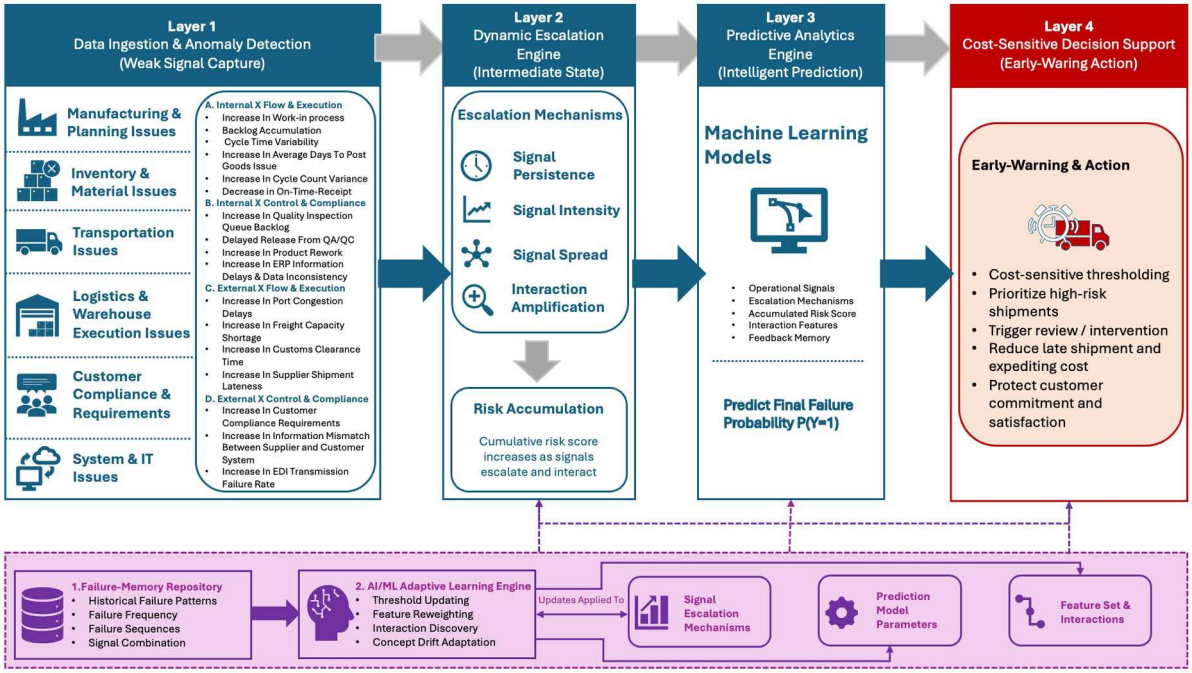


Figure 1. Feedback-driven operational reliability framework for late-delivery risk prediction.

Layer 1 represents data ingestion and anomaly detection, or weak-signal capture. This layer combines static shipment context and dynamic operational anomaly indicators. Static context features include shipment mode, product category, market, region, country, customer segment, and scheduled shipping days. This layer also captures dynamic operational anomalies, such as backlog increase, supplier schedule drift, cycle-time variability, quality queue build-up, port or customs delay, and information transmission failure.

Layer 2 represents the dynamic risk escalation engine. This layer converts captured weak signals into interpretable risk accumulation. The framework identifies four escalation mechanisms: signal persistence, signal intensity, cross-dimensional risk breadth, and interaction amplification. These mechanisms explain how small operational disturbances can develop into a cumulative risk condition.

The output of Layer 2 is a cumulative risk score, or RiskScore. The RiskScore compresses temporal and cross-dimensional weak signals into an interpretable risk-accumulation feature. It does not replace the machine learning model; instead, it serves as a structured feature that summarizes how recent operational anomalies, historical late-delivery patterns, and interaction effects contribute to late-delivery risk.

Layer 3 represents the predictive analytics engine. In this layer, a nonlinear machine learning model estimates the probability of late-delivery risk. The model uses operational signals, the accumulated RiskScore, interaction features, and feedback memory variables as inputs. Because late-delivery risk may be nonlinear and interaction-driven, this study uses a histogram-based gradient boosting decision tree model to estimate the final late-delivery risk probability (Friedman 2001; Hastie et al. 2009; Ke et al. 2017).

Layer 4 represents cost-sensitive decision support. The predicted late-delivery probability from Layer 3 is converted into operational decisions through cost-sensitive thresholding. This thresholding process considers the asymmetric costs of false negatives and false positives, following the cost-sensitive learning view that different classification errors may carry different operational penalties (Elkan 2001).

The bottom dashed box in Figure 1 represents a conceptual cross-layer feedback-learning mechanism. In the two empirical studies, the implemented components are matured historical memory, rolling feedback, risk accumulation, model scoring, ranking, and validation-selected decision thresholds. Adaptive feature reweighting, automated threshold updating, live interaction discovery, and concept-drift adaptation are deployment extensions rather than separately validated components in the reported experiments. Thus, Figure 1 should be read as a framework architecture rather than evidence that all adaptive feedback functions have been deployed in real time.

In the empirical feature set, this cross-dimensional breadth is operationalized as Spread, which counts how many lagged operational feedback channels show above-average late-delivery-risk signals.

3.2 Feature Engineering and Prediction Method

Feature engineering follows the structure of the proposed framework. Baseline operational features represent Layer 1 disturbance drivers and include shipment mode, market, region, country, customer segment, product category, order quantity, scheduled shipping days, sales variables, discount information, and order timing variables. These features describe the static operational context of each shipment.

The enhanced feedback features represent historical memory, escalation, and risk accumulation. Historical late-rate signals are calculated across operational groups such as region, shipping mode, market, country, and product category. These signals are then transformed into four escalation variables.

Table 1. Enhanced feedback features and model equations.

Feature / equation	Definition
Persistence_i	mean(historical late-rate signals across region, ship mode, market, country, category)
Intensity_i	log(1 + max(0, historical delay gap_i))
Spread_i	Count of operational feedback channels with above-average lagged late-delivery-risk signals
Amplification_i	Persistence_i x (1 + Intensity_i) x (1 + Spread_i / K)
RiskScore_i	0.40 x Persistence_i + 0.25 x Intensity_i + 0.20 x Spread_i/K + 0.15 x Amplification_i
EWMA_g,t	$\alpha \cdot y_{g,t-1} + (1-\alpha) \cdot EWMA_{g,t-1}$

The enhanced model adds 7/14/30-day rolling windows and EWMA memory constructed on an as-of basis. For a focal prediction at time t_i, only outcomes with an observation time strictly earlier than t_i update the rolling or EWMA state; same-time and still-open records are excluded. It also adds interaction variables such as risk score × scheduled days, persistence × intensity, and ship-mode × region risk. The RiskScore weights are expert-specified heuristic coefficients used to construct an interpretable summary feature. They are not fitted on the test set and are not presented as uniquely optimal or enterprise-calibrated values. The fixed weighting scheme is treated as a design choice for organizing persistence, intensity, spread, and amplification into one composite risk feature, while the usefulness of the feedback feature set is assessed through downstream model comparison and ablation results. A sensitivity analysis of alternative weight vectors is left for future work. The main nonlinear classifier is a histogram-based gradient boosting model. The model estimates $P(Y_i = 1 \mid X_i)$, where X_i includes baseline operational variables, feedback memory variables, escalation features, and interaction terms.

4. Data Collection

The primary DataCo experiment uses 180,519 order-line fulfillment records from the DataCo SMART Supply Chain dataset, representing 65,752 unique Order Id entities. The original DataCo release also includes 469,977 tokenized access-log records (Constante et al. 2019), but those records are not used in the final predictor set. The evaluation orders entities by their earliest order timestamp and assigns 70%/15%/15% of unique Order Id entities to train/validation/test partitions, yielding 138,062/29,511/12,946 order-line rows; no Order Id appears in more than one partition.

The unit of analysis is an order-line fulfillment record, identified at the row level by Order Item Id and linked to its business entity by Order Id. The binary target variable is defined using the Late_delivery_risk field: $Y_i = 1$ indicates late-delivery risk and $Y_i = 0$ indicates no observed late-delivery risk. The observed late-delivery-risk rate is 54.83%. All rows sharing an Order Id remain in the same partition to prevent same-entity leakage across train, validation, and test splits.

The DataCo prediction timestamp is order_dt, after order-level planned shipment information is available and before the realized late-delivery outcome is known. The outcome-availability timestamp is shipping date (DateOrders), stored as ship_dt/outcome_available_dt; the final label is treated as observable only after this shipment timestamp. For focal record i at time t_i, outcome-based historical memory is constructed only from prior business entities whose outcome_available_dt is strictly earlier than t_i. An earlier-created order that is still open at t_i is excluded, and same-time outcomes are also excluded by a conservative tie rule. Direct outcome variables, including Delivery Status and realized shipping duration, remain excluded from the ordinary predictor set.

Table 2. Feature availability and leakage-control design.

Feature group	Example variables	Prediction-time availability	Leakage risk	Treatment

Direct outcome variables	Delivery Status, real shipping duration	Not available before outcome	High	Removed
Static shipment context	shipping mode, region, market, country, category	Available before completion	Low	Used
Planned service variables	scheduled shipping days	Available before completion	Low	Used
Sales/order variables	sales, discount, order quantity	Available at order time	Low/Medium	Used
Historical late-rate	region/mode/category late rate	Available only after prior outcomes mature	Medium	As-of outcome-time construction; outcome_available_dt < focal order_dt
Rolling memory	7d/14d/30d rolling late rate	Available only from matured outcomes in the lookback window	Medium	7/14/30-day windows indexed by outcome-observation time; strict past-only rule
Same-record outcome	actual shipping duration, same-record late status	Not available before outcome	High	Removed

All outcome-based historical rates, 7/14/30-day rolling rates, EWMA signals, delay-gap memory, Persistence, Intensity, Spread, Amplification, and RiskScore use the same as-of outcome-availability rule. Order-line rows from the same Order Id do not contribute the entity's final outcome to one another before that outcome matures. Threshold-dependent comparisons use validation-selected policies: each model receives its own F1-optimal threshold for balanced classification and its own FN5 cost-optimal threshold for matched cost comparison; top-k comparisons use the same 5%, 10%, 20%, and 30% review capacities.

5. Results and Discussion

5.1 Numerical Results

Table 3 reports the main model comparison on the chronological test set. The purpose of this comparison is to evaluate whether the proposed feedback-driven risk accumulation framework improves late-delivery risk prediction when implemented with both linear and nonlinear learning models. The models include a linear baseline, a linear feedback model, a gradient boosting baseline, and three enhanced gradient boosting feedback models with different decision thresholds. The F1-based model emphasizes balanced classification performance, the F2-based model places greater weight on recall, and the cost-optimized model selects the decision threshold based on the assumption that missed late deliveries are more costly than false alerts. ROC-AUC evaluates discrimination across decision thresholds, while PR-AUC is useful because the positive late-delivery-risk class is operationally important (Fawcett 2006; Saito and Rehmsmeier 2015).

ROC-AUC and PR-AUC are used for threshold-independent score-quality comparison, while precision, recall, F1, FN, FP, and illustrative scenario loss reflect the selected decision threshold for each policy.

Table 3. Enhanced model comparison on the chronological test set.

Model	ROC-AUC	PR-AUC	Prec.	Recall	F1	FN	FP	Scenario Loss	Policy	Thres.
GB Feedback F1	0.8013	0.8481	0.6505	0.8930	0.7527	758	3,398	7,188	F1-opt	0.370
Linear Feedback	0.7613	0.8202	0.5798	0.9462	0.7190	381	4,858	6,763	F1-opt	0.220
GB Feedback F2	0.8013	0.8481	0.5720	0.9966	0.7269	24	5,281	5,401	F2-opt	0.180
GB Feedback CostOpt	0.8013	0.8481	0.5720	0.9966	0.7269	24	5,281	5,401	FN5-cost-opt	0.180
GB Base	0.7674	0.8332	0.6184	0.8597	0.7194	994	3,757	8,727	F1-opt	0.380
Linear Base	0.7430	0.8124	0.6319	0.8032	0.7073	1,394	3,314	10,284	F1-opt	0.300

The results show that the feedback gradient boosting model improves threshold-independent discrimination and F1 performance relative to the gradient-boosting baseline. It achieves ROC-AUC = 0.8013, PR-AUC = 0.8481, precision = 0.6505, recall = 0.8930, and F1 = 0.7527 at its validation-selected F1 threshold of 0.370. The gradient-boosting baseline achieves ROC-AUC = 0.7674, PR-AUC = 0.8332, and F1 = 0.7194 at threshold 0.380.

The linear feedback model improves F1 from 0.7073 to 0.7190 and increases ROC-AUC from 0.7430 to 0.7613. This indicates that matured feedback-memory features add predictive information beyond the static linear baseline, although the magnitude of improvement is smaller after enforcing outcome availability and entity-safe partitioning.

Because the observed late-delivery-risk rate is 54.83%, the results should also be interpreted against a simple majority-class baseline. An all-positive baseline would classify every shipment as a late-delivery-risk case and would achieve perfect recall but only moderate precision. The feedback gradient boosting model's F1-score of 0.7527 and PR-AUC of 0.8481 indicate that the gain is not simply caused by the high positive-class rate.

The ablation study evaluates which components of the enhanced framework contribute most to the improvement in late-delivery risk prediction. Table 4 reports five model variants. A1 uses only baseline operational features. A2 adds simple historical feedback features. A3 adds rolling feedback memory features. A4 further adds escalation variables. A5 represents the full feedback framework with rolling memory, escalation features, and interaction terms.

Table 4. Ablation study of feedback feature layers.

Model	ROC-AUC	PR-AUC	Prec.	Recall	F1	FN	FP
A1 Base	0.7674	0.8332	0.6184	0.8597	0.7194	994	3,757
A2 + Hist	0.7663	0.8326	0.5989	0.9050	0.7208	673	4,293
A3 + Rolling	0.8014	0.8484	0.6486	0.8940	0.7518	751	3,431
A4 + Escalation	0.8020	0.8486	0.6436	0.9055	0.7524	669	3,552
A5 Full	0.8013	0.8481	0.6505	0.8930	0.7527	758	3,398

The ablation results continue to show that rolling feedback memory is the main enhancement. Adding rolling memory in A3 raises ROC-AUC to 0.8014, PR-AUC to 0.8484, and F1 to 0.7518. A4 and A5 produce only small additional aggregate changes. Thus, most of the gain is associated with rolling matured-outcome memory; escalation and interaction layers provide smaller refinements and diagnostic organization.

Ranking evaluation is more conservative after the leakage and entity corrections. At the same top-20% review capacity, the gradient-boosting feedback model captures 34.05% of late-delivery cases with precision@top20% = 0.9313 and lift = 1.7021, while the gradient-boosting baseline captures 34.03% with precision@top20% = 0.9305 and lift = 1.7007. The near-equality at this capacity indicates that the principal benefit is improved overall discrimination and thresholded classification rather than a large top-k ranking gain at 20%.

Table 5. Illustrative asymmetric classification-loss function.

Item	Definition
Scenario loss	C_FN x FN + C_FP x FP
Main paper setting	C_FN = 5, C_FP = 1

Table 6. Matched validation-selected FN5 scenario-loss policy on the DataCo chronological test set.

Model	Thres.	Prec.	Recall	F1	FN	FP	Scenario Loss	Scenario Loss/1k
Linear Base	0.125	0.5476	0.9989	0.7074	8	5,846	5,886	454.7
Linear Feedback	0.100	0.5481	1.0000	0.7081	0	5,841	5,841	451.2
GB Base	0.205	0.5602	1.0000	0.7181	0	5,560	5,560	429.5
GB Feedback	0.180	0.5720	0.9966	0.7269	24	5,281	5,401	417.2

Late-delivery risk prediction is not a symmetric classification problem because different error types can carry different operational consequences. To make the FN5 comparison policy-matched, each model independently selects its own validation-set threshold minimizing 5×FN + FP and is evaluated once on the same held-out test set.

Under this matched FN5 policy, the gradient-boosting feedback model selects threshold 0.180 and obtains an illustrative scenario loss of 5,401, compared with 5,560 for the gradient-boosting baseline. The corresponding linear feedback and linear baseline scenario losses are 5,841 and 5,886, respectively. These values are relative classification-loss units under the assumed 5:1 false-negative-to-false-positive penalty, not realized monetary savings. Therefore, the FN5 analysis should be interpreted as a retrospective threshold-policy simulation rather than evidence of observed operational cost reduction.

Together with the matched-capacity ranking analysis, the matched FN5 comparison shows that the empirical contribution should be interpreted in three parts: the feedback model improves overall discrimination and F1, produces only marginal change in matched top-20% capture, and reduces illustrative FN5 scenario loss modestly under the assumed relative penalty. Therefore, the evidence supports retrospective triage-readiness and threshold-policy evaluation rather than demonstrated field-level operational impact.

5.2 Graphical Results

The graphical results summarize the main model comparison, PR-AUC changes across ablation variants, and cost-sensitive threshold behavior.

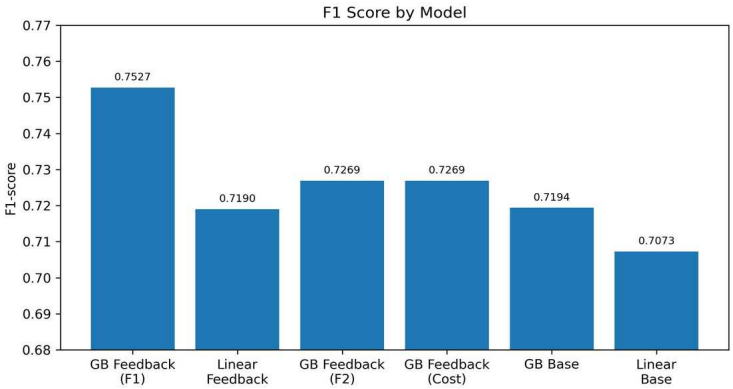


Figure 2. F1 comparison across baseline, linear feedback, and gradient boosting feedback models.

Figure 2 visualizes the F1-score comparison across the main model variants. The enhanced feedback gradient boosting model has the highest F1-score, followed by the recall-oriented and cost-optimized feedback variants, the linear feedback model, and the baseline models. This pattern suggests that the proposed framework is most effective when feedback-driven feature engineering is combined with nonlinear learning.

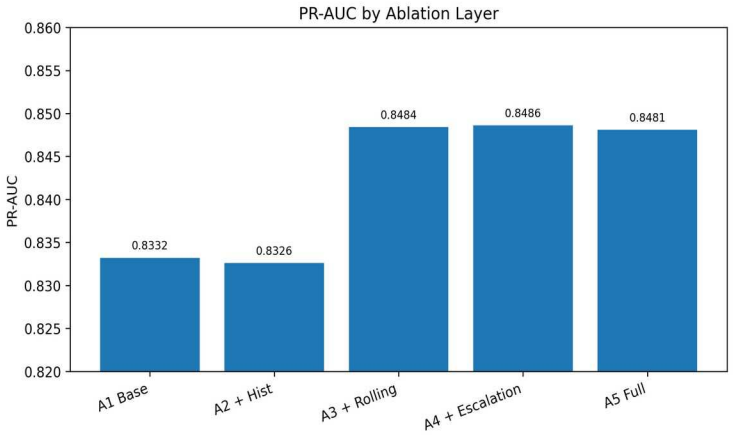


Figure 3. PR-AUC by ablation layer.

Figure 3 shows that the main threshold-independent performance improvement occurs when rolling matured-outcome memory is added. PR-AUC increases from 0.8332 in A1 to 0.8484 in A3; A4 is 0.8486 and A5 is 0.8481, confirming that escalation and interaction layers add little additional PR-AUC beyond rolling memory.

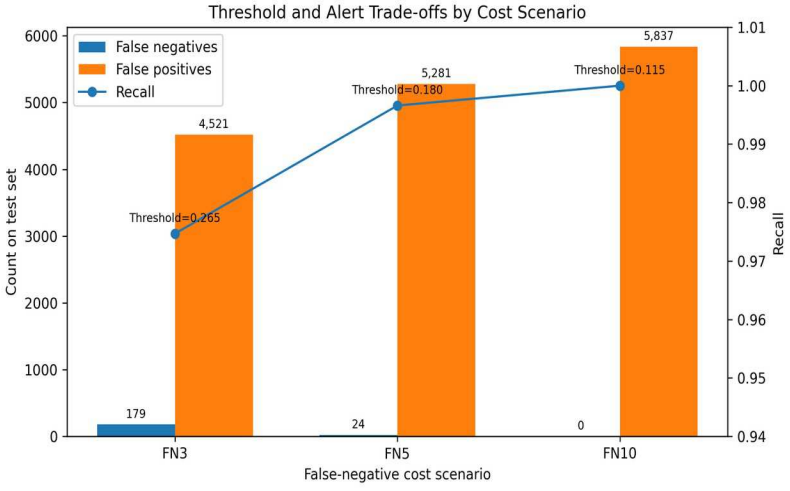


Figure 4. Threshold and alert trade-offs under illustrative false-negative cost scenarios.

Figure 4 shows how the validation-selected threshold changes under alternative false-negative cost assumptions for the feedback model. As the assumed false-negative cost increases, the selected threshold decreases, recall increases, and false negatives decrease. The figure illustrates threshold and alert trade-offs rather than ranking the scenarios on a common cost scale.

5.3 Managerial Implications and Future Research

The proposed framework improves conventional late-delivery risk prediction by moving from static shipment attributes to dynamic feedback-driven risk accumulation. A shipment may appear normal as an isolated record but become risky when recent operational memory and accumulated weak signals are considered. Future research should validate the framework using semiconductor-specific operational data, such as fab cycle-time variance, wafer start delays, inventory levels, assembly/test backlog, supplier reliability, and customer commitment performance. Future work could also extend the framework from prediction to intervention by evaluating actions such as carrier switching, shipment expediting, production rescheduling, or customer-commitment adjustment. Future field validation should evaluate whether model-guided alerts improve actual decision quality, using endpoints such as planner acceptance, time-to-action, expedite or carrier-switch decisions, rescheduling actions, late-delivery reduction, avoided expediting cost, false-alert review workload, customer-commitment changes, and planner override rates.

5.4 Validation

The validation protocol combines entity-safe chronological holdout testing, outcome-availability control, ablation analysis, matched-policy threshold evaluation, and matched-capacity ranking. The DataCo split contains 46,026/9,863/9,863 unique Order Id entities in train/validation/test, with zero entity overlap across all partition pairs. The key leakage-control rule is temporal observability: only an outcome with outcome_available_dt strictly earlier than the focal order_dt may update outcome-based memory.

Several limitations remain. The DataCo dataset is not semiconductor-specific, the evaluation uses one chronological holdout, and shipping date is used as the DataCo outcome-availability timestamp for the late-delivery label. SMT2020 strengthens cross-context transferability evidence but does not constitute semiconductor field deployment. Future work should test rolling-origin windows or paired bootstrap uncertainty, calibrated intervention costs, live operational signals, and model-explanation tools and local feature-contribution diagnostics.

5.5 Cross-Context Semiconductor Simulation Extension Using SMT2020

Study 2 applies the feedback-driven risk accumulation pipeline to SMT2020 AutoSched fab-lot simulation outputs (Kopp et al. 2020). It provides a secondary cross-context semiconductor simulation setting for evaluating whether outcome-availability-safe feedback memory can support lot-level due-date-miss prediction. The transfer from DataCo to SMT2020 maps the unit of analysis from order-line fulfillment record to fab lot, the target from late-delivery risk to due-date miss, the prediction timestamp from order_dt to STARTDATE, and the outcome-availability timestamp from ship_dt/outcome_available_dt to COMPDATE. The SMT2020 target is lot tardiness, defined as $Y_i = 1$ when the lot completion date exceeds the due date and $Y_i = 0$ otherwise.

The experiment uses SMT2020 dataset 4, LVHM_E_Model, and excludes initial WIP lots with incomplete pre-simulation histories. The resulting sample contains 97,428 fab lots, with an overall tardy rate of 6.37% and a held-out test tardy rate of 7.40%. In both studies, Persistence summarizes matured historical outcome rates, Intensity captures recent delay or cycle-time pressure, Spread counts elevated risk channels, Amplification represents interaction among risk dimensions, and rolling memory is indexed by the corresponding outcome-observation clock.

In SMT2020, all historical rates, rolling windows, EWMA memory, and prior cycle-time memory are constructed only from lots completed strictly before the focal lot's release time, with same-time completions excluded conservatively.

Table 7. SMT2020 fab-lot due-date miss prediction results

Model	Thres.	ROC-AUC	PR-AUC	Prec.	Recall	F1	FN	FP	Scenario Loss
Linear Base	0.44	0.8208	0.2154	0.2075	0.8252	0.3316	189	3,407	4,352
Linear Feedback	0.40	0.8254	0.2435	0.1996	0.8372	0.3223	176	3,630	4,510
GB Base	0.59	0.7983	0.2788	0.4134	0.1480	0.2180	921	227	4,832
GB Feedback F1	0.53	0.8372	0.3333	0.4955	0.1517	0.2323	917	167	4,752
GB Feedback F2	0.16	0.8372	0.3333	0.2200	0.7974	0.3448	219	3,057	4,152
GB Feedback CostOpt	0.85	0.8372	0.3333	0.6128	0.1332	0.2188	937	91	4,776

The SMT2020 results provide simulation-based cross-context evidence that the proposed outcome-availability-safe feedback-memory construction can be instantiated for semiconductor fab-lot due-date-miss prediction. The feedback gradient boosting model achieves ROC-AUC = 0.8372 and PR-AUC = 0.3333; the F2-oriented feedback policy increases recall to 0.7974 and F1 to 0.3448 for high-recall due-date-miss monitoring. Because thresholds are selected on the validation set rather than optimized on the test set, the row labeled F1-opt does not necessarily produce the highest realized test-set F1.

Table 8. SMT2020 risk prioritization performance under different review capacities

Review capacity	Due-date miss capture	Precision	Lift
Top 5%	24.33%	0.3598	4.8642
Top 10%	39.04%	0.2886	3.9025
Top 20%	61.70%	0.2282	3.0851
Top 30%	85.75%	0.2114	2.8581

The top 20% highest-risk lots capture 61.70% of due-date misses, showing that the feedback-risk score can concentrate future tardiness into a smaller review queue for dispatcher attention and bottleneck monitoring.

Table 9. SMT2020 cost-sensitive threshold results

Scenario	C_FN	Thres.	Prec.	Recall	F1	FN	FP	Scenario Loss	Scenario Loss/1k
FN3	3	0.85	0.6128	0.1332	0.2188	937	91	2,902	198.6
FN5	5	0.85	0.6128	0.1332	0.2188	937	91	4,776	326.8
FN10	10	0.18	0.2311	0.5698	0.3288	465	2,050	6,700	458.4

The cost-sensitive SMT2020 results show the same decision logic as DataCo. In the highest false-negative-cost scenario, the selected threshold decreases from 0.85 to 0.18, recall increases from 0.1332 to 0.5698, and false positives increase from 91 to 2,050, illustrating the expected high-recall trade-off.

Overall, the SMT2020 extension complements the DataCo analysis by showing that the same outcome-availability-safe risk-accumulation formulas and validation protocol can be instantiated in semiconductor fab simulation as lot tardiness prediction. The simulation-based setting provides a structured path for future validation with semiconductor-specific operational data.

6. Conclusion

This study develops a feedback-driven risk accumulation framework for late-delivery risk prediction and exception triage. Its central empirical hypothesis is tested with outcome-availability-safe operational memory: a prior final label can enter feedback only after that outcome is observable at the focal prediction time. The ablation evidence indicates that rolling memory is the main source of predictive improvement, while persistence, intensity, spread, amplification, and interactions primarily provide smaller refinements and diagnostic organization. Matched validation-selected thresholds and matched top-k capacities connect score quality to retrospective triage-readiness evaluation without conflating model changes with policy changes.

The paper uses a two-study design: DataCo provides the primary public fulfillment evaluation, while SMT2020 provides a secondary cross-context semiconductor simulation test. Together, the two studies show how matured feedback-memory construction can support late-delivery risk prediction and lot-level due-date-miss prioritization under leakage-controlled evaluation.

Declaration of Generative AI Use

During the preparation of this manuscript, the authors used ChatGPT and Codex (OpenAI) to refine wording and improve grammatical consistency. The authors reviewed and edited the AI-assisted text and take full responsibility for the content of the manuscript.

References

Ansoff, H. I., Managing strategic surprise by response to weak signals, *California Management Review*, vol. 18, no. 2, pp. 21-33, 1975.

Baryannis, G., Dani, S. and Antoniou, G., Predicting supply chain risks using machine learning: The trade-off between performance and interpretability, *Future Generation Computer Systems*, vol. 101, pp. 993-1004, 2019.

Christopher, M. and Peck, H., Building the resilient supply chain, *The International Journal of Logistics Management*, vol. 15, no. 2, pp. 1-14, 2004.

Constante, F., Silva, F. and Pereira, A., DataCo SMART Supply Chain for Big Data Analysis [Dataset], Mendeley Data, Version 5, 2019.

Elkan, C., The foundations of cost-sensitive learning, *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, pp. 973-978, Seattle, Washington, USA, August 4-10, 2001.

Fawcett, T., An introduction to ROC analysis, *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861-874, 2006.

Friedman, J. H., Greedy function approximation: A gradient boosting machine, *The Annals of Statistics*, vol. 29, no. 5, pp. 1189-1232, 2001.

Hastie, T., Tibshirani, R. and Friedman, J., *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed.*, Springer, New York, 2009.

Ho, W., Zheng, T., Yildiz, H. and Talluri, S., Supply chain risk management: A literature review, *International Journal of Production Research*, vol. 53, no. 16, pp. 5031-5069, 2015.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. and Liu, T.-Y., LightGBM: A highly efficient gradient boosting decision tree, *Advances in Neural Information Processing Systems*, vol. 30, pp. 3146-3154, 2017.

Kopp, D., Hassoun, M., Kalir, A. and Mönch, L., SMT2020—A semiconductor manufacturing testbed, *IEEE Transactions on Semiconductor Manufacturing*, vol. 33, no. 4, pp. 522-531, 2020.

Pan, H., Zou, N., Wang, R., Ma, J. and Liu, D., Artificial intelligence usage and supply chain resilience: An organizational information processing theory perspective, *Systems*, vol. 13, no. 9, article 724, 2025.

Roman, E.-A., Stere, A.-S., Roșca, E., Radu, A.-V., Codroiu, D. and Ilie, A., State of the art of digital twins in improving supply chain resilience, *Logistics*, vol. 9, no. 1, article 22, 2025.

Saito, T. and Rehmsmeier, M., The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets, *PLOS ONE*, vol. 10, no. 3, e0118432, 2015.

Shao, B., Su, X., Li, X. and Zhuo, X., Mixed-frequency data-driven forecasting port throughput: a novel attention-DeepAR-MIDAS model, *International Journal of Shipping and Transport Logistics*, vol. 20, no. 3, pp. 338-358, 2025.

Sheffi, Y. and Rice, J. B. Jr., A supply chain view of the resilient enterprise, *MIT Sloan Management Review*, vol. 47, no. 1, pp. 41-48, 2005.

Su, X., Zeng, L., Shao, B. and Lin, B., Data-driven optimization for production planning with multiple demand features, *Kybernetes*, vol. 54, no. 1, pp. 110-133, 2025.

Sultana, N., Muhammad Moosa, A., Faisal, S. M. F. and Noor Chowdhury, T., Machine learning models for optimizing online order fulfillment - Forecasting lead time and late delivery risk, *Journal of Optimization and Supply Chain Management*, vol. 2, no. 2, pp. 70-81, 2025.

Ülkü Küp, B., Küp, E. T., Koçak, G., Yücekaya, A. D. and Hekimoğlu, M., Real-time prediction of delivery delay in supply chains using machine learning approaches, *SSRN Working Paper No. 5062672*, 2024.

Wyrembek, M., Baryannis, G. and Brintrup, A., Causal machine learning for supply chain risk prediction and intervention planning, *International Journal of Production Research*, vol. 63, no. 15, pp. 5629-5648, 2025.

Xiong, W., Wu, D. D. and Yeung, J. H. Y., Semiconductor supply chain resilience and disruption: insights, mitigation, and future directions, *International Journal of Production Research*, vol. 63, no. 9, pp. 3442-3465, 2025.