AI and Data Science Journal
ISSN: 3078-7394
DOI: https://doi.org/10.61784/adsj4041

XOR-EW: AN EXPLAINABLE AND COST-AWARE ARTIFICIAL INTELLIGENCE FRAMEWORK FOR OPERATIONAL RISK EARLY WARNING IN FINANCIAL INSTITUTIONS

YiZhen Lin[1], Ying Wang[2], MengLin Bian[3*]
[1]*Stanford University, Stanford 94305, CA, USA.*
[2]*Pepperdine University, Malibu 90263, CA, USA.*
[3]*University of Missouri, Columbia 65211, MO, USA.*
[*]*Corresponding Author: MengLin Bian*

Abstract: Operational-risk early-warning models in financial institutions must do more than rank rare events: they must remain stable under temporal change, express meaningful probabilities, respect finite investigation capacity, and provide auditable reasons. This paper presents XOR-EW (eXplainable Operational-Risk Early Warning), a decision-oriented framework that treats transaction as an observable proxy for a class of process- and control-related operational loss events. XOR-EW combines strict chronological controls, stability-aware champion selection, post-hoc probability calibration, an exposure-sensitive threshold under an alert-capacity constraint, ensemble-disagreement triage, and SHAP-based reason codes with explanation-stability checks. Experiments were executed on a publicly accessible mirror of the ULB/Worldline credit-card data. After exact deduplication, 227,102 transactions were split chronologically into training, calibration, policy, and test windows. A balanced random forest was selected without consulting the test set because it achieved the highest temporal-robustness score across the calibration and policy windows. On the 45,421-transaction test window, calibration reduced the Brier score from 0.000511 to 0.000421 and ten-bin expected calibration error from 0.003736 to 0.000304. Relative to the raw champion with an F1-selected threshold, the complete policy increased recall from 0.561 to 0.737, increased F1 from 0.719 to 0.785, and reduced normalized exposure cost from 0.731 to 0.358. An additional eight disagreement cases raised total capture to 0.772 while total workload remained 0.128% of transactions. The results support XOR-EW as a reproducible governance architecture, not as a claim of universal superiority or production readiness.
Keywords: Operational risk; Early warning; Explainable artificial intelligence; Detection; Probability calibration; Cost-sensitive learning; Selective prediction; SHAP

1 INTRODUCTION

Financial institutions face operational losses from failed processes, human actions, systems, external events. Supervisory principles emphasize governance, risk identification, monitoring, control, and disclosure rather than treating operational-risk analytics as an isolated modeling exercise [1-2]. Model-risk guidance likewise requires robust development, independent validation, ongoing monitoring, and effective challenge [3]. These requirements make a conventional "train a classifier and report accuracy" workflow insufficient for an early-warning system.

Transaction offers a useful, observable test bed for a narrower operational-risk problem: detecting loss events early enough to support human intervention. It is not a complete representation of operational risk. Nevertheless, it captures several difficult properties shared by operational-risk monitoring: extreme rarity, asymmetric consequences, evolving behavior, delayed labels, limited investigation capacity, and a need for explanations. The NIST AI Risk Management Framework further treats validity, reliability, transparency, explainability, accountability, and knowledge limits as connected characteristics of trustworthy AI [4-5].

Three gaps motivate this study. First, random train-test splits can leak future transaction regimes into model selection. Second, high ranking performance does not guarantee usable risk probabilities, especially under class weighting or undersampling [6]. Third, a single score does not specify which transactions should be blocked, reviewed, or cleared when analyst capacity and transaction exposure vary. Prior streaming and hybrid systems have addressed scale, nonstationarity, latency, and supervised-unsupervised combinations [7-8], while cost-sensitive studies have shown the value of transaction-specific consequences [9-11]. However, temporal selection, calibration, capacity-aware action, uncertainty routing, and explanation stability are rarely evaluated as one auditable pipeline.

This paper proposes XOR-EW, an integrated early-warning framework with four contributions. (1) It introduces a temporal-robustness score that selects a champion using two pre-test windows and penalizes unstable average-precision performance. (2) It separates ranking, probability calibration, and policy optimization, enabling a threshold that minimizes an exposure-sensitive scenario cost subject to an alert-capacity limit. (3) It routes borderline cases to human review using disagreement across groups of trees, while urgent alerts receive calibrated probabilities and SHAP reason codes. (4) It reports full reproducibility artifacts and explicitly distinguishes observed results from deployment claims. The framework is evaluated with strict chronological splits, four model families, calibration, ablation, capacity sensitivity, drift diagnostics, and explanation-stability resampling.

© By the Author(s) 2024, under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0).

2 RELATED WORK AND RESEARCH GAP

2.1 Imbalanced and Cost-Sensitive Detection

detection is a canonical imbalanced-learning problem. Random forests provide nonlinear interactions and averaging across randomized trees [12], while XGBoost and LightGBM offer efficient gradient-boosted alternatives [13-14]. General reviews of imbalanced learning show that class distributions, overlap, small disjuncts, and asymmetric errors can all alter model behavior [15]. Data-level remedies such as SMOTE are influential [16], but synthetic or resampled training distributions can distort the relationship between scores and real event probabilities. XOR-EW therefore uses algorithm-level class weighting or balanced subsampling for model development and reserves an untouched chronological window for calibration (Figure 1).

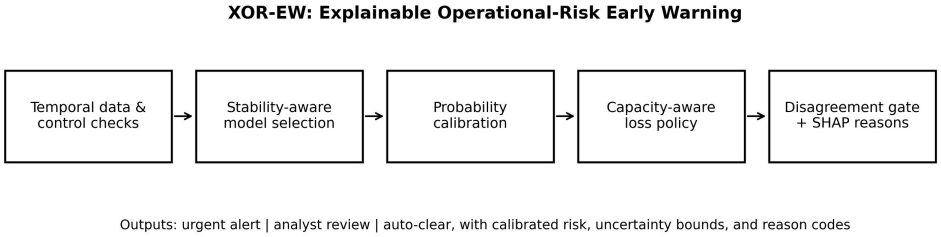

Figure 1 XOR-EW Decision and Governance Architecture

For rare events, area under the precision-recall curve (AUPRC, also called average precision in this paper) is more informative than accuracy and often more operationally interpretable than AUROC [17-19]. Yet AUPRC is threshold-free; it cannot alone decide whether an alert is worth the investigation cost. Cost-sensitive learning provides the decision-theoretic foundation for asymmetric actions, and specific work has shown that transaction-level amounts can be incorporated into losses [9-11]. XOR-EW extends this idea with a transparent scenario cost and an explicit workload constraint rather than presenting cost as an implicit training weight.

2.2 Calibration, Explainability, and Governance

A probability used for escalation should mean approximately the same thing across score ranges and time windows. Studies of probability estimation and modern classifiers show that can coexist with poor calibration [20-21]. Platt scaling maps a raw score through a logistic calibration model [22]. In XOR-EW, the calibrator is fit on a dedicated chronological window; the policy threshold is then selected on a later window, preventing the same labels from serving both calibration and action optimization.

Post-hoc explanation methods provide local and global views of model behavior. SHAP supplies additive feature attributions [23], and TreeSHAP enables efficient, consistent explanations for tree ensembles [24]. LIME is a widely used model-agnostic alternative [25]. Broader XAI taxonomies stress that explanation quality depends on audience, fidelity, scope, and purpose [26]. In high-stakes settings, explanation cannot compensate for a poorly governed black box [27-28]. XOR-EW therefore treats SHAP as an audit and reason-code layer, not as proof of causality, and reports explanation stability under repeated transaction resampling [29-30].

2.3 Research Gap

The literature contains powerful elements but an operational gap remains. A bank needs a versioned chain from data controls to model choice, calibrated score, decision policy, analyst queue, explanation, and monitoring. The proposed framework addresses this chain and uses temporal controls motivated by concept-drift research [28]. Its novelty is architectural and decision-oriented: it combines established methods so that every transformation has a separate validation window, metric, and governance output.

3 XOR-EW FRAMEWORK

3.1 Temporal Data Control and Feature Layer

Transactions are sorted by event time after schema checks, missing-value checks, and exact duplicate removal. Four non-overlapping windows are created: training (50%), calibration (20%), policy selection (10%), and final testing (20%). The final test labels are never used for model choice, calibration, or threshold selection. The feature vector contains anonymized principal components V1-V28, log(1 + Amount), and sine/cosine encodings of time-of-day. The raw sequential Time value is excluded from prediction to avoid learning a trivial boundary tied to the split itself.

3.2 Stability-Aware Champion Selection

Each candidate model is fit only on the training window. Let AP_cal,m and AP_pol,m denote the AUPRC of model m on the calibration and policy windows. XOR-EW selects the model maximizing the following temporal-robustness score (TRS):

$$TRS_m = (AP_{cal,m} + AP_{pol,m})/2 - |AP_{cal,m} - AP_{pol,m}| \quad (1)$$

The first term rewards consistently high rare-event ranking before the test period, while the penalty discourages a model whose apparent quality changes sharply between adjacent windows. TRS is intentionally simple and auditable. It is not a statistical guarantee; it is a governance rule that makes temporal stability explicit and prevents test-set-driven champion selection.

3.3 Probability Calibration

Let s(x) be the champion probability-like score. A logistic calibrator is fit on the calibration window using the logit of the clipped score. The calibrated probability is

$$\hat{p}(x) = \mathrm{sigmoid}[a \cdot \log(s(x)/(1-s(x))) + b]. \quad (2)$$

Calibration is evaluated on the final test window with the Brier score and ten-bin equal-frequency expected calibration error (ECE). Equal-frequency bins are used because prevalence is approximately one event per thousand transactions; fixed-width bins would leave most bins empty.

3.4 Exposure- and Capacity-Aware Policy

The policy window converts calibrated probabilities into an action threshold. A false alert consumes one investigation unit. A missed costs one unit plus its transaction amount divided by the median training amount. This dimensionless scenario is deliberately not labeled as an accounting loss; institutions should replace the coefficients with validated internal costs. For threshold t,

$$C(t) = \sum_{FP(t)} 1 + \sum_{FN(t)} [1 + \mathrm{Amount}_i / \mathrm{median}_{train}(\mathrm{Amount})]. \quad (3)$$

XOR-EW chooses the threshold minimizing C(t) on the policy window subject to an alert-rate limit kappa. The main experiment sets kappa = 0.5%. A normalized exposure cost divides test cost by the all-clear policy cost, so values below one indicate improvement over clearing every transaction.

3.5 Ensemble-Disagreement Triage

A point estimate can hide model instability near the threshold. The selected random forest contains 260 trees, partitioned deterministically into 13 groups of 20. Each group produces a probability that is transformed by the same calibration model. The 10th and 90th percentiles form a disagreement interval [L(x), U(x)]. Transactions with $\hat{p}(x) \geq t$ are urgent alerts. Transactions below t whose interval crosses t are sent to analyst review; all others are auto-cleared:

$$\mathrm{review}(x) = 1[\hat{p}(x) < t \text{ and } L(x) < t \leq U(x)]. \quad (4)$$

This interval is an ensemble disagreement diagnostic, not a formal confidence interval. Its purpose is selective triage: expose a small, uncertain boundary region to human judgment without inflating the urgent-alert queue.

3.6 Explanation and Monitoring Layer

For each alert, TreeSHAP produces signed feature contributions and the top six absolute contributors become local reason codes. Global importance is the mean absolute SHAP value over all and a random sample of legitimate test transactions. Explanation stability is evaluated by bootstrapping this SHAP matrix 20 times and reporting Spearman rank correlation and top-ten Jaccard overlap. Population stability index (PSI) is also computed between training and test features and between raw risk-score distributions. PSI is treated as a diagnostic trigger for investigation, not as proof of performance degradation.

3.7 Decision Trace and Governance Outputs

A warning is useful only when its origin can be reconstructed. XOR-EW therefore defines the decision record for transaction i as D_i = {data version, event time, model version, raw score, calibrator version, calibrated probability, policy version, threshold, disagreement bounds, route, and reason codes}. The route takes one of three mutually exclusive values: urgent alert, analyst review, or auto-clear. This record separates predictive evidence from the action rule. A model can retain its ranking while a cost policy changes, or a policy can remain fixed while recalibration changes the probability scale; the trace makes either change visible to validation and audit teams.

Governance artifacts are produced at four levels. The data layer records row counts, hashes, schema checks, duplicate handling, and chronological boundaries. The model layer records the candidate set, random seed, package versions, pre-test scores, and champion rule. The policy layer records calibration coefficients, threshold search, capacity limit, cost definition, and queue volumes. The monitoring layer records calibration, ranking, workload, PSI, explanation stability, and realized outcomes when labels arrive. Each artifact has an explicit interpretation boundary: disagreement quantifies ensemble dispersion rather than statistical coverage; SHAP describes model attribution rather than causality; and the scenario cost supports controlled comparison rather than financial reporting.

This decomposition also supports champion-challenger governance. A challenger can be evaluated without changing the production policy, and a new policy can be replayed on frozen scores without retraining the model. Approval gates should therefore be attached to changes in data, model, calibrator, or policy as separate objects. Such modularity is important in financial institutions because model developers, operational-risk owners, investigators, independent validators, and internal audit may require different evidence from the same decision chain.

4 EXPERIMENTAL DESIGN

4.1 Data Provenance and Scope

The official OpenML record describes 284,807 European card transactions collected over two days in September 2013, with 492, V1-V28 PCA components, Time, Amount, and Class [6]. Direct large-file access was inconsistent in the execution environment, so the experiment used two downloadable CSV files from a public repository mirror. Combined, the mirror contained 227,844 rows, approximately 80% of the official record, and spanned the official time range. Its omitted-row selection mechanism is undocumented. Accordingly, this paper reports a public subset/mirror experiment and does not claim results on the complete original file.
Table 1 reports the strict chronological split. rates differ across windows, particularly in the policy and test periods, providing a realistic stress on temporal selection and calibration. No missing values were present.

Table 1 Chronological Data Partition

Window	Rows	Fraud	Rate (%)	Time min	Time max
Train	113,551	214	0.1885	0	84,796
Calibration	45,420	86	0.1893	84,796	132,985
Policy	22,710	24	0.1057	132,985	145,267
Test	45,421	57	0.1255	145,269	172,792

4.2 Models and Reproducibility

Four candidates were evaluated: class-weighted logistic regression with standardized features; a balanced random forest with 260 trees, maximum depth 14, and minimum leaf size 3; XGBoost with 380 depth-4 trees, learning rate 0.045, subsampling and column sampling of 0.85; and LightGBM with 420 trees, learning rate 0.035, 31 leaves, and minimum child size 40. Positive-class weights for boosting were computed only from training prevalence. Hyperparameters were fixed before inspecting the final test outputs; no claim of exhaustive optimization is made.
All experiments used seed 2023. The supplied script records package versions, platform information, thresholds, split statistics, and runtime in JSON. It caches trained models but deterministically regenerates metrics, figures, explanation files, and triage details. Tests ran with Python 3.13.5, pandas 2.2.3, NumPy 2.3.5, scikit-learn 1.8.0, XGBoost 3.1.3, LightGBM 4.6.0, and SHAP 0.50.0. The cached rerun completed in 10.0 s; model-training time will vary by hardware.

4.3 Evaluation Protocol

Ranking metrics are AUROC and AUPRC. Threshold metrics are precision, recall, F1, Matthews correlation coefficient (MCC), alert count, and confusion-matrix counts. Probability metrics are Brier score and ECE. Decision metrics are normalized exposure cost and workload rate. Candidate thresholds for each baseline are selected by F1 on the policy window; XOR-EW uses the exposure/capacity objective. This distinction is reported explicitly because comparing different threshold objectives can otherwise create misleading claims.

4.4 Leakage Controls and Reproduction Checks

The experimental script enforces window roles rather than relying on manual discipline. Model fitting reads only the training indices. Platt parameters are estimated only from calibration labels. The stability score and all candidate thresholds use only calibration and policy outputs, respectively. Final-test labels enter only the reporting functions after the champion, calibrator, and policy are fixed. Raw Time is not a predictor, and preprocessing statistics for logistic regression are estimated within its training pipeline. These controls reduce three common sources of optimistic bias: random temporal mixing, reuse of calibration labels for threshold selection, and normalization on future observations.
Reproduction is checked at both input and output levels. The data manifest stores byte size and SHA-256 digest for each downloaded CSV, while the split summary exposes row counts, class counts, prevalence, and time bounds. The run metadata stores the seed, software versions, champion, thresholds, calibration parameters, and execution time. Model comparison, calibration, ablation, capacity sensitivity, disagreement triage, drift, global SHAP importance, local reason codes, and bootstrap stability are emitted as separate CSV or JSON files. Thus every table and figure in the paper can be traced to a machine-readable result rather than being transcribed from an unrecorded notebook session.

5 RESULTS

5.1 Stability-Aware Model Selection

Table 2 shows that XGBoost achieved the highest test AUROC (0.957), while the balanced random forest achieved the highest pre-test temporal robustness (0.814). Weighted logistic regression achieved the highest test F1 among the four model-specific F1 policies (0.808) and a low normalized scenario cost (0.361), but its Brier score (0.0498) and ECE (0.0940) were orders of magnitude worse than the tree ensembles, and its selected threshold was numerically near one. Therefore, XOR-EW did not select models from final-test F1; it selected the random forest using only adjacent pre-test AUPRC values. This result illustrates the difference between a governance champion and a single-metric winner (Figure 2).

Table 2 Candidate Model Results; Test Thresholds Maximize Policy-Window F1

Model	APcal	APpol	TRS	APtest	AUC	F1	Cost
RF	0.817	0.815	0.814	0.778	0.944	0.719	0.731
Logit	0.806	0.773	0.756	0.774	0.936	0.808	0.361
XGB	0.774	0.821	0.751	0.755	0.957	0.755	0.400
LGBM	0.734	0.784	0.709	0.699	0.904	0.758	0.511

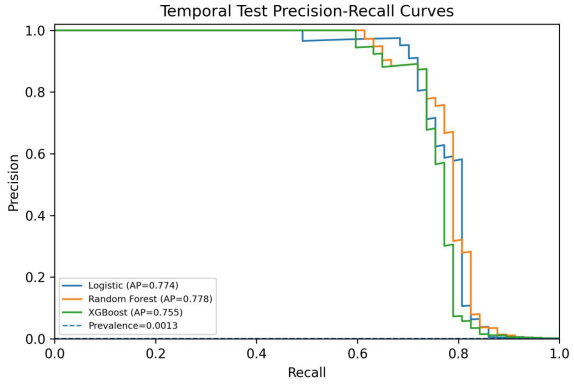

Figure 2 Precision-Recall Curves on the Chronological Test Window

5.2 Calibration and Policy Ablation

Platt calibration preserved ranking, as expected, but improved probability quality. The Brier score decreased by 17.66%, from 0.000511 to 0.000421, and ECE decreased by 91.87%, from 0.003736 to 0.000304. Figure 3 shows that the calibrated curve is closer to the ideal relationship across equal-frequency score bins. These improvements matter because the downstream threshold is interpreted as a risk policy rather than merely a rank cutoff.

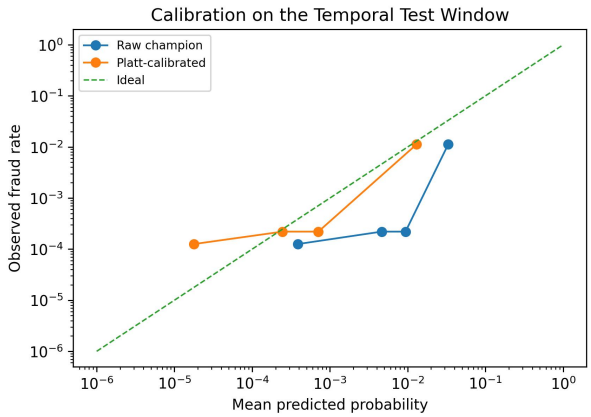

Figure 3 Raw and Platt-Calibrated Reliability on the Test Window (Log-Log Axes)

Table 3 Framework Ablation on the Test Window

Variant	Prec.	Rec.	F1	Brier	ECE	Cost	Alerts
Raw + F1	1.000	0.561	0.719	0.000511	0.003736	0.731	32
Cal. + F1	1.000	0.561	0.719	0.000421	0.000304	0.731	32
XOR-EW	0.840	0.737	0.785	0.000421	0.000304	0.358	50

Table 3 isolates the effect of calibration and policy choice. The test alert rate is 0.110%, substantially below the 0.5% policy constraint because the cost optimum occurs before capacity is exhausted.

5.3 Capacity Sensitivity and Selective Triage

Capacity sensitivity did not behave monotonically because the policy objective minimizes exposure cost under an upper bound; a looser bound does not force more alerts. At caps of 0.2%, 0.5%, and 1.0%, test F1 values were 0.800, 0.785, and 0.768, while normalized costs were 0.353, 0.358, and 0.536. The 0.2% setting happened to generalize slightly better on the test window, but the primary 0.5% setting was fixed in advance and is retained to avoid test-based policy selection (Table 4).

Table 4 Capacity Sensitivity (Thresholds Selected on Policy Window)

Cap	Alerts	Prec.	Rec.	F1	Cost
0.2%	0.106%	0.875	0.737	0.800	0.353
0.5%	0.110%	0.840	0.737	0.785	0.358
1.0%	0.092%	0.905	0.667	0.768	0.536

Urgent alerts plus review captured 44 of 57 (77.19%). Total urgent-plus-review workload was 58 transactions, or 0.1277% of the test window, leaving 99.8723% auto-cleared. Figure 4 makes the operational trade-off visible: the review queue is small relative to the urgent queue and the complete stream.

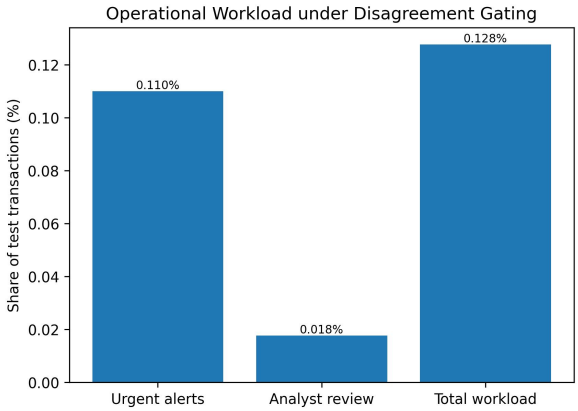

Figure 4 Workload after Urgent-Alert and Disagreement-Review Routing

5.4 Explanations, Stability, and Drift Diagnostics

Global TreeSHAP importance ranked V14, V12, V10, V4, V17, and V11 as the six leading anonymous components. Because the source features are PCA-transformed for confidentiality, they cannot be translated into causal business statements. Local reason-code files therefore report feature identifier, feature value, signed contribution, raw probability, and rank, enabling case-level audit without inventing semantic labels (Figure 5).

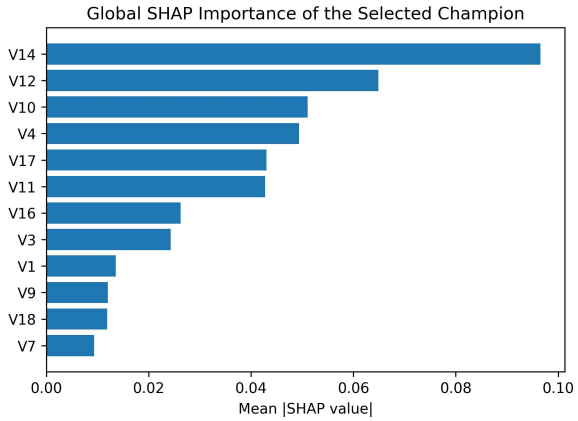

Figure 5 Mean Absolute SHAP Importance for the Selected Random Forest

Across 20 bootstrap resamples of the explanation matrix, mean Spearman rank correlation was 0.989 (standard deviation 0.004), and mean top-ten Jaccard overlap was 0.927 (standard deviation 0.091). This supports sampling stability of the reported global ranking, although it does not establish causal validity or robustness to feature perturbations.

PSI diagnostics identified large distribution changes in time-of-day encodings and several anonymized components; the raw risk-score PSI was 0.390. Given the two-day horizon and incomplete mirror, these values are descriptive rather than

universal warning thresholds. They nonetheless reinforce why the champion was selected from chronological windows and why a production implementation should monitor score and feature distributions alongside realized loss outcomes.

5.5 Operational Error Routing and Auditability

For operational use, the queues should be monitored separately. Urgent-alert precision indicates the immediate investigative yield, whereas review yield measures the value of deferral. The auto-clear population should be sampled for quality assurance and reconciled when delayed labels arrive. A rising review rate with unchanged urgent-alert rate may indicate increasing model disagreement near the policy boundary; a rising urgent rate may instead reflect score shift, recalibration, policy change, or genuine loss escalation. Because XOR-EW stores score, bounds, route, and reason codes for each case, these hypotheses can be tested rather than inferred from aggregate alert counts alone.

The result files implement this traceability. The row-level triage file includes calibrated probability, lower and upper disagreement bounds, route indicators, and observed class for all 45,421 test transactions. The local-explanation file supplies ranked signed contributions for selected cases, while the global and bootstrap files support model-level review. These artifacts permit independent recomputation of queue counts and explanation summaries, and they make negative results—such as remaining false negatives—available for challenge rather than hiding them behind a single headline metric.

6 DISCUSSION

6.1 What the Results Demonstrate

XGBoost had the highest AUROC, logistic regression had the highest F1 under its own F1 policy, and the random forest was the most stable across pre-test windows. A transparent selection rule prevents retrospective metric shopping. Second, probability calibration is valuable even when ranking does not change. Third, an operational policy can trade a small number of extra investigations for materially greater event capture and lower exposure-weighted scenario cost.

The proposed design is especially useful for model governance. Each artifact has a clear owner and validation question: data control asks whether the time ordering and schema are valid; model selection asks whether ranking is temporally stable; calibration asks whether probabilities are meaningful; policy selection asks whether actions respect exposure and capacity; disagreement asks when the model should defer; explanations ask which features drove each score; and monitoring asks whether inputs or scores changed. This decomposition aligns technical evidence with operational-risk and model-risk governance expectations [1-5].

6.2 Comparison with Simpler Alternatives

A simpler logistic model deserves serious consideration in high-stakes settings [27]. In this experiment, weighted logistic regression achieved F1 = 0.808 and normalized cost = 0.361, close to XOR-EW cost = 0.358. Its poor uncalibrated probability metrics and near-one threshold were weaknesses, but a separately calibrated logistic pipeline could be a strong challenger. XOR-EW's advantage is not that a random forest dominates every metric; it is that the framework makes stability, calibration, action cost, uncertainty routing, and explanations jointly testable. Institutions should retain interpretable baselines and reject added complexity that does not produce governance or decision value.

6.3 Limitations and Threats to Validity

This study has important limitations. The accessible mirror is not the complete official dataset, and the omitted-row mechanism is unknown. The record covers only two days in 2013, so it cannot establish long-term performance. PCA anonymity prevents business-semantic reason codes and fairness analysis over protected attributes. is only one observable proxy for operational risk; the framework has not been tested on cyber incidents, processing errors, conduct events, outages, or internal loss databases. The exposure function is a dimensionless scenario, not a bank's validated accounting cost. Labels are treated as immediately available, whereas real feedback can be delayed. The ensemble disagreement interval is heuristic and should not be interpreted as statistical coverage. Finally, no live analyst study, latency benchmark, adversarial test, or external institutional validation was performed.

These limitations constrain the claim: XOR-EW is a reproducible research framework demonstrating how governance layers can be evaluated together. Before production use, a financial institution would need representative multi-period data, validated cost coefficients, label-delay treatment, reason-code mapping to non-anonymous features, subgroup and fairness review, challenger testing, independent validation, security controls, and post-deployment monitoring.

6.4 Deployment and Validation Blueprint

A controlled deployment should begin with shadow operation. Scores and proposed routes would be generated without changing customer or investigator actions, allowing calibration, latency, workload, and reason-code usefulness to be measured against the incumbent process. The next phase could expose only the review queue to analysts while preserving existing urgent rules. Promotion of the new urgent policy should require predefined acceptance criteria for

capture, false-alert burden, stability across time and business segments, operational resilience, and investigator feedback. Rollback must be possible at the policy layer without deleting the underlying score history.

Independent validation should reproduce the chronological split, challenge the cost coefficients and capacity assumption, test alternative calibrators and interpretable challengers, and examine sensitivity to delayed or corrected labels.. Where protected or customer-impact attributes are available, subgroup performance and potential disparate effects require separate analysis; the anonymous public data used here cannot support that assessment.

Post-deployment monitoring should distinguish leading and lagging indicators. Leading indicators include input quality, feature and score drift, calibration proxy checks, alert volume, disagreement volume, latency, and reason-code availability. Lagging indicators include confirmed event capture, exposure-weighted loss, false-alert handling time, analyst overrides, and control failures. A breach should trigger a defined response—investigation, recalibration, policy rollback, challenger activation, or model redevelopment—rather than an automatic retraining loop. This human-governed lifecycle is the intended operational meaning of early warning in XOR-EW.

7 CONCLUSION

XOR-EW integrates temporal model selection, calibrated probabilities, exposure- and capacity-aware action, uncertainty-based deferral, stable SHAP explanations, and drift diagnostics for operational-risk early warning. On a real public transaction subset, the framework selected a temporally stable random forest without using final-test labels, improved calibration, raised recall from 56.14% to 73.68%, and reduced normalized exposure cost from 0.731 to 0.358 relative to the raw champion's F1 policy. A small disagreement-review queue increased total capture to 77.19% while total workload remained 0.128%. The main contribution is not a new classifier but a reproducible, auditable decision architecture. Future work should evaluate multi-month institutional loss data, delayed feedback, calibrated interpretable challengers, formal selective-risk guarantees, and human-centered reason-code utility.

COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

REFERENCES

[1] Basel Committee on Banking Supervision. Principles for the Sound Management of Operational Risk. Basel, Switzerland: Bank for International Settlements, 2011.
[2] Basel Committee on Banking Supervision. Revisions to the Principles for the Sound Management of Operational Risk. Basel, Switzerland: Bank for International Settlements, 2021.
[3] Board of Governors of the Federal Reserve System, Office of the Comptroller of the Currency. Supervisory Guidance on Model Risk Management, SR Letter 11-7 / OCC Bulletin 2011-12, 2011.
[4] Tabassi E. Artificial Intelligence Risk Management Framework (AI RMF 1.0), NIST AI 100-1. Gaithersburg, MD, USA: National Institute of Standards and Technology, 2023.
[5] Phillips P J, Hahn C, Fontana P, et al. Four Principles of Explainable Artificial Intelligence, NISTIR 8312. National Institute of Standards and Technology, 2021.
[6] Dal Pozzolo A, Caelen O, Johnson R A, et al. Calibrating probability with undersampling for unbalanced classification. In: 2015 IEEE Symposium Series on Computational Intelligence. Cape Town, South Africa, 2015: 159-166.
[7] Carcillo F, Dal Pozzolo A, Le Borgne Y A, et al. SCARFF: A scalable framework for streaming credit card detection with Spark. Information Fusion, 2018, 41: 182-194.
[8] Carcillo F, Le Borgne Y A, Caelen O, et al. Combining unsupervised and supervised learning in credit card detection. Information Sciences, 2021, 557: 317-331.
[9] Bahnsen A C, Aouada D, Stojanovic A, et al. Feature engineering strategies for credit card detection. Expert Systems with Applications, 2016, 51: 134-142.
[10] Bahnsen A C, Aouada D, Ottersten B. Example-dependent cost-sensitive decision trees. Expert Systems with Applications, 2015, 42(19): 6609-6619.
[11] Sahin Y, Bulkan S, Duman E. A cost-sensitive decision tree approach for detection. Expert Systems with Applications, 2013, 40(15): 5916-5923.
[12] Breiman L. Random forests. Machine Learning, 2001, 45: 5-32.
[13] Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, CA, USA, 2016: 785-794.
[14] Ke G, Meng Q, Finley T, et al. LightGBM: A highly efficient gradient boosting decision tree. In: Advances in Neural Information Processing Systems, 2017: 3146-3154.
[15] He H, Garcia E A. Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 2009, 21(9): 1263-1284.
[16] Chawla N V, Bowyer K W, Hall L O, et al. SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 2002, 16: 321-357.
[17] Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLOS ONE, 2015, 10(3): e0118432.

[18] Davis J, Goadrich M. The relationship between precision-recall and ROC curves. In: Proceedings of the 23rd International Conference on Machine Learning. Pittsburgh, PA, USA, 2006: 233-240.
[19] Fawcett T. An introduction to ROC analysis. Pattern Recognition Letters, 2006, 27(8): 861-874.
[20] Niculescu-Mizil A, Caruana R. Predicting good probabilities with supervised learning. In: Proceedings of the 22nd International Conference on Machine Learning. Bonn, Germany, 2005: 625-632.
[21] Guo C, Pleiss G, Sun Y, et al. On calibration of modern neural networks. In: Proceedings of the 34th International Conference on Machine Learning, 2017: 1321-1330.
[22] Platt J C. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In: Smola A J, Bartlett P, Schölkopf B, et al., eds. Advances in Large Margin Classifiers. Cambridge, MA, USA: MIT Press, 1999: 61-74.
[23] Lundberg S M, Lee S I. A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems 30, 2017: 4765-4774.
[24] Lundberg S M, Erion G, Chen H, et al. From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence, 2020, 2: 56-67.
[25] Ribeiro M T, Singh S, Guestrin C. "Why should I trust you?": Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, CA, USA, 2016: 1135-1144.
[26] Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, et al. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 2020, 58: 82-115.
[27] 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.
[28] Webb G I, Hyde R, Cao H, et al. Characterizing concept drift. Data Mining and Knowledge Discovery, 2016, 30(4): 964-994.
[29] Alvarez-Melis D, Jaakkola T S. On the robustness of interpretability methods. arXiv:1806.08049, 2018.
[30] Elkan C. The foundations of cost-sensitive learning. In: Proceedings of the 17th International Joint Conference on Artificial Intelligence. Seattle, WA, USA, 2001: 973-978.