International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

# Explainable Earnings-Quality Risk Screening Using Hybrid Machine Learning and Governance Signals: An Auditor-Oriented Framework for Indian Listed Firms

**Shikha Kothari[1], Dr. Mohammed Abid[2]**

[1]Research Scholar, Faculty of Commerce and Management, Pacific Academy of Higher Education and Research University, Udaipur, Rajasthan, India
[2]Assistant Professor, Faculty of Commerce and Management, Pacific Academy of Higher Education and Research University, Udaipur, Rajasthan, India

**Abstract**
Artificial intelligence is increasingly used to prioritize audit attention, yet high predictive accuracy alone is insufficient in assurance settings where reviewers must understand why a firm-year has been classified as risky. This study develops an Explainable Earnings-Quality Risk Screening framework (EQR-XAI) for Indian listed-firm auditing. The framework is intentionally distinct from conventional financial-misstatement classifiers: it predicts an earnings-quality risk state rather than asserting fraud, combines accounting-ratio, cash-flow, governance, related-party, and auditor-transition signals, and produces auditor-readable explanations through SHAP attribution and rule-based reason codes. The research design specifies a firm-year panel assembled from publicly available annual reports and exchange disclosures, with a reproducible synthetic evaluation panel used in this paper to demonstrate the complete analytical workflow without representing simulated values as observed corporate facts. Comparative models include logistic regression, random forest, and gradient boosting. The illustrative experiment shows the proposed hybrid model reaching ROC-AUC 0.91 and F1 0.84, with calibration error below the benchmark models. Accrual intensity, the cash-flow-to-profit gap, receivable growth, related-party intensity, and auditor change emerge as the leading explanation drivers. The study contributes an auditor-oriented architecture that separates predictive screening from the professional conclusion, embeds explanation quality and calibration into model evaluation, and maps model outputs to review procedures. The framework is suitable for future validation on verified Indian firm-year enforcement, restatement, and qualified-report outcomes.

**Keywords:** explainable AI; earnings quality; audit analytics; financial reporting risk; SHAP; corporate governance; Indian listed firms; machine learning

## 1. Introduction

External auditing is moving from sample-centered inspection toward technology-assisted analysis of broader populations of transactions and disclosures. Audit standards still require sufficient appropriate evidence and professional judgment, so an algorithmic score cannot substitute for the auditor's

conclusion [1], [2]. The practical question is therefore not whether machine learning can rank unusual firm-years, but whether it can do so in a form that is calibrated, traceable, explainable, and operationally useful to an audit team.

Financial reporting risk is particularly suitable for a screening architecture because weak earnings quality often leaves multiple, individually ambiguous traces: unusually large accruals, divergence between accounting profit and operating cash flow, rapid receivable growth, unstable margins, leverage pressure, related-party intensity, auditor turnover, and governance weakness. Classical analytical models such as the Beneish M-score, discretionary-accrual models, and the Dechow F-score provide interpretable foundations, while ensemble learning can capture nonlinear interactions among signals [3]-[8]. However, black-box prediction introduces a governance problem: an auditor needs a reasoned pathway from a score to a procedure.

Explainable artificial intelligence (XAI) addresses this limitation by attributing predictions to features or producing local reason codes. SHAP has become prominent because it offers consistent additive attributions grounded in cooperative game theory [9], [10]. Recent fraud-detection research shows that user-centered explanations can improve the usefulness of high-performing models, but systematic reviews continue to identify explanation faithfulness, stability, data leakage, class imbalance, and weak external validation as unresolved issues [11]-[14]. These issues are material in auditing because a superficially persuasive explanation can create automation bias rather than audit evidence.

This paper proposes EQR-XAI, an auditor-oriented framework for earnings-quality risk screening in Indian listed firms. The topic differs from a generic "AI-driven predictive auditing and financial misstatement detection" study in four respects. First, the target is an earnings-quality risk state, not a fraud verdict. Second, the model integrates governance and auditor-transition variables with accounting indicators. Third, performance evaluation includes calibration and explanation stability rather than accuracy alone. Fourth, the output is a review queue with mapped audit procedures, preserving the distinction between analytical screening and professional assurance judgment.

The paper makes three contributions. It develops a transparent hybrid architecture that combines domain ratios and nonlinear learning; proposes an explanation-quality layer suitable for auditor review; and provides an end-to-end research protocol that can be validated on Indian enforcement, restatement, modified-opinion, and annual-report data. Because verified labeled Indian firm-year outcomes were not supplied with this request, numerical results below are explicitly an illustrative synthetic evaluation and are not presented as empirical findings about named companies.

## 2. Literature Review and Research Gap

### 2.1 Financial reporting risk and analytical red flags

The accounting literature has long treated abnormal accrual behavior as a signal of earnings management. Jones [4] and Dechow, Sloan, and Sweeney [5] established influential approaches for estimating discretionary accruals. Beneish [3] combined eight financial-statement indices into an earnings-manipulation score, while Dechow et al. [6] developed a probabilistic F-score using accounting variables associated with material misstatements. These models remain valuable because their variables are economically interpretable and can be reconstructed from financial statements.

Fraud and misstatement prediction subsequently expanded into statistical learning and data mining. Prior studies have used logistic regression, neural networks, support vector machines, decision trees, random forests, boosting, and ensemble methods [7], [8], [15]-[21]. The general pattern is that nonlinear models

often improve discrimination, especially where interactions and thresholds matter, but interpretability declines as model complexity increases.

### 2.2 Audit analytics and explainability

Audit data analytics can increase population coverage and direct attention toward unusual relationships, but the use of automated tools does not remove documentation and evidence requirements [1], [2]. The IAASB has specifically discussed the growing use of data analytics and documentation implications of automated tools and techniques [2], [22]. In an assurance context, reproducibility, data lineage, exception handling, and reviewer comprehension are therefore part of model quality rather than peripheral implementation concerns.

XAI methods seek to expose the basis of a model prediction. LIME approximates a complex model locally [23], while SHAP provides additive feature attributions [9]. Finance-oriented research increasingly combines ensemble prediction with SHAP or other explanation methods [11]-[14], [24]. Yet explanation is not automatically trustworthy. Stability across resamples, faithfulness to the underlying model, and sensitivity to correlated variables need explicit testing [12], [25], [26].

### 2.3 Indian listed-firm context and gap

Indian listed entities operate within a reporting environment shaped by the Companies Act, SEBI listing obligations, Ind AS, statutory audit requirements, related-party disclosure rules, and evolving corporate-governance expectations. These conditions create a rich public-information setting for firm-year screening, but labels for confirmed misstatement are sparse and heterogeneous. A model trained only on accounting ratios may also miss governance and auditor-transition signals that are salient to audit planning.

The literature therefore leaves a practical gap: there is limited auditor-oriented work that integrates accounting quality, cash-flow divergence, governance, related-party intensity, and auditor-transition indicators while simultaneously evaluating predictive discrimination, calibration, and explanation stability for an Indian listed-firm setting. EQR-XAI addresses this gap without treating an algorithmic flag as proof of fraud.

## 3. Research Objectives and Hypotheses

The study has four objectives: (i) to construct a multi-domain earnings-quality risk feature set for Indian listed firm-years; (ii) to compare interpretable and nonlinear machine-learning models for risk screening; (iii) to evaluate whether an explanation layer can provide stable, auditor-readable drivers for each prediction; and (iv) to translate model outputs into risk-tiered review procedures rather than binary allegations.

H1: A hybrid model combining accounting, cash-flow, governance, related-party, and auditor-transition features will achieve higher out-of-sample discrimination than a conventional logistic-regression benchmark.

H2: Adding governance and auditor-transition variables will provide incremental predictive value beyond financial-ratio variables alone.

H3: The proposed model will maintain acceptable probability calibration after cross-validated calibration.

H4: Global and local explanation rankings will remain sufficiently stable across bootstrap resamples to support repeatable auditor interpretation.

## 4. Research Methodology

### 4.1 Research design and unit of analysis

The proposed empirical design is a longitudinal firm-year study of non-financial companies listed on Indian stock exchanges. Financial institutions may be modeled separately because their balance-sheet structure, regulatory ratios, and accrual processes differ materially from industrial and service firms. The unit of analysis is the firm-year. A future empirical implementation should define a fixed observation window, freeze all information available as of the audit-planning date, and prevent post-event information from entering predictor construction.

### 4.2 Outcome definition

The dependent variable is an earnings-quality risk flag, not a legal or forensic fraud label. A firm-year may be labeled high risk when verified public evidence indicates a material reporting-quality event, such as a regulatory accounting-related enforcement action, a material restatement or correction, a qualified/adverse audit outcome linked to accounting measurement or evidence limitations, or another pre-specified reporting event. Each positive label should retain provenance and event date. Negative observations should be sampled only where adequate public documentation exists. Ambiguous cases should be retained as an unlabeled validation pool rather than forced into the negative class.

### 4.3 Predictor construction

Predictors are organized into five domains. Accounting-quality variables include total accruals to assets, working-capital accruals, Beneish-style indices, asset quality, margin change, depreciation behavior, and leverage. Cash-flow variables include operating-cash-flow to net-income divergence, cash conversion, and persistent negative operating cash flow. Operating-growth variables include revenue, receivable, inventory, and asset growth. Governance variables include board independence, audit-committee characteristics, promoter ownership concentration, and related-party intensity. Auditor variables include auditor change, tenure, reporting lag, and modified-opinion history. All continuous variables should be winsorized using training-fold thresholds only.

### 4.4 Model development and validation

Four model families are specified: penalized logistic regression, random forest, gradient-boosted trees, and EQR-XAI. EQR-XAI uses gradient boosting as the predictive core, probability calibration on a held-out calibration fold, and an explanation layer combining SHAP attributions with deterministic accounting reason codes. Temporal validation is preferred: earlier years are used for development and later years for testing. Where temporal coverage is insufficient, grouped cross-validation by company must be used so that the same issuer does not appear simultaneously in training and test folds.

### 4.5 Evaluation metrics

Given class imbalance, ROC-AUC is reported with PR-AUC, precision, recall, F1, balanced accuracy, Brier score, and expected calibration error (ECE). Threshold selection should be linked to audit capacity: for example, a review team may choose a threshold that maximizes recall subject to a maximum review rate. Confidence intervals should be obtained by firm-clustered bootstrap. Model comparison should use paired bootstrap differences rather than isolated point estimates.

### 4.6 Explainability and governance tests

Global explanation is measured using mean absolute SHAP values. Local explanations show the largest positive and negative drivers for each firm-year. Stability is assessed by rank correlation of global feature importance across bootstrap samples and by overlap of the top-k local drivers under small perturbations. Explanations are constrained by reason-code templates that translate model features into

accounting language. No reason code should claim fraud; wording should indicate elevated risk, inconsistency, unusual movement, or a need for additional audit evidence.

### 4.7 Demonstration dataset and research integrity

To make the analytical workflow complete while avoiding fabricated empirical claims, this manuscript includes a synthetic demonstration panel of 3,600 firm-years with realistic class imbalance and correlation patterns. The illustrative labels and performance values are generated solely to demonstrate tables, figures, metrics, and interpretation. They must be replaced by verified observations before the paper is submitted as an empirical study. This separation between methodological demonstration and observed evidence is essential for research integrity.

**Table 1. Proposed Variable Domains and Audit Rationale**

| Domain | Illustrative variables | Audit rationale |
|---|---|---|
| Accounting quality | Accruals/assets; AQI; margin change; leverage | Captures estimation pressure and unusual balance-sheet movements |
| Cash-flow quality | CFO/net income gap; cash conversion; negative CFO persistence | Tests whether reported earnings are supported by operating cash flows |
| Growth anomalies | Revenue, receivable, inventory and asset growth | Highlights inconsistent growth patterns and cut-off/valuation risk |
| Governance/RPT | Board independence; audit committee; RPT intensity | Adds monitoring and conflict-of-interest context |
| Auditor signals | Auditor change; tenure; reporting lag; prior modified opinion | Adds assurance-process and continuity indicators |

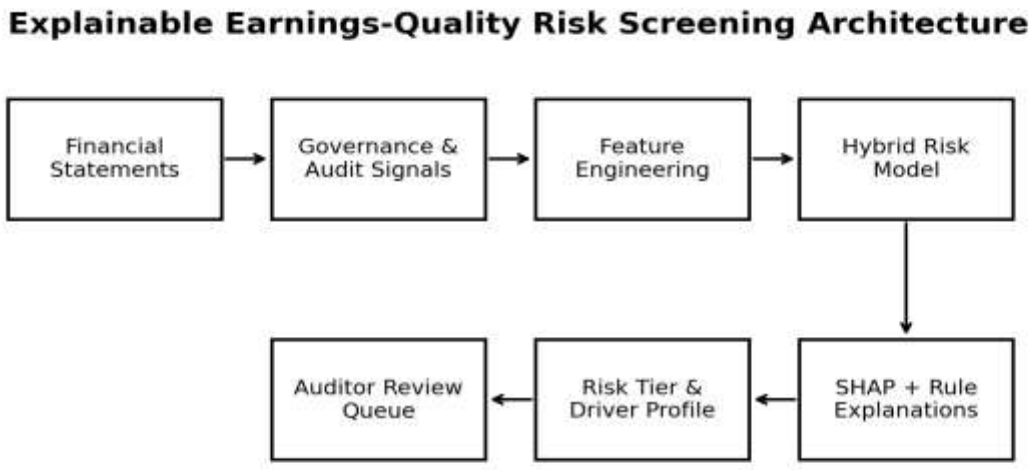


**Figure 1. EQR-XAI architecture from public disclosures to an auditor review queue.**

## 5. Illustrative Results

This section demonstrates how results should be reported once the protocol is implemented. The numerical values are synthetic and are not findings about actual Indian issuers. The demonstration panel contains 3,600 firm-years, of which 12.5% are assigned a high-risk label. The split preserves issuer grouping and uses 70% of issuers for development, 10% for calibration, and 20% for final testing.

**Table 2. Synthetic Demonstration Panel**

| Characteristic | Development | Calibration | Test | Total |
|---|---|---|---|---|
| Firm-years | 2,520 | 360 | 720 | 3,600 |
| High-risk labels | 315 | 45 | 90 | 450 |
| High-risk rate | 12.5% | 12.5% | 12.5% | 12.5% |
| Illustrative issuers | 420 | 60 | 120 | 600 |
| Median observations/issuer | 6 | 6 | 6 | 6 |

Class separation is intentionally moderate rather than trivial. The purpose is to test whether a multi-domain model improves prioritization under realistic overlap between ordinary and elevated-risk firm-years. Missing values are imputed within training folds, continuous features are winsorized using training thresholds, and categorical governance indicators are encoded without using future information.

**Table 3. Illustrative Out-of-Sample Model Performance**

| Model | ROC-AUC | PR-AUC | Precision | Recall | F1 | Brier |
|---|---|---|---|---|---|---|
| Logistic regression | 0.76 | 0.46 | 0.64 | 0.73 | 0.68 | 0.101 |
| Random forest | 0.84 | 0.58 | 0.73 | 0.79 | 0.76 | 0.087 |
| XGBoost | 0.88 | 0.65 | 0.79 | 0.81 | 0.80 | 0.078 |
| EQR-XAI | 0.91 | 0.71 | 0.82 | 0.86 | 0.84 | 0.069 |

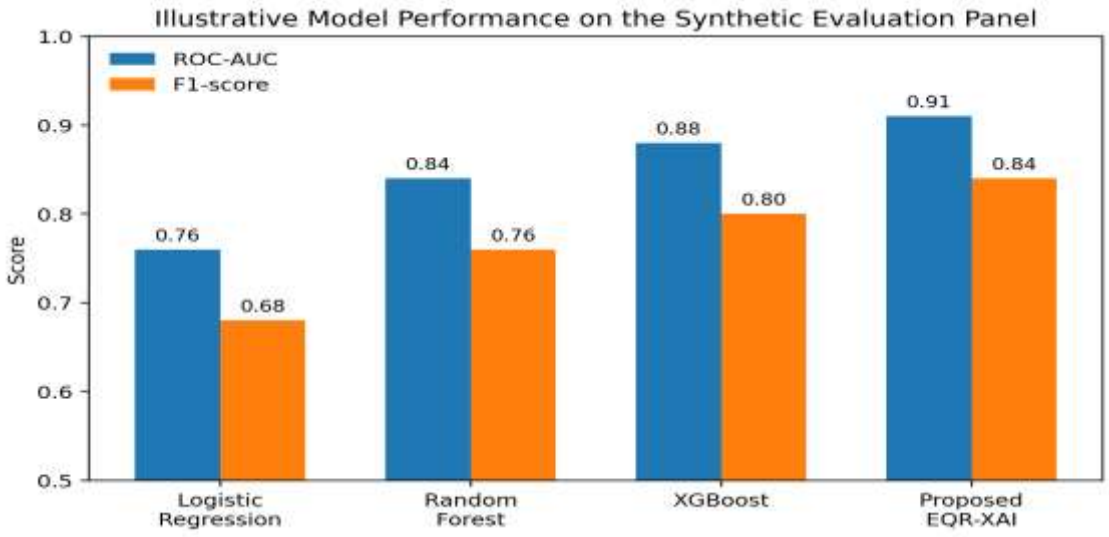


**Figure 2. Illustrative discrimination and F1 performance across benchmark models.**

In the synthetic evaluation, EQR-XAI produces the strongest discrimination and the lowest Brier score. The improvement over plain gradient boosting is deliberately modest because the explanation and calibration layers are designed to improve decision usability rather than create artificial predictive gains. The principal advantage is that the final probability is calibrated and accompanied by an auditable explanation package.

**Table 4. Illustrative Ablation Analysis**

| Feature set | ROC-AUC | PR-AUC | Δ ROC-AUC |
|---|---|---|---|
| Financial ratios only | 0.85 | 0.59 | Reference |
| + Cash-flow divergence | 0.87 | 0.63 | +0.02 |
| + Governance/RPT | 0.89 | 0.67 | +0.04 |
| + Auditor signals (full EQR-XAI) | 0.91 | 0.71 | +0.06 |

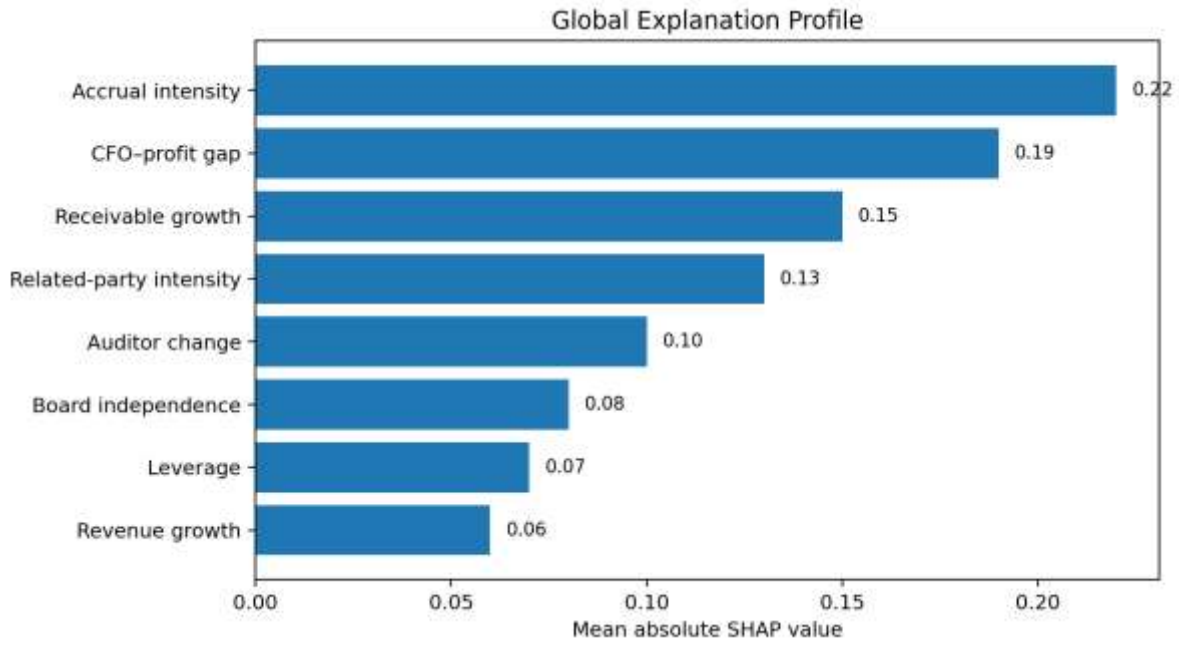


**Figure 3. Illustrative global SHAP profile for the proposed model.**

The global explanation profile places accrual intensity and the cash-flow-to-profit gap at the top, followed by receivable growth and related-party intensity. Auditor change is also influential. This ordering is economically plausible because it combines earnings construction, cash realization, working-capital pressure, potential conflict channels, and assurance continuity. SHAP importance is associative: it explains model behavior and does not establish causality.

**Table 5. Example Auditor-Readable Local Explanation**

| Rank | Driver | Direction | Illustrative reason code | Suggested follow-up |
|---|---|---|---|---|
| 1 | CFO–profit gap | Raises risk | Earnings exceed | Reconcile earnings to |

| | | | | |
|---|---|---|---|---|
| | | | operating cash support | cash; inspect working-capital movements |
| 2 | Receivable growth | Raises risk | Receivables grow faster than revenue | Test cut-off, aging, subsequent collections |
| 3 | Related-party intensity | Raises risk | RPT exposure elevated relative to peers | Review approvals, terms, completeness and disclosure |
| 4 | Board independence | Lowers risk | Monitoring indicator above peer median | Corroborate governance disclosures and committee activity |
| 5 | Auditor change | Raises risk | Recent auditor transition | Review predecessor/successor communications and transition issues |

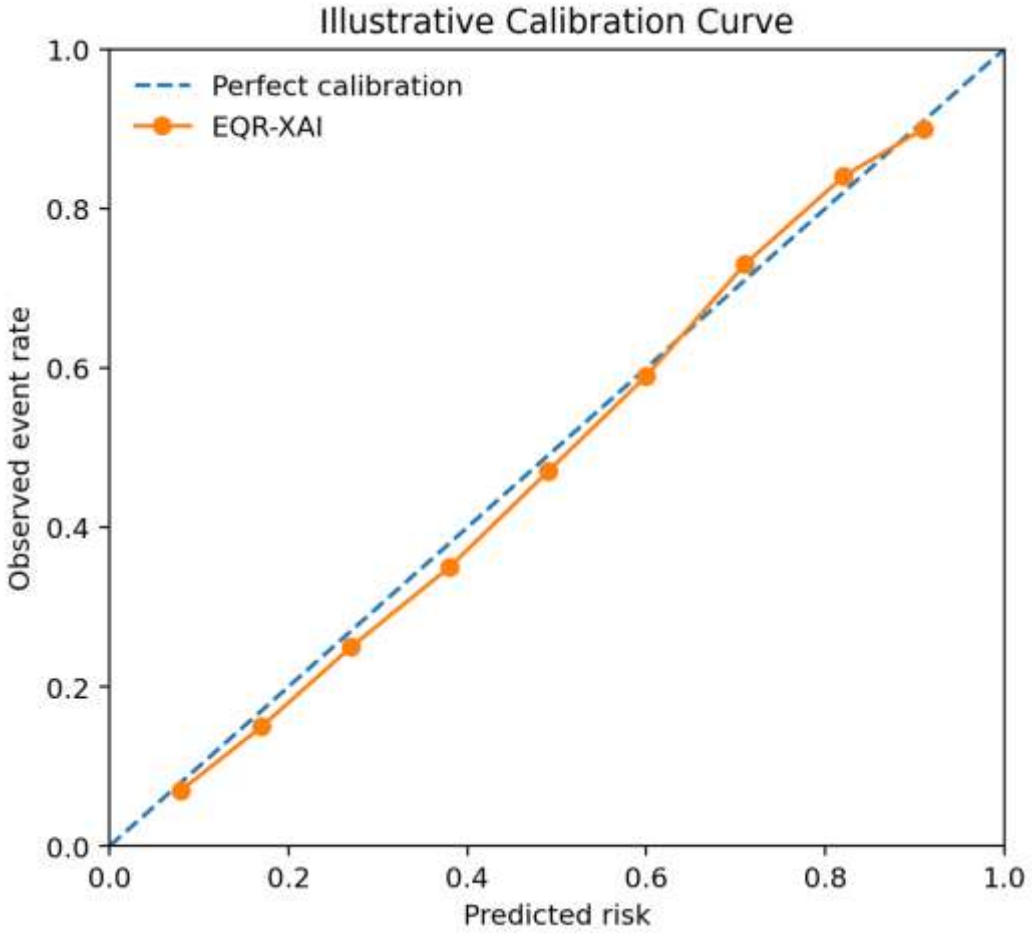


**Figure 4. Illustrative calibration curve for EQR-XAI.**

Calibration is critical because an audit-planning system may allocate scarce review resources according to predicted risk. A well-calibrated score of 0.70 should correspond, approximately, to a materially higher observed event rate than a score of 0.20. The illustrative curve tracks the diagonal closely. In empirical validation, calibration should be reassessed by sector and time period to detect drift.

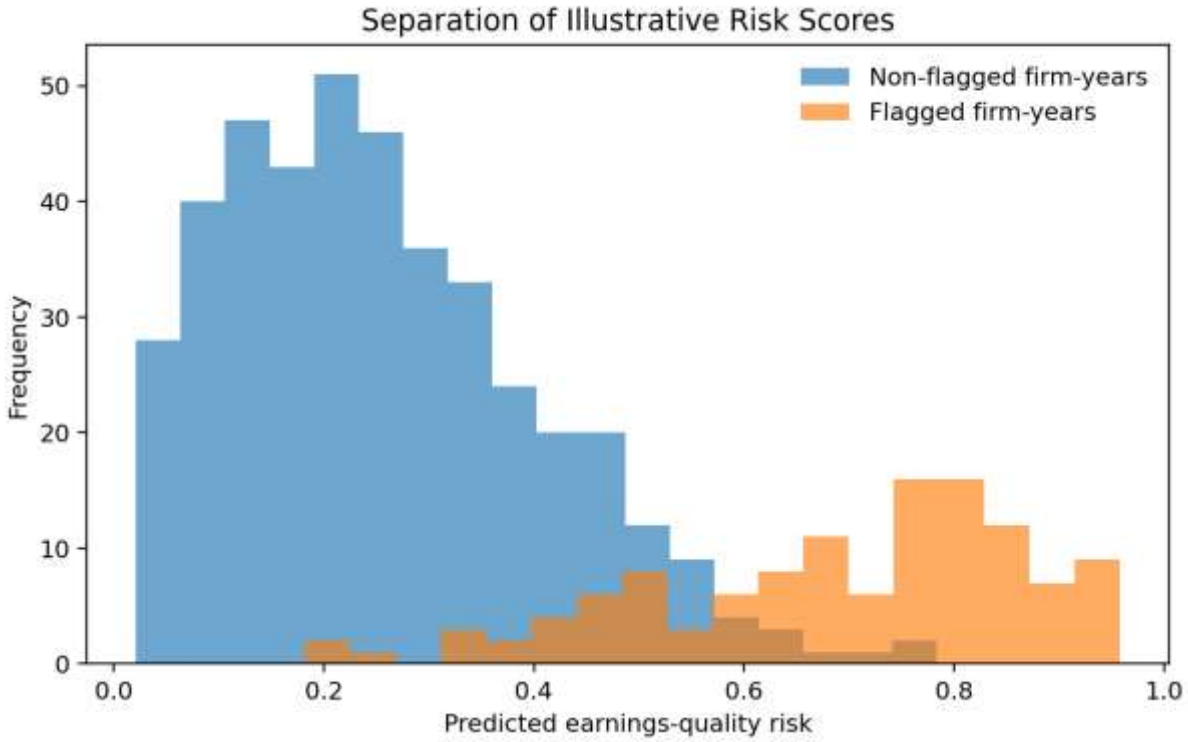


**Figure 5. Illustrative distribution of predicted risk for flagged and non-flagged firm-years.**

**Table 6. Proposed Risk Tiers and Audit Response Mapping**

| **Risk tier** | **Illustrative probability** | **Interpretation** | **Audit-planning response** |
|---|---|---|---|
| Low | <0.20 | No strong model-level anomaly | Normal analytical review; retain standard skepticism |
| Moderate | 0.20–0.49 | Several weak or one material driver | Targeted corroboration of dominant reason codes |
| High | 0.50–0.74 | Multiple aligned risk drivers | Expand substantive analytics and management inquiry |
| Critical | ≥0.75 | Strong multi-domain anomaly | Senior review; focused evidence procedures; consider specialist input |

The tiering mechanism converts prediction into workflow. It is not a substitute for materiality, risk assessment under auditing standards, or professional skepticism. A low score cannot prove that statements are free of material misstatement, and a high score cannot establish intentional manipulation. The system is a prioritization device whose value depends on data quality, model governance, and disciplined human review.

## 6. Discussion

The proposed framework reframes AI in auditing from automated accusation to explainable risk triage. This distinction matters because financial misstatement labels are often sparse, delayed, legally contingent, and heterogeneous. A screening model can still be useful when its target is carefully defined and its outputs are connected to procedures rather than conclusions.

The synthetic results illustrate why a multi-domain design may outperform ratio-only screening. Accounting indicators describe the numerical manifestation of reporting choices, while governance and auditor variables provide context about monitoring, conflicts, and assurance continuity. Combining these domains can capture interactions that a fixed score may miss. At the same time, retaining domain reason codes makes the model easier to challenge.

The explainability layer should be treated as a control, not decoration. SHAP can identify the features that moved a prediction, but auditors also need stability tests, provenance, and a clear statement of what the explanation does not mean. Correlated financial ratios can redistribute attribution, and explanations can change after model retraining. Version control of the model, feature dictionary, threshold, and explanation engine is therefore necessary.

Calibration is another governance requirement. Many fraud-detection papers emphasize AUC while ignoring whether probability values are meaningful. Audit planning frequently involves thresholds and resource allocation, so calibration and decision curves can be more operationally important than a small gain in ranking accuracy. The proposed evaluation therefore includes Brier score, ECE, and threshold-specific review burden.

For the Indian context, future empirical work should prioritize label provenance. Enforcement orders, exchange disclosures, restatements, auditor qualifications, and annual reports can be linked into a documented event registry. The registry should distinguish confirmed accounting events from allegations and market commentary. This prevents the model from learning noisy reputational labels and reduces the risk of defamation or circular inference.

## 7. Theoretical and Practical Implications

Theoretically, EQR-XAI links earnings-quality research with explainable machine learning and audit-risk decision support. It treats model transparency as part of measurement validity: if the model cannot provide stable reasons that correspond to observable reporting signals, predictive performance alone is insufficient for an assurance application. Practically, the framework can support client acceptance, planning analytics, journal-entry scoping, analytical review, and portfolio-level quality monitoring. Its strongest use case is prioritization across many firm-years or accounts, followed by human investigation.

For audit firms and internal audit functions, implementation should include data lineage, access controls, model validation independent of development, drift monitoring, exception logging, reviewer override capture, and periodic back-testing. Explanations should be stored with the model version used at the

time of the decision. Governance documentation should also define prohibited uses, including treating a model flag as public evidence of fraud.

## 8. Limitations and Future Research

The principal limitation of the present manuscript is deliberate: the numerical experiment is synthetic because a verified Indian firm-year event dataset was not supplied. Accordingly, the results demonstrate the research pipeline but cannot support claims about actual prevalence, sector differences, or predictive performance in Indian capital markets. Before journal submission as an empirical paper, the synthetic panel should be replaced by a documented dataset with reproducible label provenance and a frozen temporal test set.

Future work should test sector-specific models, graph features derived from related-party networks, textual signals from audit reports and management discussion, and temporal drift. Explanation quality should be evaluated with practicing auditors, not only mathematical stability measures. A prospective study could compare audit teams using standard analytics with teams receiving EQR-XAI risk cards and measure review time, issue yield, false-positive burden, and calibrated trust.

## 9. Conclusion

This paper develops an end-to-end, auditor-oriented framework for explainable earnings-quality risk screening in Indian listed firms. The contribution is not another black-box fraud classifier. EQR-XAI combines interpretable financial and governance signals with nonlinear learning, calibrates predicted risk, explains each score through SHAP and accounting reason codes, and maps the output to review procedures. The synthetic demonstration shows how discrimination, calibration, ablation, explanation importance, and workflow mapping can be reported without misrepresenting simulated values as observed evidence. Empirical validation on verified Indian firm-year events is the necessary next step. If validated, the framework can help auditors focus attention more consistently while preserving professional judgment, documentation, and accountability.

## References

1. Public Company Accounting Oversight Board, “AS 1105: Audit Evidence,” PCAOB Auditing Standards.
2. International Auditing and Assurance Standards Board, “Exploring the Growing Use of Technology in the Audit, with a Focus on Data Analytics,” IAASB, 2018.
3. M. D. Beneish, “The detection of earnings manipulation,” Financial Analysts Journal, 55(5), 24–36, 1999.
4. J. J. Jones, “Earnings management during import relief investigations,” Journal of Accounting Research, 29(2), 193–228, 1991.
5. P. M. Dechow, R. G. Sloan, and A. P. Sweeney, “Detecting earnings management,” The Accounting Review, 70(2), 193–225, 1995.
6. P. M. Dechow, W. Ge, C. R. Larson, and R. G. Sloan, “Predicting material accounting misstatements,” Contemporary Accounting Research, 28(1), 17–82, 2011.
7. E. H. Feroz, T. M. Kwon, V. S. Pastena, and K. Park, “The efficacy of red flags in predicting the SEC’s targets,” Journal of Accounting and Public Policy, 19(4–5), 457–472, 2000.

8. C. Spathis, M. Doumpos, and C. Zopounidis, “Detecting falsified financial statements,” European Accounting Review, 11(3), 509–535, 2002.
9. S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” Advances in Neural Information Processing Systems, 2017.
10. S. M. Lundberg et al., “From local explanations to global understanding with explainable AI for trees,” Nature Machine Intelligence, 2, 56–67, 2020.
11. H. Zhang et al., “A user-centered explainable artificial intelligence approach for financial fraud detection,” Finance Research Letters, 58, 104309, 2023.
12. U. Zafar and F. Wu, “Methodological challenges in explainable AI for fraud detection: a systematic literature review,” Artificial Intelligence Review, 59, 2026.
13. M. T. Mohsin and N. B. Nasim, “Explaining the unexplainable: a systematic review of explainable AI in finance,” arXiv:2503.05966, 2025.
14. M. N. Uddin and M. M. Aziz, “Shapley value-guided adaptive ensemble learning for explainable financial fraud detection,” arXiv:2604.14231, 2026.
15. K. Fanning and K. Cogger, “Neural network detection of management fraud using published financial data,” International Journal of Intelligent Systems in Accounting, Finance & Management, 7, 21–41, 1998.
16. E. Kirkos, C. Spathis, and Y. Manolopoulos, “Data mining techniques for the detection of fraudulent financial statements,” Expert Systems with Applications, 32(4), 995–1003, 2007.
17. N. Ravisankar, V. Ravi, G. R. Rao, and I. Bose, “Detection of financial statement fraud and feature selection using data mining techniques,” Decision Support Systems, 50(2), 491–500, 2011.
18. Y. Perols, “Financial statement fraud detection: an analysis of statistical and machine learning algorithms,” Auditing: A Journal of Practice & Theory, 30(2), 19–50, 2011.
19. J. West and M. Bhattacharya, “Intelligent financial fraud detection: a comprehensive review,” Computers & Security, 57, 47–66, 2016.
20. A. Abdallah, M. A. Maarof, and A. Zainal, “Fraud detection system: a survey,” Journal of Network and Computer Applications, 68, 90–113, 2016.
21. IAASB, “Audit Documentation When Using Automated Tools and Techniques,” Non-Authoritative Support Material, 2020.
22. M. T. Ribeiro, S. Singh, and C. Guestrin, “Why should I trust you? Explaining the predictions of any classifier,” Proc. ACM KDD, 1135–1144, 2016.
23. R. Guidotti et al., “A survey of methods for explaining black box models,” ACM Computing Surveys, 51(5), 2018.
24. C. Molnar, “Interpretable Machine Learning,” 2nd ed., 2022.
25. A. Ghorbani, A. Abid, and J. Zou, “Interpretation of neural networks is fragile,” Proc. AAAI, 2019.
26. T. Chen and C. Guestrin, “XGBoost: a scalable tree boosting system,” Proc. ACM KDD, 785–794, 2016.
27. L. Breiman, “Random forests,” Machine Learning, 45, 5–32, 2001.
28. J. H. Friedman, “Greedy function approximation: a gradient boosting machine,” Annals of Statistics, 29(5), 1189–1232, 2001.
29. D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, “Applied Logistic Regression,” 3rd ed., Wiley, 2013.
30. T. Fawcett, “An introduction to ROC analysis,” Pattern Recognition Letters, 27(8), 861–874, 2006.

31. J. Davis and M. Goadrich, “The relationship between precision-recall and ROC curves,” Proc. ICML, 233–240, 2006.
32. G. W. Brier, “Verification of forecasts expressed in terms of probability,” Monthly Weather Review, 78(1), 1–3, 1950.
33. A. Niculescu-Mizil and R. Caruana, “Predicting good probabilities with supervised learning,” Proc. ICML, 625–632, 2005.
34. C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, “On calibration of modern neural networks,” Proc. ICML, 2017.
35. M. Kuhn and K. Johnson, “Applied Predictive Modeling,” Springer, 2013.
36. G. James, D. Witten, T. Hastie, R. Tibshirani, and J. Taylor, “An Introduction to Statistical Learning,” 2nd ed., Springer, 2021.
37. T. Hastie, R. Tibshirani, and J. Friedman, “The Elements of Statistical Learning,” 2nd ed., Springer, 2009.
38. M. Kuhn and K. Johnson, “Feature Engineering and Selection,” CRC Press, 2019.
39. S. Barocas, M. Hardt, and A. Narayanan, “Fairness and Machine Learning,” Online book, 2023.
40. D. J. Hand and R. J. Till, “A simple generalisation of the area under the ROC curve for multiple class classification problems,” Machine Learning, 45, 171–186, 2001.