

Hybrid Explainable AI Framework for Predictive Operational Risk Management in Financial Services

Harshendra Gite,

Senior Director, LTIMindtree Ltd., Greensboro, NC, USA, hgite@proton.me,

ORCID: 0009-0000-1577-4322

Abstract

The digital transformation of financial services has increased organizational dependence on complex information systems, automated processes, third-party service providers, algorithmic decision systems, and interconnected digital infrastructures. Although these developments improve efficiency and customer service, they simultaneously amplify operational exposures arising from process failures, internal and external fraud, cyber incidents, system disruptions, human error, regulatory non-compliance, and third-party dependencies. Traditional operational risk management approaches remain predominantly retrospective and frequently depend on historical loss databases, risk and control self-assessments, key risk indicators, and expert judgment. Such mechanisms are valuable for governance but are often insufficient for predicting rapidly evolving and low-frequency operational risk events. This study proposes a Hybrid Explainable Artificial Intelligence Framework for Predictive Operational Risk Management (HXAI-ORM) that integrates supervised machine learning, anomaly detection, interpretable statistical modelling, explainable artificial intelligence, and human-in-the-loop governance. The framework combines logistic regression, Random Forest, XGBoost, and Isolation Forest outputs to produce probability-, severity-, and anomaly-adjusted operational risk scores. SHapley Additive exPlanations (SHAP), Local Interpretable Model-Agnostic Explanations (LIME), rule-based explanations, and counterfactual reasoning are incorporated to translate predictive outputs into auditable risk drivers and actionable control recommendations. The framework is aligned with the Basel Committee's principles for operational risk and operational resilience and with the NIST Artificial Intelligence Risk Management Framework. Unlike conventional black-box prediction architectures, the proposed framework explicitly incorporates explanation fidelity, probability calibration, model stability, human review, model-risk controls, and continuous monitoring. The study contributes to the literature by repositioning explainability from a supplementary visualization mechanism to an integral component of predictive operational risk governance. A structured empirical validation protocol is also developed for subsequent implementation with financial-institution data. The framework offers financial institutions a practical foundation for moving from retrospective loss reporting toward anticipatory, transparent, and governance-compatible operational risk management.

Keywords: operational risk; explainable artificial intelligence; financial services; machine learning; predictive risk management; SHAP; XGBoost; anomaly detection; operational resilience; model risk management

1. Introduction

Operational risk has become increasingly important to financial institutions as banking activities have become more digital, automated, interconnected, and dependent on information and communication technologies. Operational risk generally arises from inadequate or failed internal processes, people and systems, or from external events, and encompasses diverse events including fraud, business disruption, technology failures, execution errors, inappropriate business practices, and damage to physical assets (Basel

Committee on Banking Supervision [BCBS], 2021a). The growing reliance of financial institutions on digital platforms, cloud infrastructure, application programming interfaces, artificial intelligence (AI), outsourced services, and real-time payments has expanded both the scale and complexity of operational risk.

Traditional operational risk management generally combines internal loss-event databases, risk and control self-assessments, scenario analysis, key risk indicators, audit findings, and managerial judgment. Although

these mechanisms remain fundamental to effective governance, many are inherently retrospective. Risk managers may therefore recognize patterns only after incidents have accumulated sufficiently to become visible through threshold-based reporting. This creates a gap between contemporary high-frequency operational environments and conventional periodic risk assessment processes.

The growing availability of granular transactional, process, employee, technology, control, vendor, customer, and incident data offers an opportunity to shift operational risk management from predominantly retrospective identification toward predictive early-warning systems. Machine learning (ML) is particularly promising because it can identify nonlinear interactions and complex dependencies that are difficult to specify through conventional statistical models (Breiman, 2001; Chen & Guestrin, 2016). AI and ML have consequently received substantial attention throughout financial services, including risk assessment, fraud detection, compliance, credit decisions, surveillance, portfolio management, and automated customer interactions (Financial Stability Board [FSB], 2017; Milana & Ashta, 2021; Cao & Zhai, 2022).

Nevertheless, predictive accuracy alone is insufficient in a highly regulated financial environment. Complex algorithms can generate predictions whose underlying logic is difficult for risk managers, model validators, internal auditors, senior executives, and regulators to evaluate. The interpretability problem becomes particularly significant when algorithmic outputs trigger investigations, control interventions, customer restrictions, employee scrutiny, capital allocations, or regulatory reporting. Research on explainable artificial intelligence (XAI) emphasizes that understandable prediction logic is essential for transparency, accountability, validation, and appropriate reliance on AI-assisted decisions (Carvalho et al., 2019; Arrieta et al., 2020; Rudin, 2019).

Explainability requirements in finance are also stakeholder-dependent. A data scientist may require detailed feature attribution and diagnostic information, whereas a business manager may need a short explanation linking an alert to specific operational drivers. Internal auditors require evidence of traceability and control effectiveness, while senior management may be primarily concerned with emerging concentrations and aggregate exposure. Hadji Misheva et al. (2021) emphasize that explanation mechanisms

should therefore reflect differences in stakeholder information requirements rather than assuming a single universal form of explanation.

Existing research on XAI in finance largely concentrates on credit risk, lending, fraud detection, investment decisions, and financial-market applications. Bussmann et al. (2020), for example, demonstrate how Shapley-value-based explanations can clarify predictive credit-risk models. Much less attention has been devoted to developing an integrated explainable framework specifically for enterprise operational risk prediction, where data are heterogeneous, events may be rare, risk categories differ substantially, causality is difficult to establish, and regulatory governance is essential.

The present study addresses this gap through a Hybrid Explainable Artificial Intelligence Framework for Predictive Operational Risk Management (HXAI-ORM). Instead of depending upon a single algorithm, HXAI-ORM integrates an interpretable statistical model, supervised ensemble learning, unsupervised anomaly detection, explanation methods, probability calibration, risk severity, control information, and human judgment.

The study addresses three research questions. First, how can heterogeneous financial-services data be combined to provide reliable early-warning predictions of operational risk? Second, how can model accuracy and complexity be reconciled with the explainability, traceability, and model-governance requirements of financial institutions? Third, how can prediction and explanation outputs be integrated into operational risk processes so that AI generates actionable risk-management decisions rather than merely analytical scores?

The principal contribution is a governance-integrated hybrid architecture in which prediction, anomaly detection, explainability, validation, and human escalation constitute a unified operational-risk decision process. In contrast to approaches where XAI is applied after modelling merely to visualize a black-box prediction, the framework treats explanation quality as one of the formal dimensions of model performance.

2. Literature Review

Operational risk is structurally different from conventional credit or market risk. Credit risk is generally linked to counterparty default, while market

risk emerges primarily from adverse changes in market variables. Operational losses, in contrast, can originate from diverse interactions among processes, personnel, technologies, external actors, legal obligations, and organizational controls.

The Basel framework distinguishes seven broad categories of operational loss events: internal fraud;

external fraud; employment practices and workplace safety; clients, products, and business practices; damage to physical assets; business disruption and system failures; and execution, delivery, and process management. This taxonomy demonstrates that operational-risk prediction cannot be reduced to conventional transaction-fraud modelling alone.

Table 1. Operational-risk categories and candidate predictive indicators

Operational-risk category	Representative events	Potential predictive indicators	Appropriate analytical signal
Internal fraud	Unauthorized transactions, asset misappropriation, intentional policy circumvention	Privileged-access frequency, unusual employee activity, override rates, transaction deviations	Supervised classification + anomaly detection
External fraud	Account takeover, payment fraud, identity abuse, cyber-enabled fraud	Login anomalies, transaction velocity, geolocation variation, device mismatch	XGBoost/Random Forest + Isolation Forest
Employment practices	Policy violations, workplace claims, employee-related incidents	Turnover, overtime, unresolved complaints, policy exceptions	Statistical and tree-based classification
Clients, products, and business practices	Mis-selling, disclosure failure, unsuitable products, regulatory violations	Complaint growth, exception rates, product complexity, compliance breaches	Classification + text-derived indicators
Physical asset damage	Natural events, vandalism, facility incidents	Location exposure, environmental alerts, physical-security incidents	Scenario model + external indicators
Business disruption and system failures	Application outages, cyber disruption, infrastructure failures	Latency, system errors, failed jobs, capacity utilization, incident frequency	Time-series anomaly detection
Execution, delivery, and process management	Processing failures, reconciliation errors, missed deadlines, vendor failures	Manual-adjustment rate, reconciliation breaks, failed transactions, SLA breaches	Ensemble classification + process analytics

The 2021 BCBS revisions emphasize governance, risk identification, monitoring, change management, information and communication technology, and appropriate control environments. Closely associated principles for operational resilience emphasize the

capacity of banks to continue critical operations during severe but plausible disruptions (BCBS, 2021a, 2021b). Predictive operational-risk systems should therefore support resilience and control decisions rather than operate solely as loss-prediction mechanisms.

Machine learning enables predictive algorithms to learn complex relationships from historical observations without requiring analysts to specify every interaction beforehand. Random Forest, introduced by Breiman (2001), combines multiple decision trees to improve predictive stability and resistance to noise. Gradient-boosting models sequentially combine weak learners to reduce predictive error, while XGBoost provides a computationally scalable implementation capable of handling sparse and heterogeneous data (Chen & Guestrin, 2016).

Operational-risk modelling presents several challenges particularly suitable for hybrid learning. First, material operational incidents are often rare relative to normal observations. Rare-event classification can result in deceptively high accuracy when a model simply predicts the majority class (King & Zeng, 2001). Methods such as class weighting, cost-sensitive learning, threshold optimization, and carefully controlled resampling can therefore be required. Synthetic Minority Over-sampling Technique (SMOTE) is one established approach for generating synthetic minority observations, although temporal leakage and unrealistic synthetic cases must be controlled when it is applied to financial data (Chawla et al., 2002).

Second, previously unseen operational incidents cannot always be represented adequately within supervised training labels. Unsupervised anomaly detection can complement supervised prediction by identifying unusual observations without requiring confirmed historical labels. Isolation Forest, for instance, detects anomalies based on the principle that unusual observations are generally easier to isolate using random partitions than normal observations (Liu et al., 2008).

Third, probabilities matter in risk management. A model may classify events correctly while producing unreliable probabilities. Calibration techniques are therefore important when predicted probabilities determine escalation thresholds or expected-loss estimates (Niculescu-Mizil & Caruana, 2005).

These characteristics motivate a hybrid approach that combines supervised event prediction, unsupervised anomaly identification, severity estimation, and interpretable baseline models.

Explainable artificial intelligence seeks to make machine-generated outcomes understandable to

humans. XAI methods can be categorized according to whether the underlying model is inherently interpretable or whether explanations are produced after predictions through post-hoc techniques (Carvalho et al., 2019; Arrieta et al., 2020; Linardatos et al., 2021).

LIME approximates the behaviour of a complex predictive model in the local region surrounding a particular observation using an interpretable surrogate model (Ribeiro et al., 2016). Its principal advantage is model independence; however, explanations can be sensitive to local sampling assumptions.

SHAP builds upon cooperative game theory and assigns individual variables contributions to a specific model prediction (Lundberg & Lee, 2017). SHAP supports both local and global analysis and has been applied successfully to financial-risk prediction (Bussmann et al., 2020). For operational risk, SHAP can identify which variables increased or decreased the predicted probability of an incident.

Explainability nevertheless involves more than producing feature-importance plots. Rudin (2019) argues that high-stakes applications should employ interpretable models whenever suitably accurate interpretable alternatives are available rather than automatically relying on black-box systems. This observation motivates the inclusion of logistic regression and rule-based controls in HXAI-ORM. A complex model should be deployed only when it creates sufficient incremental predictive value to justify additional governance and model-risk costs.

Financial institutions require model development, validation, documentation, monitoring, change control, and effective challenge. The Federal Reserve and Office of the Comptroller of the Currency's model-risk guidance emphasizes rigorous development, validation, and governance surrounding models used in banking decisions (Board of Governors of the Federal Reserve System & Office of the Comptroller of the Currency, 2011).

The NIST AI Risk Management Framework similarly conceptualizes trustworthy AI through governance, mapping, measurement, and management functions and identifies characteristics including validity, reliability, safety, security, resilience, accountability, transparency, explainability, interpretability, privacy, and fairness (Tabassi, 2023).

The implication for operational-risk analytics is important: a technically strong prediction model does not automatically constitute a well-governed risk system. Institutions require evidence that data are appropriate, models are stable, predictions are explainable, outputs are monitored, human intervention is possible, and model limitations are understood.

3. Theoretical Foundation and Research Propositions

The proposed framework draws on three complementary perspectives: risk sensing, ensemble learning, and accountable AI governance.

Risk sensing assumes that operational incidents produce weak signals before formal loss recognition. These signals may emerge from transaction activity, process exceptions, employee behaviour, application performance, vendor service deterioration, complaints, reconciliation failures, audit observations, or control deficiencies. An early-warning system should combine these signals across traditionally separated organizational data sources.

Ensemble learning suggests that predictive robustness can improve when models with different inductive properties are combined. A linear statistical model provides transparency and a stable reference point; tree ensembles capture complex nonlinear relationships; and anomaly detection provides sensitivity to previously unseen patterns.

Accountable AI governance requires predictive intelligence to remain subject to human oversight and formal control. Consequently, the proposed model does not conceptualize AI as an autonomous operational-risk decision maker. It functions as an evidence-generating layer that supports accountable risk owners.

The first proposition is therefore that combining supervised, interpretable, and anomaly-based models will provide a more robust representation of operational-risk conditions than reliance on a single predictive model.

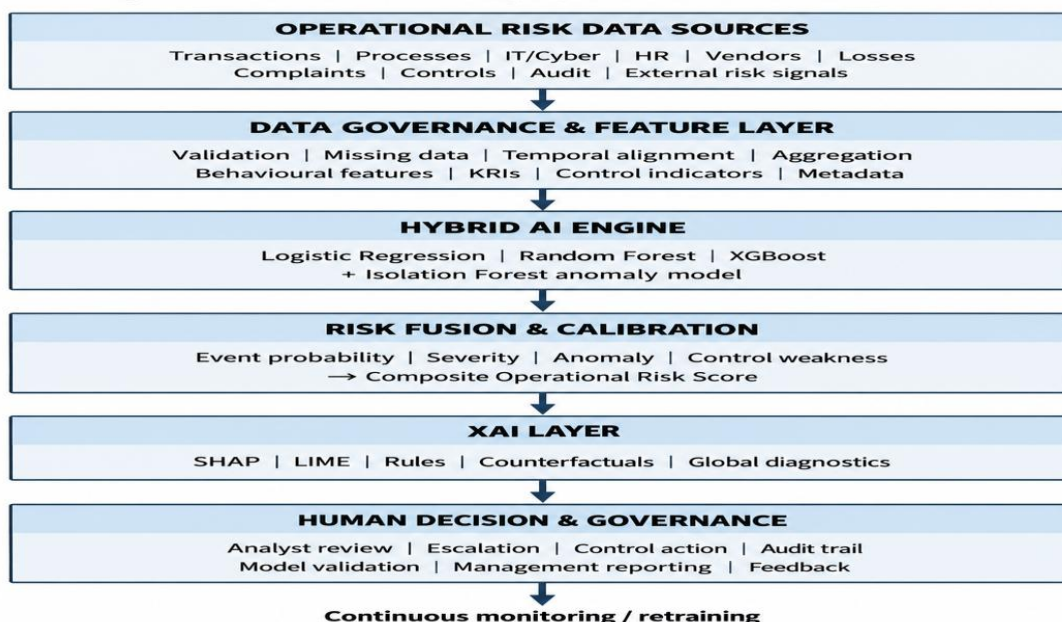
The second proposition is that explanation quality moderates the practical usefulness of predictive accuracy because operational-risk alerts that cannot be translated into understandable risk drivers are less suitable for investigation, escalation, independent validation, and regulatory review.

The third proposition is that human-in-the-loop governance increases the organizational value of operational-risk AI by translating prediction and explanation outputs into control decisions while providing feedback for model improvement.

4. Proposed Hybrid Explainable AI Framework

HXAI-ORM consists of six interconnected layers: data acquisition and governance, feature engineering, hybrid prediction, explanation, decision orchestration, and continuous monitoring.

Figure 1. Architecture of the proposed HXAI-ORM framework



The distinguishing characteristic of this architecture is that explainability and governance are included inside the predictive process rather than appended after deployment.

Table 2. Layers and responsibilities within HXAI-ORM

Layer	Core function	Representative techniques	Governance objective
Data layer	Integrate heterogeneous operational information	ETL, temporal alignment, entity resolution, quality checks	Data lineage and reliability
Feature layer	Generate predictive risk indicators	Rolling statistics, velocity features, process exceptions, control scores	Business relevance
Prediction layer	Estimate event probability and identify abnormal behaviour	Logistic regression, Random Forest, XGBoost, Isolation Forest	Predictive discrimination
Risk-fusion layer	Convert model outputs into decision-oriented risk scores	Probability calibration, severity weighting, ensemble aggregation	Consistent prioritization
Explainability layer	Identify prediction drivers	SHAP, LIME, interpretable rules, counterfactual analysis	Transparency and challenge
Governance layer	Convert outputs into accountable decisions	Human review, escalation, override, feedback, validation	Accountability and auditability

Operational-risk prediction requires data beyond historical losses because realized loss events represent only a fraction of the organization's evolving risk environment. The framework therefore distinguishes lagging and leading indicators.

Lagging indicators include recorded loss events, incidents, breaches, complaints, failed controls, audit findings, and historical outages. Leading indicators can include increasing manual overrides, rising transaction volumes, unusual access patterns, employee turnover, delayed reconciliations, infrastructure capacity, software defects, vendor service-level failures, unresolved control issues, and repeated near-miss events.

For entity i during period t , a feature vector may be represented as

$$X_{i,t} = [T_{i,t}, P_{i,t}, C_{i,t}, H_{i,t}, S_{i,t}, V_{i,t}, L_{i,t}],$$

where T represents transactional characteristics, P process indicators, C control measures, H human-

resource factors, S system and cybersecurity indicators, V vendor measures, and L historical loss information.

Time-aware construction is essential. Every predictive variable for time t must use only information that would have been available before the prediction date. Random train-test splitting should generally be avoided when it permits future information to influence past predictions.

The predictive layer combines models with complementary strengths.

Logistic regression provides an interpretable benchmark:

$$P(Y_{i,t} = 1|X_{i,t}) = \frac{1}{1 + \exp \left[- \left(\beta_0 + \sum_{j=1}^p \beta_j X_{j,i,t} \right) \right]}.$$

The logistic component helps identify whether complex algorithms materially outperform a transparent baseline and provides directional coefficient information useful for model validation.

Random Forest captures nonlinear relationships while reducing variance through aggregation across randomized decision trees (Breiman, 2001). XGBoost captures complex interactions through sequential gradient boosting and is particularly suitable for structured tabular financial data (Chen & Guestrin, 2016).

Isolation Forest operates independently of event labels and produces an anomaly score $A_{i,t}$. This enables the framework to detect unusual operational conditions even where historical labels are incomplete.

The supervised ensemble probability can be expressed as

$$P_{i,t}^H = w_L P_{i,t}^L + w_R P_{i,t}^{RF} + w_X P_{i,t}^{XGB},$$

where $w_L + w_R + w_X = 1$, and weights are estimated through validation performance rather than assigned arbitrarily.

A composite operational-risk score is subsequently calculated as

$$ORS_{i,t} = P_{i,t}^H \times S_{i,t} \times (1 + \lambda A_{i,t}) \times (1 + \gamma C_{i,t}),$$

where $S_{i,t}$ represents predicted or assigned severity, $A_{i,t}$ represents normalized anomaly intensity, $C_{i,t}$ represents control weakness, and λ and γ determine anomaly and control adjustments.

This formulation separates the probability of an event from the possible impact of that event. A high-frequency low-severity processing error and a rare but potentially catastrophic system outage therefore need not receive equivalent escalation priority.

Table 3. Model components and their roles in the hybrid framework

Model/component	Primary strength	Limitation	Role in HXAI-ORM
Logistic regression	Direct interpretability and stable benchmark	Limited nonlinear modelling	Transparent challenger/baseline
Random Forest	Robust nonlinear interactions	Less transparent than linear models	Main supervised ensemble component
XGBoost	Strong tabular predictive performance	Complexity and tuning requirements	Main nonlinear predictive engine
Isolation Forest	Detects unknown/unlabelled anomalies	Anomaly does not automatically imply loss	Novel-pattern detection
Probability calibration	Aligns scores with observed event likelihood	Requires representative validation data	Reliable risk thresholds
SHAP/LIME	Explains complex predictions	Explanation quality requires validation	Local and global explanation
Human risk review	Contextual and control expertise	Subjectivity and capacity constraints	Final escalation and feedback

The explainability layer produces both local and global outputs.

Global explanations identify recurring patterns across the portfolio, such as whether unresolved reconciliation breaks, unusual privileged access, declining system availability, increasing customer complaints, or vendor service failures are consistently associated with elevated risk.

Local explanations address individual alerts. For observation i , the SHAP decomposition can be written as

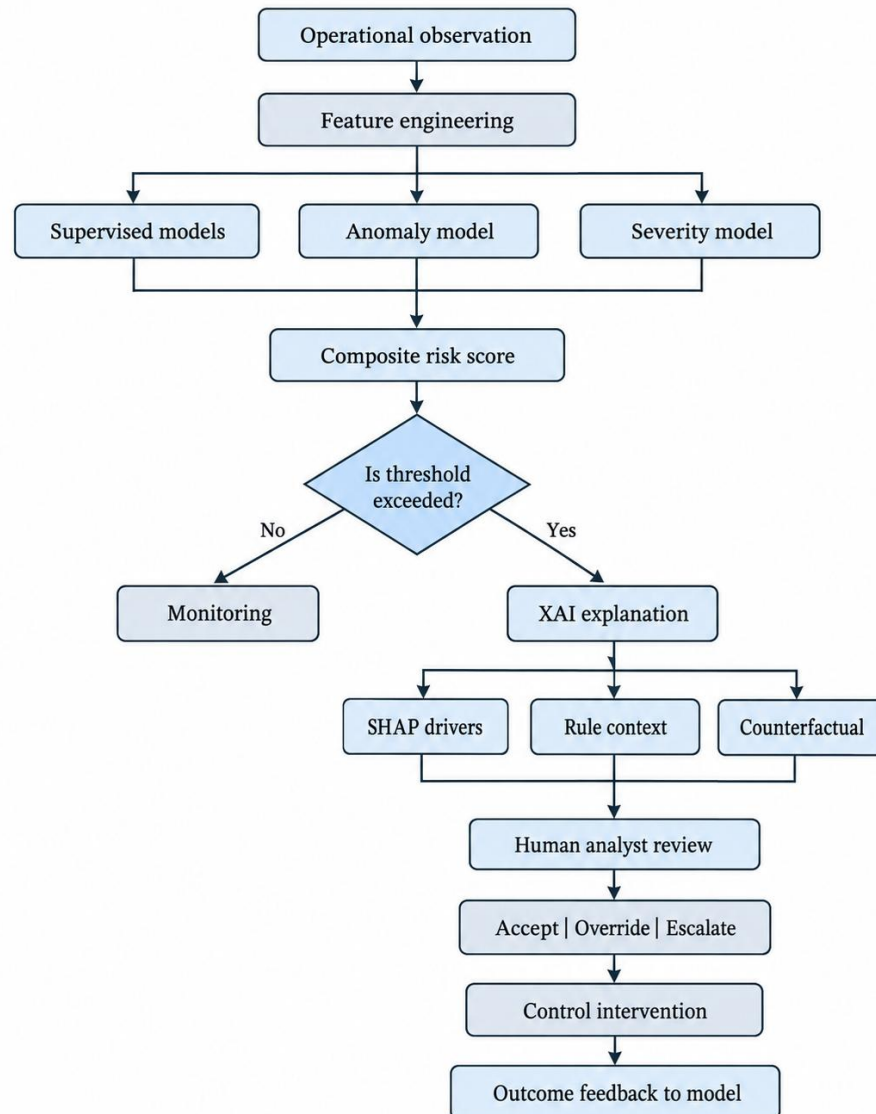
$$f(x_i) = \phi_0 + \sum_{j=1}^p \phi_{j,i},$$

where ϕ_0 represents the model's baseline output and $\phi_{j,i}$ represents the contribution of feature j to the particular prediction (Lundberg & Lee, 2017).

The explanation should not simply communicate that an alert has a risk probability of, for example, 0.78. It should communicate why. An operational-risk analyst might receive an explanation indicating that risk increased because failed reconciliations doubled during

the preceding seven days, privileged-access activity exceeded its historical range, a critical vendor repeatedly breached its SLA, and two relevant controls were overdue for remediation.

Figure 2. Prediction-to-explanation decision workflow



LIME can provide a secondary explanation to support robustness testing where SHAP explanations appear unstable. Counterfactual explanations can additionally identify practical changes that would reduce the score. For example, an explanation could indicate that restoring system-error rates below a defined level and resolving overdue reconciliation breaks would move the observation below the escalation threshold.

Explanation agreement itself becomes a model-control indicator. If different valid explanation methods

consistently identify contradictory drivers for the same predictions, model-risk investigation may be required.

The system is not designed to replace operational-risk experts. Automated predictions influence risk prioritization, while human reviewers maintain responsibility for material escalation and control decisions.

Low-risk observations can remain under automated monitoring. Medium-risk cases can be assigned to

analysts for review, whereas high-risk or high-severity cases can trigger immediate escalation. Analysts should receive the predicted risk level, principal explanation drivers, anomaly status, relevant historical incidents, data-quality flags, and confidence information.

Human overrides should themselves become structured data. Persistent disagreement between model recommendations and experienced analysts may reveal changing risk patterns, missing variables, threshold problems, or analyst bias. Monitoring override frequency therefore provides information about both the AI system and the governance process.

5. Research Methodology and Empirical Validation Protocol

Because access to internal operational-loss and control data is institution-specific and frequently confidential, this paper develops a validation protocol rather than presenting invented empirical observations. The protocol can be implemented using anonymized data from a commercial bank, insurer, payments institution, securities firm, or other regulated financial institution.

The appropriate unit of observation depends on the institution and risk category. Possible units include transaction-day, account-day, branch-week, process-day, application-hour, employee-month, vendor-month, or business-unit-month.

The primary outcome $Y_{i,t+h}$ should represent whether a qualifying operational-risk event occurs during a future horizon h . Separate models may be preferable for different Basel event categories because drivers of external fraud are unlikely to be identical to drivers of system disruption.

A second target should capture severity:

$$Z_{i,t+h} = \log(1 + \text{Operational Loss}_{i,t+h}).$$

For nonfinancial incidents, severity may incorporate downtime, affected customers, regulatory significance, critical-service disruption, or control impact rather than monetary loss alone.

Model evaluation should simulate genuine future prediction. If observations cover January 2017 through December 2023, an illustrative structure could train models on 2017–2020, tune hyperparameters on 2021, test initially on 2022, and maintain 2023 as an untouched out-of-time validation sample.

The approach avoids leakage associated with random splitting and provides evidence regarding model stability during changing operating conditions.

Operational incidents frequently represent a small fraction of observations. Accuracy is therefore an inappropriate principal criterion.

The framework recommends class weighting as the initial approach. Controlled resampling such as SMOTE may be evaluated strictly within the training partition, never prior to the temporal split. Precision-recall measures should accompany ROC-AUC because they are particularly informative where positive outcomes are rare.

The framework evaluates technical performance, probability reliability, explanation performance, and operational utility simultaneously.

Table 4. Proposed multidimensional evaluation framework

Dimension	Metric/test	Interpretation
Discrimination	ROC-AUC	Ability to rank risk events above non-events
Rare-event detection	PR-AUC / average precision	Performance on the minority operational-event class
Sensitivity	Recall	Percentage of actual risk events detected
Alert efficiency	Precision	Percentage of alerts representing actual events
Probability reliability	Brier score / calibration curve	Agreement between predicted and realized probabilities

Ranking utility	Recall@Top-k	Events captured within analyst capacity
Severity performance	MAE/RMSE for loss severity	Accuracy of predicted impact
Stability	PSI, rolling performance, feature drift	Model robustness over time
Explanation fidelity	Local surrogate/perturbation consistency	Whether explanations reflect underlying model behaviour
Explanation stability	Rank correlation across similar cases	Consistency of identified drivers
Governance utility	Override rate, analyst acceptance, time-to-action	Practical usefulness for decision making
Control effectiveness	Post-alert incident reduction	Whether interventions reduce subsequent exposure

Model selection should therefore not be based solely on the highest ROC-AUC. A small increase in discrimination may be insufficient to justify a substantial deterioration in calibration, stability, or explainability.

A financial institution should validate explanations rather than assuming that use of a recognized XAI method automatically generates reliable interpretability.

Three forms of validation are proposed. Technical fidelity assesses whether identified explanation drivers genuinely influence the model. Stability examines whether small, operationally irrelevant perturbations cause unreasonably large changes in explanations. Human usefulness examines whether operational-risk professionals can correctly understand the explanation and use it to determine an appropriate action.

A structured analyst study could compare four interfaces: risk score only; risk score plus top SHAP drivers; risk score plus SHAP and counterfactual explanation; and conventional KRI information without AI. Outcome measures could include decision accuracy, investigation time, confidence calibration, and override frequency.

6. Framework Outcomes and Analytical Interpretation

As a conceptual-methodological study, the principal result is the development of a unified mechanism linking operational-risk sensing, predictive modelling, explanation, and governance.

First, the framework addresses model incompleteness by combining supervised and unsupervised learning. Supervised models identify patterns similar to known historical incidents, while Isolation Forest identifies previously unseen deviations. This distinction is important because operational risk continuously evolves.

Second, the framework addresses the accuracy–interpretability tension by preserving an interpretable logistic benchmark. Complex algorithms are not automatically treated as preferable. Their adoption depends on demonstrated incremental value relative to an interpretable baseline.

Third, the framework converts generic model probability into a more operationally meaningful signal through severity, anomaly, and control adjustments. Operational-risk prioritization therefore depends on expected consequences rather than prediction likelihood alone.

Fourth, explainability is linked directly to decision making. SHAP identifies risk contributions, LIME provides complementary local approximation, and counterfactual reasoning supports the identification of potential corrective actions.

Fifth, the framework establishes a closed feedback loop in which analyst decisions, overrides, confirmed incidents, false alarms, and control remediation outcomes become inputs to future validation.

7. Discussion

The main theoretical contribution of HXAI-ORM lies in treating operational-risk prediction as a socio-technical governance problem rather than a conventional binary classification task. Existing financial AI literature demonstrates the ability of ML to produce sophisticated predictions, but predictive performance alone does not resolve questions concerning governance, model validation, transparency, and organizational accountability.

The framework extends operational-risk research by distinguishing four analytical objects: event likelihood, severity, abnormality, and control weakness. These components are conceptually different. An unusual event is not necessarily harmful, and an event with a high likelihood is not necessarily material. The hybrid formulation prevents any single signal from dominating the organization's understanding of exposure.

The study additionally contributes to XAI by emphasizing explanation performance as a model-validation dimension. XAI research frequently evaluates explanations visually or through case studies, while financial institutions require more systematic assurance concerning explanation fidelity and stability (Carvalho et al., 2019; Ding et al., 2022).

The framework further supports audience-specific explanation. Data scientists can receive full feature-contribution distributions and model diagnostics; operational-risk analysts can receive primary causal indicators and historical context; executives can receive portfolio-level concentration trends; and validators can receive stability, calibration, and explanation-fidelity results.

For financial institutions, the principal advantage of the framework is earlier prioritization of operational vulnerabilities. Risk teams frequently operate under capacity constraints and cannot investigate every process exception or technical anomaly. Risk-ranked alerts can direct limited expert capacity toward the most important cases.

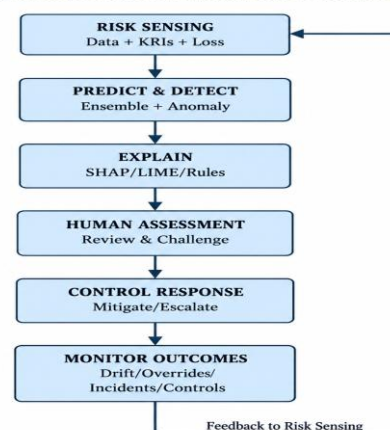
Explainability may also improve interaction between first-line business functions and second-line risk management. Instead of receiving a generic warning that a process has exceeded an algorithmic risk threshold, managers can receive evidence identifying the factors responsible for elevated risk.

Model outputs can additionally strengthen scenario design. Persistent SHAP patterns indicating vendor dependency, process exceptions, or system-capacity pressure can support stress scenarios and operational-resilience testing.

The framework aligns predictive analytics with established expectations concerning operational-risk governance and AI risk management. BCBS principles emphasize governance, risk identification, monitoring, control, ICT management, and resilience. NIST AI RMF 1.0 emphasizes governance and continuous risk measurement and management (Tabassi, 2023).

HXAI-ORM consequently requires version-controlled models, documented data sources, independent validation, threshold justification, monitoring, escalation procedures, override recording, and audit trails. These controls should apply to explanation methods as well as prediction models.

Figure 3. Continuous predictive operational-risk governance cycle



This lifecycle highlights that deployment is not the endpoint of model development. Predictive operational-risk management should continuously update its understanding as processes, threats, technologies, customers, and controls change.

8. Practical Implementation Architecture

Implementation should proceed incrementally rather than through immediate enterprise-wide automation.

An institution can initially deploy HXAI-ORM as a decision-support mechanism for one operational-risk domain, such as payment-processing errors, technology incidents, fraud operations, reconciliation failures, or third-party service disruptions. A narrow implementation permits validation of data lineage,

prediction quality, explanations, workflow integration, analyst behaviour, and governance requirements.

The first implementation stage should establish an operational-risk data mart containing temporally aligned risk events, controls, KRIs, process indicators, and relevant contextual data. The second stage should develop a transparent benchmark and evaluate whether complex algorithms generate sufficient incremental improvement. The third stage should introduce explanation mechanisms and compare machine-identified drivers against expert knowledge. The fourth stage should integrate outputs into analyst case-management systems rather than relying on standalone dashboards. The final stage should monitor whether interventions triggered by the system actually reduce subsequent incidents or exposure.

A fundamental implementation principle is that automated escalation thresholds should reflect organizational capacity and loss tolerance. A technically optimal threshold that produces thousands of alerts per day may have little operational value if analysts can investigate only a small fraction. Recall@Top-k and alert volumes should therefore be evaluated alongside conventional statistical metrics.

9. Limitations and Future Research

Several limitations should be recognized.

First, operational-loss data can be sparse, inconsistent, and affected by changes in reporting practices. A model may therefore learn differences in reporting behaviour rather than genuine risk changes.

Second, operational incidents can involve long and uncertain causal chains. XAI feature contributions should not automatically be interpreted as causal effects. SHAP and related methods explain a model's prediction; they do not independently demonstrate that changing a feature will causally prevent an incident.

Third, models trained within one financial institution may not generalize directly to another because business models, technologies, customers, controls, and reporting thresholds vary. External validation is therefore an important research opportunity.

Fourth, sophisticated attackers or internal fraudsters may adapt their behaviour in response to control mechanisms, creating adversarial model risk.

Fifth, explanation evaluation remains an evolving area. Future empirical work should examine which

combinations of global, local, rule-based, and counterfactual explanations are most effective for operational-risk professionals.

Future research should therefore test HXAI-ORM with real longitudinal bank data and compare its performance across multiple Basel operational-risk categories. Additional research could incorporate graph neural networks for connected fraud and third-party dependency analysis, natural-language processing for complaints and incident reports, and survival models for estimating time to operational failure. However, any methodological extension should be evaluated not only in terms of predictive accuracy but also interpretability, stability, governance cost, and operational value.

10. Conclusion

Financial institutions require operational-risk systems capable of identifying emerging threats before those threats become material losses or major service disruptions. Conventional backward-looking mechanisms remain essential, but the increasing speed and complexity of financial operations creates a need for complementary predictive capabilities.

This paper develops HXAI-ORM, a Hybrid Explainable Artificial Intelligence Framework for Predictive Operational Risk Management. The framework integrates interpretable statistical modelling, Random Forest, XGBoost, anomaly detection, probability calibration, severity analysis, SHAP, LIME, counterfactual reasoning, and human oversight.

The principal argument is that high-performing operational-risk AI should not be evaluated exclusively according to predictive accuracy. A financially responsible system must also provide calibrated probabilities, stable predictions, understandable explanations, appropriate escalation, human challenge, independent validation, and continuous monitoring.

Accordingly, explainability should not be treated as a cosmetic layer placed around a black-box model. It should operate as a core governance mechanism connecting predictive analytics to accountable financial decision making.

By combining prediction, interpretation, control information, human expertise, and continuous feedback, HXAI-ORM provides a pathway from retrospective operational-loss management toward proactive and explainable operational-risk intelligence. The framework also creates a testable foundation for

subsequent empirical research using institution-specific financial-services data.

References

- Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Benetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82–115. <https://doi.org/10.1016/j.inffus.2019.12.012>
- Basel Committee on Banking Supervision. (2021a). *Revisions to the principles for the sound management of operational risk*. Bank for International Settlements.
- Basel Committee on Banking Supervision. (2021b). *Principles for operational resilience*. Bank for International Settlements.
- Board of Governors of the Federal Reserve System, & Office of the Comptroller of the Currency. (2011). *Supervisory guidance on model risk management* (SR Letter 11-7).
- Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32. <https://doi.org/10.1023/A:1010933404324>
- Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2020). Explainable AI in fintech risk management. *Frontiers in Artificial Intelligence*, 3, 26. <https://doi.org/10.3389/frai.2020.00026>
- Cao, Y., & Zhai, J. (2022). A survey of AI in finance. *Journal of Chinese Economic and Business Studies*, 20(2), 125–137. <https://doi.org/10.1080/14765284.2022.2077632>
- Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8), 832. <https://doi.org/10.3390/electronics8080832>
- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321–357. <https://doi.org/10.1613/jair.953>
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). ACM. <https://doi.org/10.1145/2939672.2939785>
- Ding, W., Abdel-Basset, M., Hawash, H., & Ali, A. M. (2022). Explainability of artificial intelligence methods, applications and challenges: A comprehensive survey. *Information Sciences*, 615, 238–292. <https://doi.org/10.1016/j.ins.2022.10.013>
- Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.
- European Banking Authority. (2019). *Guidelines on ICT and security risk management* (EBA/GL/2019/04).
- Financial Stability Board. (2017). *Artificial intelligence and machine learning in financial services: Market developments and financial stability implications*. FSB.
- Hadji Misheva, B., Jaggi, D., Posth, J.-A., Gramespacher, T., & Osterrieder, J. (2021). Audience-dependent explanations for AI-based risk management tools: A survey. *Frontiers in Artificial Intelligence*, 4, 794996. <https://doi.org/10.3389/frai.2021.794996>
- King, G., & Zeng, L. (2001). Logistic regression in rare events data. *Political Analysis*, 9(2), 137–163.
- Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2021). Explainable AI: A review of machine learning interpretability methods. *Entropy*, 23(1), 18. <https://doi.org/10.3390/e23010018>
- Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation forest. In *Proceedings of the Eighth IEEE International Conference on Data Mining* (pp. 413–422). IEEE. <https://doi.org/10.1109/ICDM.2008.17>
- Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model

- predictions. *Advances in Neural Information Processing Systems*, 30, 4765–4774.
20. Milana, C., & Ashta, A. (2021). Artificial intelligence techniques in finance and financial markets: A survey of the literature. *Strategic Change*, 30(3), 189–209. <https://doi.org/10.1002/jsc.2403>
 21. Niculescu-Mizil, A., & Caruana, R. (2005). Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning* (pp. 625–632). ACM. <https://doi.org/10.1145/1102351.1102430>
 22. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). “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* (pp. 1135–1144). ACM. <https://doi.org/10.1145/2939672.2939778>
 23. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1, 206–215. <https://doi.org/10.1038/s42256-019-0048-x>
 24. Tabassi, E. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. National Institute of Standards and Technology. <https://doi.org/10.6028/NIST.AI.100-1>