Research Square

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

# SHAP and LIME Explainability in Financial Machine Learning: A Systematic Review of Methods, Evaluation Gaps, and Barriers to Deployment

**Manpreet Singh**
Boston University
**Akshatha Srikantha**
University of California, Irvine
**Keshav Kaushik**
Sharda University
**Prajwal Bajpai**
prajwalb@iitk.ac.in

Indian Institute of Technology Kanpur

**Research Article**

**Keywords:** Explainable AI, SHAP, LIME, Finance, Machine learning, Systematic literature review, Interpretability, PRISMA

**Posted Date:** August 18th, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-10464976/v1

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

**Additional Declarations:** No competing interests reported.

# SHAP and LIME Explainability in Financial Machine Learning: A Systematic Review of Methods, Evaluation Gaps, and Barriers to Deployment

Manpreet Singh [1*], Akshatha Srikantha [2], Keshav Kaushik [3], Prajwal Bajpai [4*]

[1]Boston University, Boston, MA, 02215, USA.
[2]University of California, Irvine, CA, USA.
[3]Sharda University, Greater Noida, UP, India.
[4]Indian Institute of Technology, Kanpur, India.

*Corresponding author(s). E-mail(s): manni@bu.edu; prajwalb@iitk.ac.in;
Contributing authors: asrikan4@uci.edu;
officialkeshavkaushik@gmail.com;

**Abstract**

Financial institutions rely heavily on Artificial Intelligence (AI). They use it for high-stakes decisions like credit scoring, fraud detection, algorithmic trading, risk assessment, and insurance pricing. Model complexity often results in black-box systems. Customers, auditors, managers, and regulators cannot easily inspect how these models make decisions, which creates problems for trust, fairness, accountability, and regulatory compliance. Explainable AI (XAI) tackles these problems. It offers tools to clarify model behaviour while keeping predictive performance high. In this systematic literature review, we synthesize empirical research on post-hoc XAI in finance, specifically focusing on SHAP and LIME. We analyze the literature across four areas. We look at how techniques map to tasks, whether generated explanation outputs undergo explicit evaluation, what technical and data-related barriers block deployment, and what broader validation strategies researchers use. We followed PRISMA 2020 guidelines and software engineering systematic review protocols to search Scopus, Web of Science, and IEEE Xplore in May 2026. Our search yielded 4,168 records. After removing duplicates, we screened 3,010 records and sought 325 reports for retrieval. We assessed 312 full texts. The final corpus includes 297 peer-reviewed

studies. Every included study directly applies SHAP, LIME, or both to a financial machine learning task. Classification tasks account for the vast majority of our sample (219 studies, 73.7%). The most common financial domain is credit, loan, and default modelling (140 studies, 47.1%). Grouped SHAP variants appear in 279 studies (93.9%). This massive presence results directly from TreeSHAP's native integration into gradient-boosted tree frameworks like XGBoost, LightGBM, and CatBoost, rather than representing all XAI methods in finance. We found that only 27 studies (9.1%; roughly 5.8% to 12.4%) report evidence of explanation quality. Just 21 studies (7.1%; roughly 4.2% to 10.0%) conduct a formal evaluation. Evaluation is clearly the weakest aspect of current SHAP and LIME applications in finance. The application of post-hoc XAI in finance is expanding fast. Yet, to reliably support regulated financial decision-making, the field needs much stronger validation, clearer reporting standards, and concrete evidence of deployment viability.

**Keywords:** Explainable AI, SHAP, LIME, Finance, Machine learning, Systematic literature review, Interpretability, PRISMA

# 1 Introduction

Financial machine learning now relies heavily on explainable artificial intelligence (XAI). Financial institutions deploy high-performing models for decisions that require auditability and fairness. Flexible machine learning models improve credit scoring, fraud detection, insurance pricing, risk management, algorithmic trading, and financial forecasting. We know these models use complex non-linear representations. Customers, auditors, managers, and regulators struggle to inspect them [1–5]. A practical tension emerges from this opacity. Banks and insurers want predictive performance, yet they must have explanations to ensure trust, accountability, and regulatory compliance.

Model outputs frequently feed into consequential decision chains within the financial sector. Loan approvals depend on credit-scoring models. Fraud classifiers can freeze accounts or trigger investigations. Insurance models adjust premium costs, trading algorithms shape portfolio exposure, and risk models drive capital allocation. Explanation functions as more than just a communication aid in these environments. Governance teams use it to validate, monitor, challenge, and audit models responsibly. Developers cannot simply check if an explanation looks visually plausible. They must evaluate the explanation against the specific financial task, the intended stakeholder, regulatory rules, the data regime, and deployment constraints.

Researchers have rapidly expanded the finance-specific XAI literature. Their evidence base, however, lacks cohesion. We see massive variation across studies regarding financial domains, machine learning tasks, dataset types, model families, explanation methods, evaluation metrics, and deployment environments. Some authors treat XAI strictly as a post-hoc visualization step. Others link explanations directly to model validation, regulatory compliance, and operational decision support. We now have a massive collection of empirical demonstrations. We still lack cumulative answers about

how techniques map to financial tasks, how practitioners evaluate explanation quality, and what keeps research prototypes from becoming deployable financial systems.

Researchers and practitioners face two main problems from this fragmentation. Comparing studies proves difficult when authors report completely different combinations of models, explanation methods, financial tasks, and evaluation metrics. You cannot directly compare a SHAP explanation for a tree-based credit model to a LIME explanation for a neural time-series model or a counterfactual explanation for a fraud classifier. Also, applying an XAI method does not guarantee validation. Authors frequently publish feature-importance plots and local explanations without testing fidelity, robustness, stability, or decision-support value. We see similar evaluation gaps documented in XAI research outside of finance [6]. The literature often appears more mature than the underlying evidence supports. The total number of XAI applications keeps growing while the evidentiary standard for judging explanation quality stays uneven.

Past reviews chart important parts of this research area. Černevičienė and Kabašinskas [7], Weber et al. [8], Yeo et al. [9], Khan et al. [10], and Martins et al. [11] synthesize XAI in finance from various perspectives. They cover broad technique categorization, model-agnostic methods, taxonomies, and application-specific challenges. Previous work largely misses four elements necessary for deploying financial XAI. We focus on mapping techniques to tasks, evaluating explanations, analyzing technical and data-related deployment barriers, and assessing methodological validation strategies. We address this gap by auditing the literature as an evidence base. We ask directly whether researchers validate their explanations strongly enough to support high-stakes financial operations, rather than simply cataloguing XAI applications.

## 1.1 Scope and Review Rationale

This review focuses on empirical studies of XAI in finance in which SHAP, LIME, or both are genuinely applied to a trained financial machine learning model. SHAP and LIME are used as the inclusion gate because they are among the most widely applied post-hoc attribution methods in financial machine learning and appear frequently enough to support cross-study comparison on a common methodological footing. The review also records co-present explanation and validation methods, such as counterfactual explanations, feature importance methods, causal validation, robustness testing, domain-expert assessment, and human-centered evaluation, when they appear alongside SHAP or LIME.

The SHAP/LIME focus is a deliberate scoping decision rather than a claim that these two methods exhaust the XAI field. The broader XAI literature includes intrinsically interpretable models, rule extraction, partial dependence plots, saliency methods, attention-based explanations, prototypes, counterfactual explanations, causal explanations, and interactive visual analytics [2, 12, 13]. A review that included every XAI method in finance would be broader in coverage but weaker in cross-study comparability, since heterogeneous explanation families rarely report evidence on common terms. By using SHAP and LIME as the eligibility gate, we obtain a corpus large enough for systematic mapping and coherent enough for cross-study synthesis. At the same

time, we preserve broader methodological context by coding additional explanation and validation approaches when they appear in the included studies.

The review covers peer-reviewed journal articles and conference papers published from 2016 onward, with the database search executed in May 2026. We searched Scopus, Web of Science, and IEEE Xplore, screened 3,010 unique records, assessed 312 full texts, identified 297 eligible studies that genuinely applied SHAP and/or LIME to financial tasks, and identified a 97-study high-coverage sample used to test the robustness of the main findings. The resulting corpus should be interpreted as a database and search-string-bounded snapshot of SHAP/LIME-based financial XAI as in May 2026, rather than as an exhaustive census of every finance subdomain or every explanation method. The full methodological procedure is reported in Section 3. The search strings are reported in Appendix A.

The 97-study high-coverage sample is not intended to erase the larger eligible set or to serve as a second corpus. Rather, it supports a two-level review design in which the full corpus of 297 studies is the basis for all prevalence and descriptive claims, and the high-coverage sample functions as a sensitivity analysis. The full corpus establishes the size, growth, and headline patterns of the field, while the high-coverage sample tests whether those patterns (most importantly the low rate of explanation-quality evaluation) persist among the studies that report most completely across the four review dimensions. This design is important because systematic reviews of fast-growing technical fields often face a trade-off between breadth and extraction depth; here, breadth is preserved by computing every reported statistic on the full corpus, while the high-coverage sample guards against the concern that a headline pattern is merely an artefact of thin reporting in weaker papers.

## 1.2 Research Questions

This review is guided by four research questions:

**Table 1** Research questions guiding this systematic review

| RQ | Research question | Review dimension |
|---|---|---|
| **RQ1** | Across financial applications, which XAI techniques are most commonly applied to which machine learning task types and model families, and how consistently is predictive performance reported alongside them? | Technique–task mapping |
| **RQ2** | What explicit metrics, tests, expert assessments, or user studies are used to evaluate generated SHAP/LIME explanation outputs in finance? | Explanation evaluation |
| **RQ3** | What technical and data-related barriers to deploying XAI in financial systems are reported in the literature, and how do these barriers vary across financial domains? | Deployment barriers |
| **RQ4** | What broader methodological strategies are used to validate or triangulate SHAP/LIME explanations beyond a single attribution output? | Validation strategy |

Our questions shift the focus from simple mapping to rigorous assessment. We first establish how techniques map to specific tasks in RQ1. We then investigate in RQ2 if researchers explicitly evaluate their generated explanations using functional metrics, expert assessments, or human-centered procedures. For RQ3, we analyze the obstacles preventing real-world deployment. These barriers range from computational cost and scalability to data quality, class imbalance, temporal drift, and regulatory alignment. Finally, RQ4 examines broader validation strategies such as causal testing, counterfactual analysis, adversarial checks, and domain-knowledge grounding.

We intentionally linked these four inquiries. A descriptive map of method use from RQ1 cannot tell us if the explanations actually work. We need RQ2 to identify the evaluation metrics, but we also must understand the specific task contexts where those metrics apply. Explanation practice must survive the operational constraints of financial systems. Latency, privacy rules, data access restrictions, drift, and institutional governance dictate whether a method is usable, which we cover in RQ3. We then expand our perspective in RQ4 from individual metrics to comprehensive validation designs. By structuring our review this way, we do not just ask "Which methods are used?" We investigate how practitioners assess them, what blocks their deployment, and what validation strategies might make them trustworthy.

## 1.3 Contributions of the Review

We offer four main contributions to the field of financial machine learning:

1. We map SHAP, LIME, and co-occurring XAI techniques to specific financial domains, machine learning tasks, and model families.
2. We document how researchers evaluate explanations in practice. We track everything from fidelity and stability metrics to expert reviews and user studies, explicitly noting when authors skip explanation-quality assessment entirely.
3. We analyze the technical and data-related barriers preventing deployment. We pinpoint exactly where academic demonstrations fail to meet production requirements across different financial sectors.
4. We categorize how researchers validate XAI outputs. We separate basic metric-level evaluation from comprehensive validation designs like causal reasoning, counterfactual testing, and human-in-the-loop assessment.

We explicitly avoid providing just another broad overview of AI in finance. We instead conduct a strict evidence audit of post-hoc explanations. We track how researchers apply, evaluate, and constrain these tools in financial contexts. We want to know if the academic community is producing evidence that model-risk teams, regulators, and industry practitioners can actually use.

Our findings serve specific needs for different groups. We show researchers which combinations of tasks, models, and explanation methods lack sufficient study. We help industry practitioners distinguish between purely academic XAI experiments and techniques ready for operational deployment. We give regulators and risk teams the tools to separate vague claims of explainability from hard evidence of explanation quality. Finally, we establish a reproducible framework for future reviewers to update this evidence base as new literature gets published.

### 1.4 Paper Structure

Section 2 outlines our conceptual and coding framework. We define the XAI categories, machine learning tasks, financial domains, and evaluation concepts necessary to understand the literature. Section 3 details our systematic review protocol. We specify the search strategy, eligibility criteria, study selection process, data extraction steps, and synthesis methods. We describe the resulting analytic corpus in Section 4. We report our specific findings for each research question in Section 5. Section 6 explores what our results mean for researchers, industry practitioners, and financial regulators. Finally, Section 7 examines threats to validity, and Section 8 concludes our review.

## 2 Background

We built a specific review framework for data extraction and synthesis. Rather than attempting a broad survey of all explainable AI or financial machine learning, we mapped out the conceptual distinctions necessary to interpret the included SHAP and LIME studies. We asked specific questions about each paper. What kind of explanation does the method produce? Which model and task does it explain? How does the financial setting shape the explanation's purpose? We also extracted any evidence the authors offered to prove their explanation is faithful, useful, stable, or deployable.

### 2.1 Conceptual framework for financial XAI

Explainability depends on the specific combination of a model, an explanation, and a user. It is not simply a label you can attach to a method [2, 12, 13]. Because of this, we do not consider a study strong just because it runs SHAP or LIME. We only treat an explanation as evidence-bearing when authors connect it to a trained model, a financial task, a clear governance purpose, and measurable quality criteria. This perspective separates two issues that applied finance papers often mix up. Applying an explanation method is different from evaluating the resulting explanation.

We organized our extraction framework around four distinctions. First, explanations are either intrinsic or post-hoc. Intrinsic models are interpretable by design, covering approaches like linear models and sparse scorecards. Post-hoc methods try to explain a model after training. They add an approximation layer between model behavior and human interpretation. Second, methods are model-specific or model-agnostic. TreeSHAP exploits the internal structure of tree ensembles. LIME and KernelSHAP work with any model that exposes input-output behavior. Third, explanations operate at local or global levels. Local explanations help with case-level reasoning for adverse credit decisions or suspicious transactions. Global explanations support broader model governance and risk reviews. Finally, researchers assess explanation quality either functionally with metrics like fidelity, or empirically through user studies and expert reviews [17–19].

Figure 1 maps the broader field of explainable AI. We use this figure as a scope map rather than an exhaustive taxonomy. SHAP and LIME represent one specific area: post-hoc attribution for trained models. We use them as an eligibility gate because they appear frequently enough in finance to allow systematic cross-study comparison.

We do not claim they represent all of explainability. When authors use counterfactuals, causal explanations, or rule extraction alongside SHAP or LIME, we still code those methods.

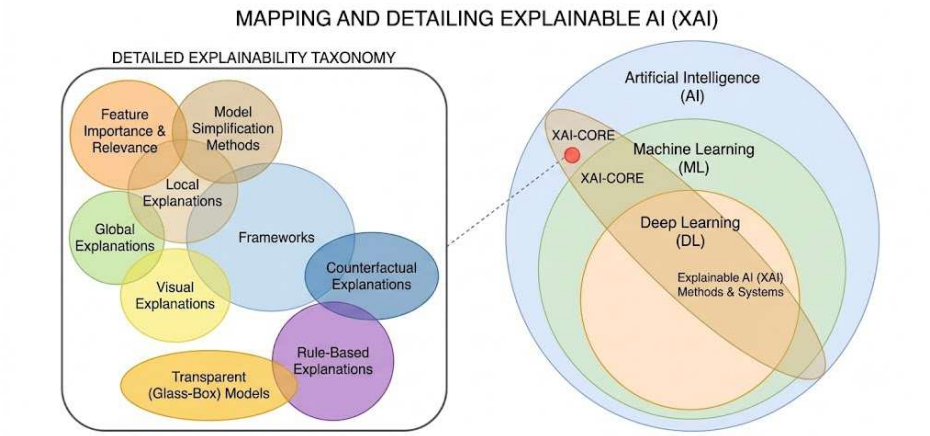



**Fig. 1** Conceptual mapping of explainable AI methods and their position within artificial intelligence, machine learning, and deep learning. Source: authors' synthesis.

## 2.2 SHAP/LIME scope and its implications

LIME [14] explains individual predictions by perturbing inputs near the case of interest. It then observes the model's response and fits a local interpretable surrogate. The method offers intuitive local outputs and works across different models. However, we found it carries significant risks regarding perturbation design and neighborhood definitions. SHAP [15] takes a different approach by assigning feature attributions based on Shapley-value logic from cooperative game theory. It provides additive attribution and allows for efficient computation on tree ensembles via TreeSHAP [16]. Yet it also struggles with computational costs for model-agnostic approximations and sensitivity to background distributions. Users might also mistakenly read association-based attributions as causal explanations [20, 21].

Using SHAP and LIME as filters creates a coherent review population while narrowing our claims. We review SHAP/LIME-based post-hoc explanations in finance, keeping broader context visible through our coding fields. When we discuss "financial XAI" in later sections, we mean this specific SHAP/LIME-gated subpopulation unless we state otherwise. We do not claim to cover every explainability method. Instead, we want to audit a large, comparable body of applied explanation studies. We ask if they produce evidence strong enough for regulated financial use.

## 2.3 Financial decision context

Financial explainability differs from generic model interpretation because explanations tie directly to institutional accountability. Different stakeholders often need entirely

different explanations from the same model. Credit officers, auditors, and customers look for different signals. A feature-attribution plot that helps a data scientist debug a classifier will rarely satisfy a customer-facing adverse-action requirement. A local fraud explanation often falls short for portfolio-level model governance. Because of this, previous literature treats financial explainability as both a technical and governance challenge [3, 7–9].

Recent banking-focused reviews frame the field as a socio-technical problem. Explanations need to be technically reliable and understandable to different stakeholder groups. They must embed deeply into governance processes instead of functioning merely as post-hoc visualisations [22, 23].

Different financial domains create distinct explanation requirements. Credit and lending focus heavily on local recourse, fairness, and model-risk documentation [4, 24–28]. Fraud and anti-money-laundering teams care about analyst triage under severe class imbalance and false-positive costs [29, 30]. Trading and forecasting demand attention to temporal dependence, regime changes, and latency. Auditing focuses on traceability and defensible reporting. FinTech introduces unique customer-facing challenges, particularly when institutions use alternative data and behavioral signals [31].

We designed our coding strategy around these domain differences. We code each study by financial domain, machine learning task, model family, and explanation method. We also record evaluation evidence, deployment barriers, and validation strategies. This approach stops us from treating all SHAP and LIME applications as equivalent. A SHAP summary plot for a credit model and a LIME explanation for a fraud alert answer fundamentally different operational questions. They demand different standards of evidence.

## 2.4 Task and model dimensions

We group financial machine learning tasks into several categories. These include classification, regression, time-series forecasting, anomaly detection, and reinforcement learning. Classification covers tasks like default prediction, loan approval, and suspicious-transaction screening. Regression handles continuous estimates for claim severity, volatility, and prices. Researchers typically use anomaly detection and clustering for fraud, market surveillance, and regime detection. Reinforcement learning shows up mostly in sequential decisions for trading, market making, and dynamic pricing.

We code the model family separately because a model's architecture changes how explanations behave. Linear models are fairly transparent but often underfit complex financial data. Tree ensembles and gradient boosting models dominate tabular financial prediction. They match perfectly with TreeSHAP. Neural networks and transformer models work well for sequential or high-dimensional data. However, their explanations are harder to validate and cost more to compute. We map tasks to models to answer RQ1 and interpret deployment readiness. Evidence showing SHAP works well for tree-based credit classification does not mean it works for neural time-series forecasting.

### 2.5 Evaluation and validation concepts

We distinguish explanation evaluation from validation strategy throughout this paper. *Explanation evaluation* means explicitly measuring the quality of a generated explanation. Researchers use functional metrics like fidelity, faithfulness, stability, and consistency to move beyond anecdotal evidence [6]. They also gather human-centered evidence through expert reviews, user studies, and decision-support experiments [17–19]. *Validation strategy* takes a broader view. It covers study-level designs that build confidence in the explanation itself or how users apply it. These strategies include causal analysis, domain-knowledge checks, adversarial tests, and multi-method triangulation.

This separation drives our review. Simply applying two explanation methods does not count as evaluation. SHAP and LIME might agree, but that agreement does not prove either explanation accurately represents the model. Generating a visually plausible attribution plot falls short of validation. We consider an explanation strong evidence only when researchers test it properly. They need to show it reflects actual model behavior or remains stable under perturbation. Real explanations must support user decisions, align with domain knowledge, and survive real-world latency and governance constraints.

### 2.6 Related reviews and remaining gap

The existing review literature on financial explainability is substantial. We position our manuscript as a targeted evidence audit rather than a first mapping exercise. Table 2 summarizes the most relevant review streams by scope, corpus logic, and evaluation focus. We do not rank prior reviews. We simply show where our work provides a distinct, corpus-level audit of explanation evidence.

Previous reviews establish why explainability matters in finance and document the available methods. We ask a more adversarial question. Does the empirical literature produce explanation evidence strong enough for high-stakes financial use? We address this by linking methods to specific tasks, models, domains, and evaluation metrics. Our contribution goes beyond running a newer search date. We separate method use from explanation evidence. We assess whether financial studies have moved past merely producing explanations to actually validating them.

## 3 SLR Methodology

### 3.1 Review Protocol

We designed our systematic literature review around Kitchenham and Charters' guidance for software engineering [37]. We also incorporated PRISMA 2020 reporting principles [38]. Before assessing full texts for eligibility, we specified our research questions, databases, search strings, and eligibility criteria. Our protocol also defined the screening process, data-extraction fields, reporting-transparency rubric, and synthesis strategy. We maintained this protocol internally. We did not prospectively register it in PROSPERO, OSF, or another public registry, which we acknowledge as a transparency limitation. We kept all eligibility criteria constant once full-text screening

**Table 2** Positioning of this review against related reviews and survey streams

| Review stream | Sources | Typical scope | Corpus logic | Evaluation focus | Added focus here |
|---|---|---|---|---|---|
| Broad XAI-in-finance reviews | [7–9] | Financial XAI methods, applications, and high-level challenges. | Broad method scope. | Usually narrative or high-level. | Joint audit of method use, evaluation, barriers, and validation. |
| Model-agnostic XAI reviews | [10] | SHAP, LIME, counterfactuals, and other model-agnostic methods. | Method-oriented. | Method properties. | Finance corpus prevalence of evaluated SHAP/LIME explanations. |
| Explanation-evaluation reviews | [6] | XAI evaluation properties and metrics across AI. | Cross-domain. | Central focus. | Finance-specific measurement of whether generated explanations are evaluated. |
| Domain-specific reviews | [24, 29] | Credit risk, lending, anti-money-laundering, or related subfields. | Domain-bounded. | Varies by domain. | Cross-domain comparison in retrieved SHAP/LIME corpus. |
| Regulatory, ethical, and managerial reviews | [31–33] | Compliance, trust, fairness, organisational decision-making, and responsible finance. | Conceptual or managerial. | Governance needs. | Separation of explanation presence from explanation-quality evidence. |
| Banking implementation and regulatory-integration reviews | [22, 23] | Adoption, stakeholder needs, regulatory pressure, and governance requirements. | Banking or financial-services bounded. | Barriers and governance. | Links barrier reporting to explanation-output evaluation. |
| Bibliometric and taxonomic reviews | [11, 34, 35] | Publication clusters, taxonomies, and field structure. | Bibliometric or taxonomy-led. | Limited substantive coding. | Study-level coding of metrics, barriers, and validation strength. |
| Task-specific or emerging-method reviews | [36] | Specialized settings such as financial time-series explainability. | Task-bounded. | Task-specific. | Cross-task audit of SHAP/LIME use and validation practice. |

began. Later clarifications only refined our coding definitions. In particular, we clarified the distinction between papers that merely mention SHAP/LIME and those that actually apply either method to trained model outputs.

We executed the review in three distinct phases. During planning, we defined the SHAP/LIME-gated scope, formulated four research questions, and built the extraction codebook. We then moved to the conduct phase. Here we searched Scopus, Web of Science, and IEEE Xplore. We screened records using Rayyan, retrieved full texts, extracted study-level data, and appraised reporting transparency. Finally, in the reporting phase, we aggregated descriptive evidence from the full corpus. We identified a high-coverage subset for sensitivity analysis and structured our findings around the four research questions.

### 3.1.1 Protocol, registration, and PRISMA reporting

Table 3 maps the main PRISMA 2020 elements to specific manuscript sections and supplementary materials. We use the extraction workbook as our primary audit file. It contains the study register, codebook, PRISMA log, and verification checks. Appendix B describes the fields within this workbook. The completed PRISMA checklist is provided as supplementary material. We made this choice because many reporting elements, such as exclusion reasons and detailed extraction conventions, take up too much space for the main text.

**Table 3** PRISMA 2020 reporting map for this review

| PRISMA element | Where addressed | Status |
|---|---|---|
| Rationale and objectives | Introduction, scope, and research questions | Reported |
| Eligibility criteria | Section 3.3; Table 4 | Reported |
| Information sources and search strategy | Section 3.2; Appendix A | Reported |
| Selection process | Section 3.4; Figure 2 | Reported |
| Data-collection process and data items | Section 3.5; Appendix B | Reported |
| Study appraisal / reporting-transparency assessment | Section 3.6; Table 6 | Reported |
| Synthesis methods | Section 3.7 | Reported |
| Excluded full texts with reasons | Figure 2; extraction workbook PRISMA log | Reported in supplement |
| Protocol registration | Section 3.1.1 | Not prospectively registered; limitation stated |
| Certainty assessment | Not applicable; heterogeneous technical SLR without pooled effect estimate | Not applicable |

## 3.2 Search Strategy

### 3.2.1 Databases

We searched three databases to ensure comprehensive coverage across finance, computer science, engineering, and information systems:

- **Scopus**: provides broad coverage of finance and computer-science journals and conference proceedings.
- **Web of Science**: indexes high-quality, peer-reviewed sources across many disciplines.
- **IEEE Xplore**: offers strong coverage of computer-science and engineering conferences.

We chose these databases because they support advanced query syntax and bibliometric export. They also cover the main archival venues for financial machine learning and XAI. We restricted eligibility to peer-reviewed journal articles and conference proceedings. During screening and full-text assessment, we excluded preprints, editorials, book chapters, surveys, reviews, position papers, and other non-primary publications.

### 3.2.2 Search terms

We constructed search strings by combining two concept blocks with AND. Within each block, we combined alternative terms using OR. Block A (Finance): "Asset evaluation", "banking", "credit risk", "finance", "financial performance", "portfolio optimi*", "portfolio selection", "financial planning", "bankruptcy", "credit scoring", "financial distress", "loan default", "Financial Advice", "Financial Market", "Option pricing". Block B (XAI, ML): "explainable artificial intelligence", "xai", "interpretable machine learning", "explainable ai", "explainable machine learning", "interpretability", "ai transparency", "SHAP", "LIME". We designed the finance block to retrieve a broad, finance-labelled SHAP/LIME corpus. We did not attempt to enumerate every possible application label. The query captures many studies in fraud, anti-money laundering, trading, insurance, accounting, and risk management when those studies use broader terminology. It may miss studies indexed only under narrower subdomain labels. We deliberately omitted terms like "fraud", "money laundering", "AML", "insurance", and "anomaly detection" to keep the search focused on credit and finance. We chose not to expand the query after screening, as this would have changed the protocol post hoc. We therefore treat the corpus as search-string bounded. Any prevalence claims apply only to this retrieved SHAP/LIME corpus. Readers should interpret fraud/AML and insurance findings as lower-bound evidence rather than completeness claims about those subfields. Appendix A provides the exact query for each database, including field tags and syntax adjustments.

### 3.2.3 Language

We restricted the search to English-language publications.

### 3.2.4 Screening and coding process

Before screening began, we converted the criteria in Table 4 into a shared decision guide. This guide defined a genuine financial modelling task, a genuine SHAP/LIME application, and a peer-reviewed empirical study. It also detailed the evidence needed for each research question. We retained borderline title-and-abstract records for full-text review to minimize false exclusions.

Two reviewers independently screened titles and abstracts. They then independently assessed full texts for eligibility. We resolved disagreements through discussion against the written criteria. A third reviewer adjudicated unresolved cases. The two primary reviewers and the adjudicator belonged to the author team. They worked from the same criteria but made initial decisions independently. We did not blind reviewers to bibliographic metadata, as we needed venue and document-type information for eligibility checks. This procedure created a clear exclusion audit trail. We paid special attention to papers that mentioned SHAP or LIME in related work but failed to apply either method to model outputs.

We did not compute a formal inter-coder reliability statistic for screening or extraction. We therefore rely on procedural consistency rather than quantitative reliability evidence. We supported this consistency through independent screening, written guides, consensus discussions, and third-reviewer adjudication. For extraction, two reviewers piloted the codebook on included studies before the final coding pass. We clarified ambiguous categories in the codebook, especially validation strategy, and coded them as consensus-adjudicated fields. The final validation-strategy codes represent reconciled judgements rather than independently double-coded labels.

We used a single codebook for data extraction across all included studies. After extraction, we checked the workbook for internal consistency. We compared technique labels with method descriptions, task labels with target variables, and domain labels with financial applications. We also verified RQ coverage against extracted evidence and checked barrier codes against supporting text. This verification pass targeted three specific errors: false-positive SHAP/LIME inclusion, confusing predictive performance with explanation-quality evaluation, and coding asserted barriers as empirically tested barriers.

## 3.3 Selection Criteria

We assessed each record retained after title-and-abstract screening against the criteria in Table 4. For inclusion, a paper had to meet all five conditions. It needed a genuine financial modelling task, empirical model results, and peer-reviewed status. It also had to be published in 2016 or later and demonstrate actual use of SHAP and/or LIME as a post-hoc explanation method. We excluded papers that only discussed explainability conceptually. We also rejected studies that used intrinsically interpretable models without post-hoc explanation, mentioned SHAP/LIME only in passing, or presented a review rather than primary empirical results.

**Table 4** Inclusion and exclusion criteria.

| Inclusion criteria | Exclusion criteria |
|---|---|
| Addresses a core financial modelling task (e.g. credit scoring, fraud detection, distress prediction, forecasting, risk assessment). | Concerns XAI philosophy or general theory with no financial application. |
| Genuinely applies SHAP and/or LIME as a post-hoc method, not merely citing them. | Applies AI in finance but uses no XAI method. |
| Reports empirical results. | Mentions SHAP or LIME only in passing, without applying either. |
| Published in a peer-reviewed journal or conference proceedings. | Survey, review, editorial, book, book chapter, or opinion piece. |
| Published in 2016 or later. | Published before 2016. |

## 3.4 Study Selection Process

We structured our study selection around the PRISMA 2020 flow shown in Figure 2. The process moved through identification, title-and-abstract screening, full-text retrieval, eligibility assessment, inclusion, and finally high-coverage subset assignment.

*Identification.* We executed the search strings described in Section 3.2 in May 2026. The three databases returned 4,168 records in total. We imported all records into Rayyan to manage the screening process. We removed 1,158 duplicate records through automated and manual de-duplication. This left 3,010 unique records for screening.

*Screening.* We screened the 3,010 unique records at the title-and-abstract level against the criteria in Section 3.3. We retained a record only if the title or abstract indicated both a genuine financial application and an explainability method. We excluded clearly out-of-scope records. These included non-financial applications, purely theoretical pieces, and studies lacking an explainability component. This initial screening removed 2,685 records. We sought the remaining 325 reports for full-text retrieval.

*Retrieval and eligibility.* We could not obtain 13 of the 325 reports through institutional access, interlibrary channels, or direct author contact. We recorded these as not retrieved. We assessed the remaining 312 full texts against our eligibility criteria. For inclusion, a study had to meet three simultaneous conditions. First, it had to address a core financial modelling task, such as credit scoring, fraud detection, or risk assessment. Second, it needed a genuine post-hoc explainability method. Third, it had to explicitly apply SHAP and/or LIME to the trained model. The requirement to actually *apply* the methods proved decisive. Merely mentioning them in related work or using them as a baseline was insufficient. At this stage, we excluded 13 reports that cited SHAP or LIME without applying them. We could not resolve two cases due to unreadable source files. This process yielded a final analytic corpus of 297 studies.

*High-coverage subset as a sensitivity analysis.* We compute all prevalence claims on the full corpus of 297 studies. This includes distributions for publication years, domains, tasks, techniques, dataset types, and deployment barriers. The full corpus serves as the denominator whenever we state how often a phenomenon occurs.

We constructed a 97-study high-coverage subset for a more specific purpose. It functions as a sensitivity analysis. We used it to test whether the low corpus-wide rate of explanation-quality evaluation simply results from incomplete reporting. A study entered this subset if it contained RQ1 evidence plus explicit evidence for at least two of RQ2, RQ3, and RQ4. This yielded 84 studies with evidence for three RQs and 13 with evidence for all four. We do not treat studies outside the subset as low quality. Many are narrower empirical applications that map to RQ1 but lack the detail to enter the subset. Because membership required extensive reporting, the subset leans towards studies that evaluate their explanations. It provides an upper-bound check rather than an unbiased estimate. We never use the subset as a denominator for field-wide prevalence. Readers should interpret its composition against the full-corpus distributions in Section 4. Table 5 outlines how we divide labour between the two evidence bases.

**Table 5** Use of the full corpus and high-coverage subset in the synthesis

| Evidence base | Size | Membership rule | Use in this review |
|---|---|---|---|
| Full included corpus | $n = 297$ | Included after full-text eligibility; genuine SHAP/LIME application in finance. | Denominator for *all* prevalence, descriptive, and headline claims (technique, domain, task, dataset, and explanation-evaluation rates). |
| High-coverage subset | $n = 97$ | RQ1 evidence plus explicit evidence for at least two of RQ2, RQ3, and RQ4 (84 with three RQs, 13 with four). | Sensitivity analysis only: tests whether the low evaluation rate persists among the most fully reported studies. Never a prevalence denominator; biased towards RQ2/3/4 reporters, providing an upper-leaning check rather than an independent estimate. |

## 3.5 Data Extraction

We extracted data into a structured workbook developed specifically for this review. For each study, we coded bibliographic metadata, financial domain, sub-application, and dataset type. We also tracked the machine learning task, model family, XAI technique, and co-present explanation methods. We piloted this form on a small subset to refine our category definitions before proceeding. To verify our extraction reliability, two reviewers independently double-coded a stratified random subsample of 50 studies ($\approx 17\%$ of the corpus). Observed agreement exceeded 92% across all key fields. We found strong chance-corrected agreement for explanation evaluation tier ($P_o = 96.0\%$, Cohen's $\kappa = 0.83$, Gwet's $\mathrm{AC}_1 = 0.95$), deployment barriers ($P_o = 94.0\%$, $\kappa = 0.81$, $\mathrm{AC}_1 = 0.89$), and validation strategy ($P_o = 92.0\%$, $\kappa = 0.78$, $\mathrm{AC}_1 = 0.86$). We report Gwet's $\mathrm{AC}_1$ alongside $\kappa$ to correct for low-prevalence skew in rare codes.

We applied strict definitions during extraction. *Genuine SHAP/LIME use* meant visible application to trained model predictions. Mere citation in related work did not qualify. For *explanation-quality evaluation*, we looked for an explicit metric,

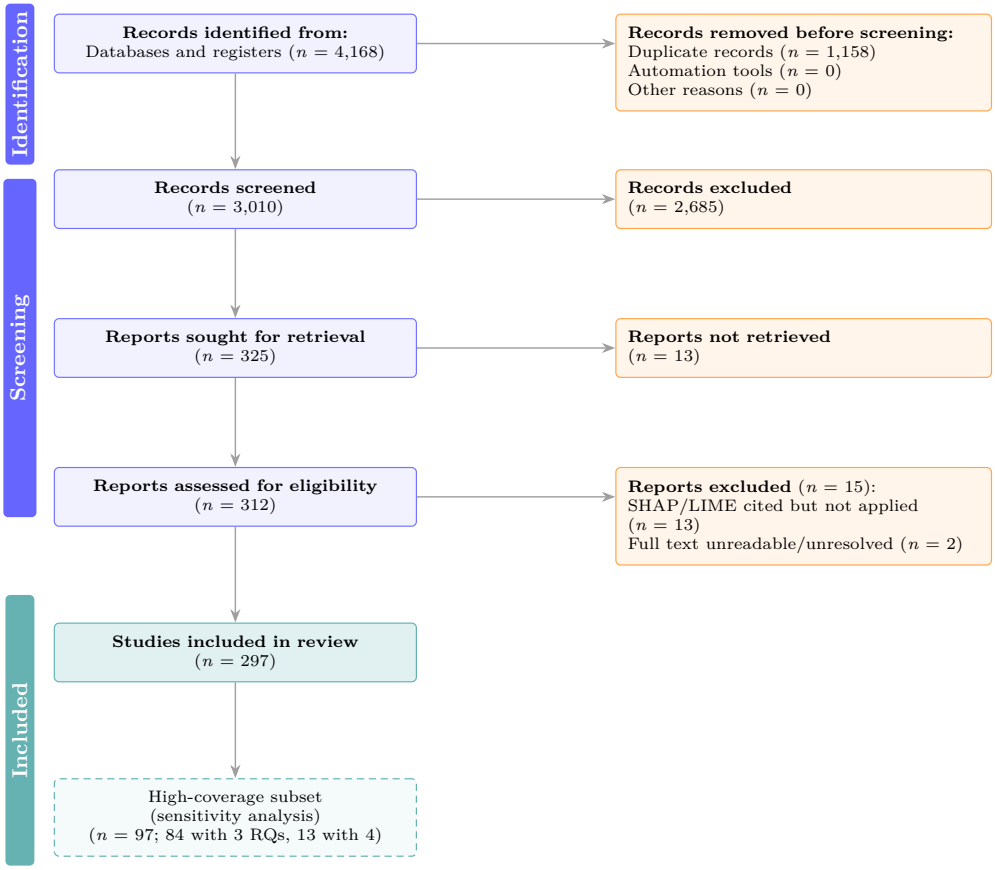



**Fig. 2** PRISMA 2020 flow diagram of study selection.

expert assessment, user study, or structured qualitative judgement about the explanation itself. Predictive accuracy or AUC alone did not count. *Formal XAI evaluation* required a reproducible design, such as fidelity testing, robustness checks, or trust-calibration tests. We coded a *partial explanation-quality signal* when we found limited evidence, like qualitative expert inspection without a full protocol. We flagged a *deployment barrier* whenever a study discussed technical, data, latency, privacy, or regulatory obstacles. Finally, we coded a *validation strategy* only if a study used a broader design to build confidence in the explanation. Examples include causal analysis, counterfactual reasoning, adversarial testing, or multi-method triangulation.

These fields map directly to our research questions. Technique, model family, and financial domain support RQ1. Explanation-quality metrics support RQ2. Barrier fields align with RQ3, while validation-strategy fields address RQ4. We derive RQ coverage strictly from extracted data. A study only contributes to a research question if it provides explicit evidence.

## 3.6 Reporting Transparency and Appraisal Signals

The corpus consists of heterogeneous applied XAI studies rather than controlled experiments. We therefore did not apply a clinical-style risk-of-bias instrument or rank overall scientific quality. We instead appraised the specific reporting signals needed for our synthesis. We checked each study for a clear dataset and financial task. We looked for extractable model-pipeline information and genuine SHAP/LIME use. We also verified reported predictive performance and methodological transparency. If we could not confirm actual SHAP/LIME use from the full text, we excluded the study during eligibility assessment.

Table 6 details our rubric for judging reporting transparency. We coded each dimension as 0 for unclear reporting, 1 for partial reporting, or 2 for adequate reporting. Readers should treat these scores as evidence-usability signals rather than definitive quality metrics. We did not want to exclude studies just for incomplete reporting. Weak reporting actually forms a key part of the evidence base. We used these signals to qualify our synthesis claims. A study might support a descriptive claim about SHAP usage without proving explanation validity.

**Table 6** Reporting-transparency and extractability rubric for included XAI-finance studies

| Dimension | Adequate extraction signal | Weak or unusable signal |
|---|---|---|
| Financial task clarity | Financial domain, target variable, and decision context are clearly described. | Financial application is vague or only loosely connected to model output. |
| Dataset transparency | Dataset source, sample structure, feature type, and public/proprietary status are stated. | Dataset provenance, sample size, or feature context is missing. |
| Core modelling signal | Model family and predictive-performance evidence are extractable; stronger cases also report training setup, validation split, and implementation detail. | Model family or predictive-performance evidence cannot be extracted, or the pipeline is too under-specified to support interpretation beyond a basic method label. |
| Genuine SHAP/LIME use | SHAP/LIME outputs are applied to trained model predictions and are inspectable in text, tables, or figures. | SHAP/LIME are mentioned only conceptually or used without visible explanation output. |
| Explanation evaluation | Fidelity, stability, robustness, comprehensibility, usefulness, expert review, or user-study evidence is reported. | Explanation plots are shown without any quality assessment. |
| Deployment relevance | Technical, data, governance, latency, privacy, drift, scalability, or regulatory barriers are discussed or tested. | Study treats explanation as a visualization step with no deployment context. |
| Validation strategy | Explanation is checked against domain knowledge, causal reasoning, counterfactual tests, robustness analysis, or human judgment. | Explanation is interpreted as self-validating. |

Table 7 outlines the reporting-transparency signals from our extraction workbook. We found strong signals for task, dataset type, model family, and genuine SHAP/LIME application. This makes sense since these fields acted as eligibility gates. These signals prove extractability for review coding, though they do not guarantee full pipeline reproducibility. We found the weakest signals around explanation-quality evaluation. Only 21 studies report formal XAI evaluation. Six more provide limited signals without a formal design. This pattern reinforces our later conclusion that many researchers apply explanation methods without validating the results.

During a final verification pass, we checked the internal consistency of each extraction record. We compared technique codes, quality signals, and RQ coverage. We watched closely for three specific failure modes. We checked for papers citing SHAP/LIME without applying them. We looked for papers confusing predictive performance with explanation quality. We also identified papers making deployment-barrier claims without empirical support. While these checks support auditability, they do not guarantee error-free extraction across the entire corpus.

### 3.7 Synthesis Strategy

We found massive heterogeneity across datasets, model families, and task formulations. This made quantitative meta-analysis impossible. We rely on descriptive and narrative synthesis instead. For RQ1, we aggregated technique, task, and performance fields into frequency tables. For RQ2, we separated explanation-quality evidence into formal evaluation and partial signals. For RQ3, we grouped technical and data barriers thematically. For RQ4, we categorized validation strategies by the assurance they provide. These range from low-assurance method agreement to higher-assurance causal or adversarial testing.

We did not pool predictive performance results. Studies use completely different datasets, class balances, and metrics. We use performance evidence purely descriptively. It helps us see whether studies report model quality before trying to interpret explanations. We treat explanation evidence separately. High predictive performance does not guarantee a faithful or useful explanation.

## 4 Descriptive Analysis of the Corpus

This section characterises the 297 studies in the corpus before we present the research-question findings in Section 5. We provide descriptive context rather than explanatory analysis. We profile the temporal, domain, task, method, dataset, and evidence-coverage patterns of the included literature. These patterns establish the empirical background for our later synthesis. They reveal where financial XAI research clusters and where evidence remains thin. We report all percentages out of the 297 included studies unless we state otherwise.

### 4.1 Publication trends over time

The corpus is heavily concentrated in recent years. Of the 297 studies, 95 were published in 2026 and 108 in 2025. Roughly 68% of the corpus appeared in these two most

**Table 7** Reporting-transparency and extractability signal summary across the included studies ($n = 297$)

| Dimension | Extractable | Partial | Weak/unclear | Main implication |
|---|---|---|---|---|
| Financial task clarity | 278 (93.6%) | 19 (6.4%) | 0 (0.0%) | Most studies clearly fit a financial task; borderline cases were retained only where the financial application remained extractable. |
| Dataset-type signal | 297 (100.0%) | 0 (0.0%) | 0 (0.0%) | Dataset source/type was extractable for all included studies, although extraction does not imply full dataset reproducibility. |
| Core modelling signal | 297 (100.0%) | 0 (0.0%) | 0 (0.0%) | Broad model family and predictive-performance evidence were extractable for all included studies; this does not imply that full training pipelines, hyperparameters, code, or validation procedures were reproducible. |
| Genuine SHAP/LIME use | 297 (100.0%) | 0 (0.0%) | 0 (0.0%) | The full-text eligibility gate removed papers that only mentioned SHAP/LIME. |
| Explanation evaluation | 21 (7.1%) | 6 (2.0%) | 270 (90.9%) | Explanation validation is the weakest reporting area in the corpus. |
| Deployment relevance | 280 (94.3%) | 12 (4.0%) | 5 (1.7%) | Most included studies situate their work in a deployment context; a codeable technical, data, or domain barrier is reported in 215 (72%) of studies (Section 5.3). |
| Validation strategy | 106 (35.7%) | 0 (0.0%) | 191 (64.3%) | Broader validation beyond showing explanation plots remains limited. |

recent years. A longer tail extends across 2024 (41), 2023 (23), 2022 (13), 2021 (11), 2020 (5), and 2019 (1). We found no eligible studies published between 2016 and 2018 after our full-text eligibility assessment, despite our 2016 onward inclusion window. The 2026 count requires cautious interpretation because we executed the search in May 2026. The 95 records dated 2026 reflect partial-year indexing, online-first publication, and database date assignment rather than a complete calendar year. We see

**Table 8** Summary statistics of the included studies

| **Main information about the corpus** | |
|---|---|
| Time interval | 2019–2026 |
| Publication sources (unique venues) | 225 |
| Included papers | 297 |
| Papers with DOI | 291 (98.0%) |
| Average age of paper, measured in 2026 | 1.36 |
| **Authors** | |
| Unique author-name strings | 1,020 |
| Author occurrences | 1,073 |
| Single-authored papers | 27 |
| Average authors per paper | 3.61 |
| Median authors per paper | 3 |
| **Dataset types** | |
| Public datasets | 241 |
| Proprietary datasets | 38 |
| Mixed public/proprietary datasets | 14 |
| Synthetic datasets | 4 |
| **XAI technique and evaluation coverage** | |
| SHAP only | 215 |
| SHAP and LIME | 64 |
| LIME only | 18 |
| Formal XAI evaluation done | 21 (7.1%) |
| Explanation quality reported | 27 (9.1%) |

that applied SHAP/LIME work in finance is recent and growing rapidly. However, one should not compare the 2026 bar mechanically with complete earlier years.

Financial XAI has transitioned from rare empirical demonstrations to a dense, fast-expanding literature. We see few studies before 2023. This suggests early work was exploratory, typically showing that researchers could attach post-hoc explanation methods to financial models. The lack of included 2016–2018 studies stems directly from our eligibility screening. It does not imply that no finance-adjacent interpretability work existed then. Studies from that early window either missed our database query or failed the final inclusion gate. They did not apply SHAP/LIME to trained model outputs for a peer-reviewed empirical financial task. Prominent 2020 finance-XAI work, such as Bussmann et al. [3], falls after this empty early interval. We use it here as contextual background. We observe a sharp rise after 2024. Empirical finance papers now routinely include SHAP or LIME. This happens especially when predictive models are hard to justify without explanation.

This recency affects our synthesis methodologically. The corpus is highly relevant to current financial AI practice. Yet many studies are too new for longitudinal validation, replication, or operational deployment evidence. The large 2026 tail makes exact replication date-sensitive. Rerunning the search later in 2026 would capture newly indexed records. We conclude from this publication trend that the field is expanding quickly. More applications, however, do not guarantee mature explanation evaluation or deployment readiness.

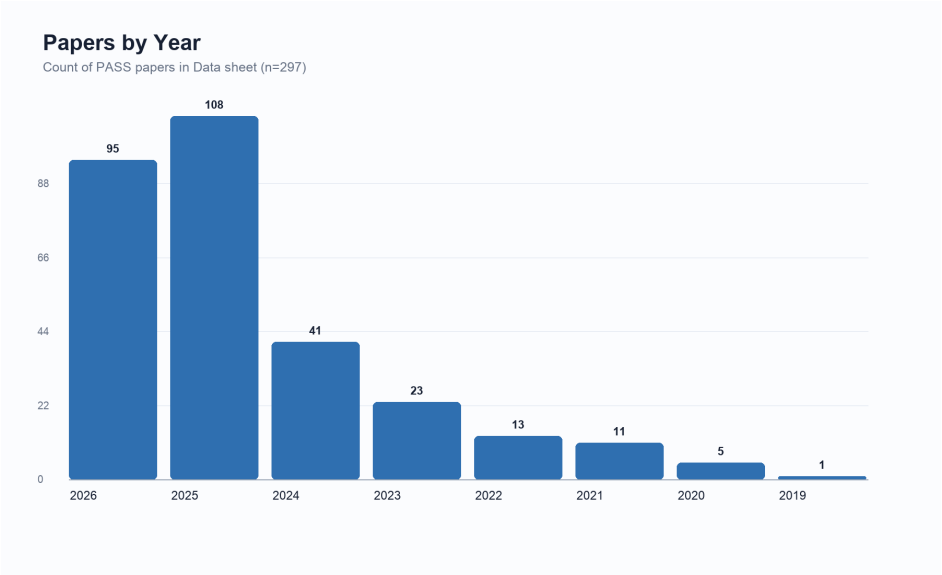



**Fig. 3** Publication-year distribution of included SHAP/LIME financial XAI studies ($n = 297$).

Table 9 cross-tabulates publication year with explanation-quality evidence. Because the early cells are very small, especially 2019-2022, they should not be interpreted as a monotonic time trend. The more important and durable pattern is that the evaluation has not caught up with the publication growth: in the two largest and most recent cells, 2025 and partial-year 2026, the formal explanation evaluation appears in only 8 of 108 studies (7.4%) and 6 of 95 studies (6.3%), respectively. Thus the evaluation gap is not confined to the early exploratory literature; it persists in the recent expansion of SHAP/LIME-based finance papers.

**Table 9** Publication year by explanation-quality evidence in the included corpus

| Year | Studies | Formal XAI eval. | % | Any quality signal | % |
|---|---|---|---|---|---|
| 2019 | 1 | 1 | 100.0 | 1 | 100.0 |
| 2020 | 5 | 1 | 20.0 | 2 | 40.0 |
| 2021 | 11 | 1 | 9.1 | 1 | 9.1 |
| 2022 | 13 | 2 | 15.4 | 3 | 23.1 |
| 2023 | 23 | 1 | 4.3 | 2 | 8.7 |
| 2024 | 41 | 1 | 2.4 | 1 | 2.4 |
| 2025 | 108 | 8 | 7.4 | 11 | 10.2 |
| 2026 | 95 | 6 | 6.3 | 6 | 6.3 |

## 4.2 Distribution by financial domain

Within our retrieved sample, credit, loan, and default modelling is the most common domain (140 studies, 47.1%). Markets, trading, and forecasting follows (58,

19.5%), alongside fraud, anti-money-laundering, and anomaly detection (52, 17.5%). Financial-distress and bankruptcy prediction (20, 6.7%) and generic finance/revisit studies (27, 9.1%) make up the remainder. These categories sum to the full included corpus ($n = 297$). We treat these domain proportions as lower-bound evidence within our search vocabulary, not an exhaustive census of all financial subdomains. The generic finance category groups cross-domain or methodological studies. We note that the concentration in credit, markets, and fraud matches the emphasis reported by prior reviews.

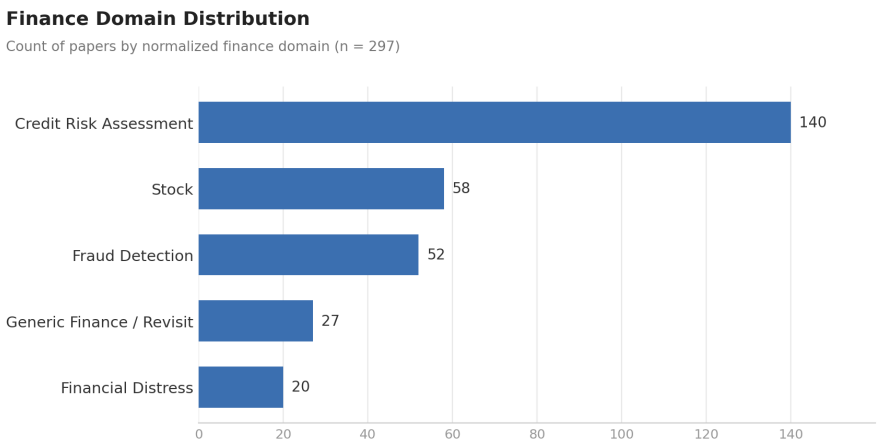



**Fig. 4** Distribution of included studies across financial domains ($n = 297$). Percentages represent proportions within the retrieved search-string-bounded sample.

We unsurprisingly find credit, loan, and default modelling dominant. These tasks combine high predictive value with strict explanation requirements. Individual and institutional consequences are severe for credit decisions. Models face internal risk review and require defensible reasoning during challenges. Consequently, credit serves as a natural testbed for SHAP and LIME. Researchers typically use tabular data and well-defined targets. They frame explanations around familiar borrower or account-level features.

Markets, trading, and forecasting represent the second-largest domain. They pose completely different explainability problems. Researchers use explanations here to interpret predictive signals, feature contributions, or model sensitivity. They rarely justify individual customer-facing decisions. Market data is non-stationary and time-dependent. This makes explanations harder to validate. Feature importance shifts across regimes. Fraud, anti-money-laundering, and anomaly detection introduce a third profile. Here, explanations must support analyst triage. The context involves class imbalance, evolving behaviour, and high false-positive costs.

Financial XAI remains unevenly distributed across the retrieved corpus. We see smaller representation for financial-distress, bankruptcy, insurance, and accounting applications. Our search vocabulary omitted several narrower subdomain labels,

particularly insurance-specific terms. Therefore, one should not interpret under-representation as definitive measurement of those subfields. These areas likely have substantial practical explanation needs. Our protocol simply retrieved fewer empirical studies meeting the criteria. This unevenness shapes our findings. Our conclusions are most robust for credit, markets, and retrieved fraud/AML studies. They remain tentative for less frequently retrieved domains.

Figure 5 shows that the post-2024 expansion was not confined to a single domain. Credit-risk studies increased most sharply, but stock-market and fraud-detection studies also expanded in 2025 and 2026. The field's growth is therefore broad, although still centred on a small number of high-volume application areas.

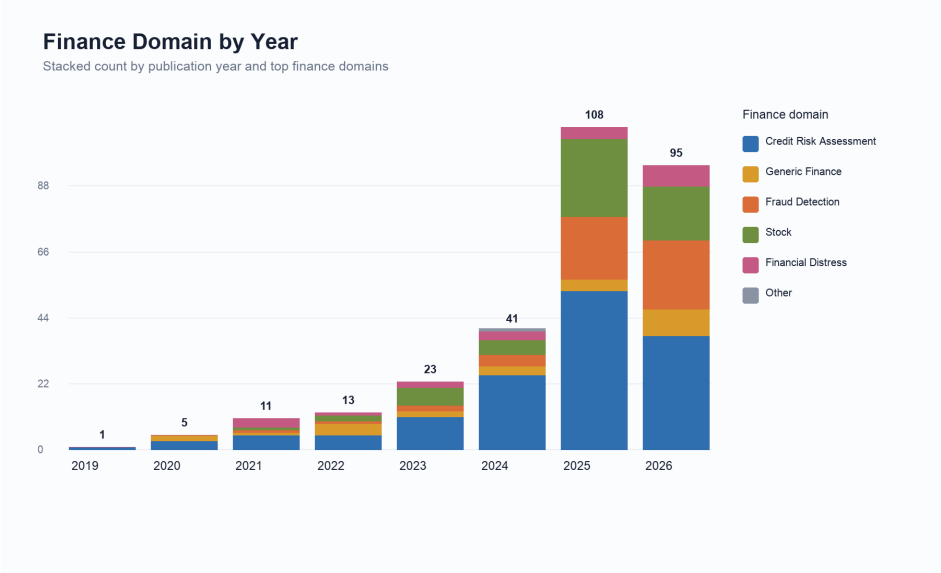



**Fig. 5** Financial-domain composition by publication year. The stacked bars show the main domains contributing to the rapid growth of the corpus after 2024.

## 4.3 Distribution by ML task type

Classification dominates the corpus (219 studies, 73.7%). This aligns with the prevalence of credit and fraud problems. Time-series forecasting accounts for 35 studies (11.8%), regression for 20 (6.7%), and anomaly detection for 8 (2.7%). The remaining 15 studies (5.1%) use multi-task or specialized formulations.

Supervised tabular prediction currently shapes financial XAI research. Researchers favour classification tasks for their clear labels and familiar performance metrics. They can explain local decisions using feature-attribution methods. Many standard financial applications fit this structure. Default prediction, fraud detection, customer churn, loan approval, and suspicious-transaction classification all work this way. SHAP and LIME apply with minimal adaptation. Users can interpret explanation outputs as feature contributions to a discrete decision.

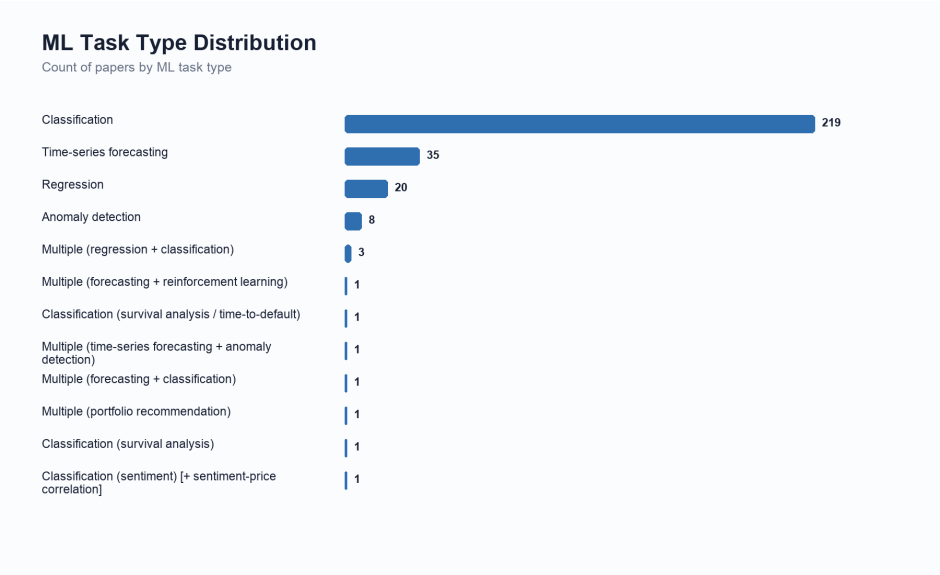



**Fig. 6** Machine-learning task types represented in the corpus. Classification dominates, while forecasting, regression, and anomaly-detection formulations are much less frequent.

Continuous and temporally structured financial problems remain less developed. We see smaller shares for time-series forecasting and regression studies. These tasks demand explanations for both which variables matter and when they matter. The influence of variables changes across time. Standard feature-attribution summaries often fail here. They struggle when explaining a price trajectory, volatility estimate, portfolio signal, or macro-financial forecast.

Anomaly detection also rarely appears as a distinct formulation. It has massive practical importance in fraud and anti-money-laundering, yet we see few studies. Explaining unsupervised or weakly supervised models is difficult. There is no stable class label to assess explanation quality against. We must avoid treating financial XAI as a single uniform field. Methods mature for classification are often unsettled for forecasting, regression, anomaly detection, or multi-task settings.

## 4.4 XAI techniques applied

Within the corpus, grouped SHAP variants are applied on their own in 215 studies (72%), SHAP and LIME are applied together in 64 studies (22%), and grouped LIME variants are applied on their own in only 18 studies (6%). Once the SHAP and LIME papers are counted, SHAP features in some form in 279 of the 297 included studies (93.9%), making it the most common method within this SHAP/LIME-focused corpus, with LIME most often used as a complement rather than just a standalone method. Figures 7 and 8 show the overall distribution and its cross-domain pattern.

The drivers of this distribution (why SHAP dominates, why standalone LIME has receded, and why co-application of both methods should not be read as validation) are analysed alongside the technique×task and technique×domain crosswalks

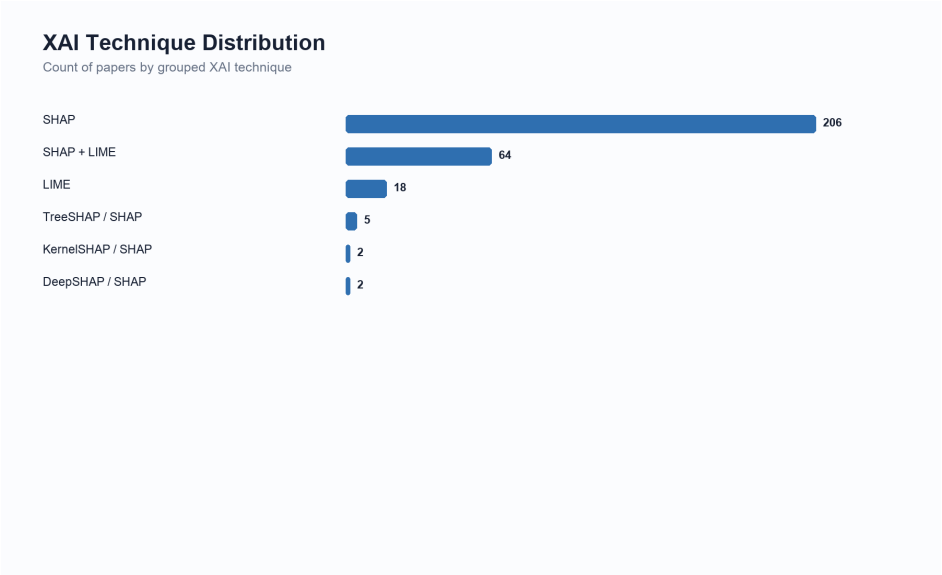



**Fig. 7** Grouped distribution of XAI techniques in the corpus. The "SHAP" bar shows studies coded to the generic SHAP label; the separately plotted TreeSHAP, KernelSHAP, and DeepSHAP bars are also SHAP variants, so *grouped* SHAP totals 215 studies, and SHAP features in 279 studies once SHAP+LIME papers are included. "SHAP+LIME" indicates papers applying both methods.

in Section 5.1, and the distinction between applying two methods and evaluating either is developed under RQ4 in Section 5.4. This section therefore reports the technique distribution descriptively and defers its interpretation to the research-question synthesis.

## 4.5 Datasets used

Most studies rely on publicly available datasets (241 studies, 81%). Authors use proprietary data in 38 studies (13%), mixed public/proprietary data in 14 (5%), and synthetic data in 4 (1%). This heavy reliance on public benchmarks directly affects external validity. We examine this limitation in Section 7.

Public datasets increase transparency and reproducibility. Yet they restrict the types of available financial XAI evidence. Researchers favor public benchmark data because they are accessible, shareable, and easy to compare across models. This convenience explains their prevalence in empirical studies. Public data supports methodological experimentation. Researchers can test explanation methods without navigating strict confidentiality constraints.

However, public datasets fail to capture operational complexity. Real financial pipelines involve temporal drift and missingness. They feature delayed labels, class imbalance, and institution-specific feature engineering. Changing business rules and governance constraints are standard. Public benchmarks rarely represent these elements. Explanations appearing stable and useful on public data often behave entirely

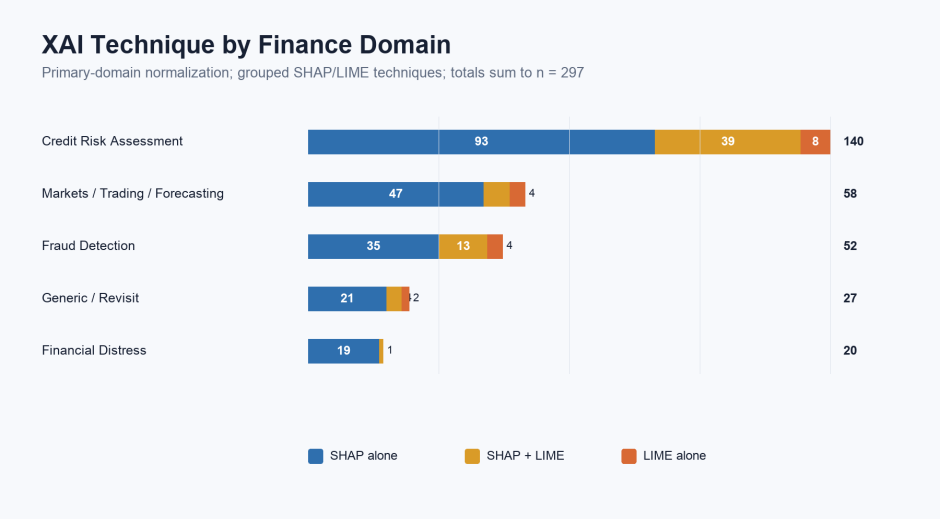



**Fig. 8** Distribution of grouped XAI techniques across financial domains in the retrieved corpus ($n = 297$). Domain counts reflect search-string-bounded sample distributions.

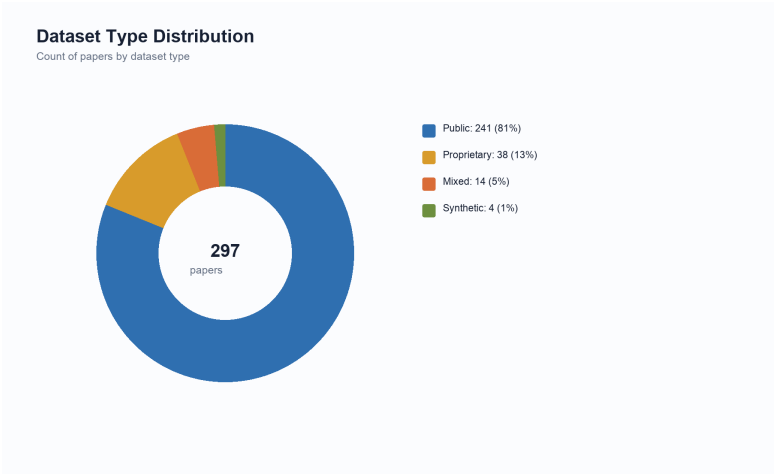



**Fig. 9** Dataset types used by included studies. Public datasets dominate the corpus, while proprietary and mixed-data studies remain comparatively rare.

differently when deployed in proprietary credit, fraud, trading, or risk-management systems.

We note the small share of proprietary and mixed-data studies. These studies matter because they approximate real deployment conditions. Unfortunately, proprietary datasets restrict reproducibility. Authors rarely disclose data access, preprocessing decisions, or feature definitions. This dataset distribution exposes a core trade-off. Public-data studies are reproducible. Proprietary-data studies are externally realistic but nearly impossible to audit.

## 4.6 Research-question coverage

Every study in the corpus contributes to RQ1 (297, 100%). We found 215 studies addressing RQ3 (72%) and 106 addressing RQ4 (36%). Only 27 studies (9.1%) provide evidence for RQ2. Fifty-nine studies address a single research question. We coded 141 addressing two, 84 addressing exactly three, and 13 addressing all four. These last two groups form the 97-study high-coverage subset. The low RQ2 coverage previews our central finding. Formal evaluation of explanation quality is extremely rare.

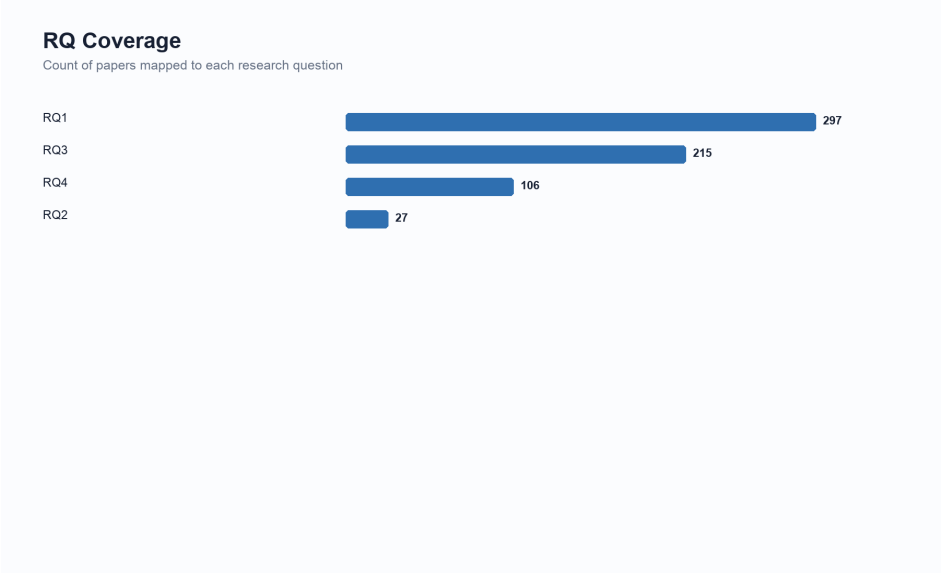



**Fig. 10** Research-question coverage across the extracted corpus. RQ1 is universal by design, while RQ2 coverage is rare.

Our inclusion criteria guarantee universal coverage of RQ1. Every included study applied SHAP, LIME, or both to a financial machine learning task. This rule makes the corpus excellent for mapping explanation techniques to task types, model families, and financial domains. Coverage of RQ2 is narrow because authors simply present explanation outputs. They rarely evaluate the quality of those explanations.

The higher coverage of RQ3 shows authors discuss deployment barriers even without evaluating explanation quality. Studies frequently mention computational cost, class imbalance, data quality, scalability, privacy, and domain-specific constraints. Authors do not always test these barriers empirically. They mention them in discussion sections, limitation statements, or qualitative interpretations. Thus, RQ3 evidence is broad but varies significantly in depth.

RQ4 sits in the middle. A substantial minority of studies report some validation practice. They use domain knowledge comparisons, robustness checks, counterfactual reasoning, causal interpretations, or human-centered assessments. Yet only 13 studies address all four review dimensions. Comprehensive reporting remains scarce. We use

the high-coverage subset of 97 studies for the sensitivity analysis in Section 5.5. We use the full corpus to estimate field-wide prevalence.

# 5 Findings by Research Question

This section presents our findings for the four research questions. We apply the two-level synthesis design from Section 3.4. We use the full corpus ($n = 297$) for all prevalence, mapping, and frequency claims. We use the high-coverage subset ($n = 97$, covering RQ1 plus at least two of RQ2–RQ4) exclusively as a sensitivity analysis on headline findings (Section 5.5). We base all percentages on the 297 included studies unless stated otherwise. As Section 4.6 notes, these numbers describe the SHAP/LIME subpopulation, not the entirety of financial XAI.

## 5.1 RQ1: XAI techniques, task types, and reported performance

Our screening criteria fix the technique axis by demanding genuine SHAP or LIME applications. We find the informative variation in how these techniques distribute across tasks and domains. Grouped SHAP variants appear alone in 215 studies (72%). Researchers apply SHAP and LIME together in 64 studies (22%). Grouped LIME variants appear alone in only 18 studies (6%). We observe SHAP in 279 of the 297 studies (93.9%). This percentage reflects the internal split within our eligibility gate. It does not measure SHAP's prevalence across all possible financial XAI methods. TreeSHAP's efficient native implementation in gradient-boosted tree frameworks (e.g., XGBoost, LightGBM, CatBoost) enables this high concentration. These frameworks dominate tabular financial prediction. LIME functions mostly as a complement rather than a standalone tool.

We cross-tabulate technique against task type in Figure 11. Classification is heavily represented. Of the 219 classification studies, 155 apply SHAP alone, 54 apply both, and 10 apply LIME alone. This SHAP-first preference appears in every other task type. Time-series forecasting uses SHAP in 26 studies, both in 4, and LIME in 5. Regression uses them 16, 3, and 1 times respectively. Anomaly detection uses them 7, 0, and 1 times. Multi-task studies use them 11, 3, and 1 times. LIME never dominates. Financial XAI clusters in a single setting: the SHAP-based explanation of classifiers. LIME serves as a secondary or corroborating method. This fits the methodological distinction from Section 2.2. LIME relies on local perturbation sampling and reacts sensitively to neighbourhood choices. Authors frequently present SHAP as more stable and theoretically grounded. Credit-risk evidence from [20] confirms we should treat the two methods as empirically distinct, not interchangeable.

The domain view in Figure 8 shows the same pattern from a different angle. SHAP applied alone is the most common choice in every normalized financial domain: credit (93 SHAP alone, 39 both, 8 LIME), markets (47, 7, 4), fraud (35, 13, 4), financial distress (19, 1, 0), and broad/generic finance or revisit studies (21, 4, 2). These cross-walk counts sum to the full corpus ($n = 297$) and use the same normalized domain grouping as Section 4.2. The combined use of SHAP and LIME is proportionally most visible in credit and fraud, where decisions are most often contested or audited

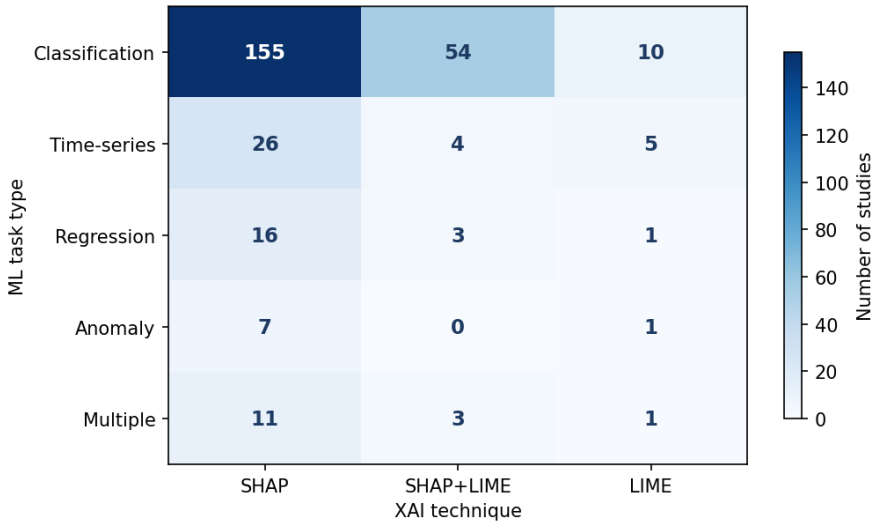



**Fig. 11** Cross-tabulation of XAI technique against machine-learning task type in the full corpus ($n = 297$). Darker shading indicates more studies; the concentration in the classification–SHAP cell shows the dominant setting, and the faint LIME column shows that LIME does not dominate any task.

and authors appear more inclined to report a second method as an informal cross-check. The uniformity across domains is itself informative: the field has converged on a shared toolset rather than developing domain-specific technique preferences. Individual studies nonetheless show variety within this landscape. In credit, SHAP is used to explain federated [39] and ensemble [40] classifiers, among many other credit-scoring and loan-default applications [41–44], while fraud studies apply both standard SHAP [45] and structured variants such as hierarchical group-based Shapley values [46], alongside a broad range of fraud- and anti-money-laundering detectors [47–50]. Beyond classification, Moustati and Gherabi [51] apply a temporal SHAP variant to an LSTM trading model, Wu et al. [52] embed SHAP and LIME together in a futures-forecasting pipeline to cross-check feature contributions, and further market and forecasting studies attach SHAP to attention-based [53] and keyword-guided [54] stock-prediction models, among many additional market-forecasting applications [55–58]. Financial-distress and bankruptcy prediction, though a smaller domain, is also represented [59–62].

Technique choice is closely tied to the model families in use. Every included study reports a model, but the families overlap because many studies use hybrid or multi-model pipelines (for example, gradient-boosted trees stacked with a neural component), so a single study can contribute to more than one family. With that caveat, tree-based ensembles and gradient boosting are the most common class (approximately 96 studies), followed by deep neural networks (approximately 70), while graph neural networks and purely linear or support-vector models are comparatively rare. The prevalence of tabular, tree-based modelling is a direct enabler of SHAP's dominance, because the efficient TreeSHAP algorithm makes high-fidelity

attribution inexpensive for exactly the model class the field most often uses; because model family is recorded as free text and pipelines are frequently hybrid, these counts are indicative rather than exact. However, recent financial-risk work also cautions that Kernel-SHAP explanations can change when dependence among input features is handled differently, so SHAP's practical appeal should not be read as automatic explanation validity [21].

The performance-reporting half of RQ1 shows the opposite pattern from explanation evaluation. All 297 included studies report some predictive-performance result, although the exact metrics differ across tasks, datasets, class balances, and validation splits. Table 10 reports non-exclusive metric mentions extracted from the performance field. Accuracy is the most common metric (142 studies, 47.8%), followed by AUC/ROC (91, 30.6%) and F1 (74, 24.9%); regression and forecasting metrics such as RMSE and MAE appear less often because classification dominates the corpus. These metrics are not pooled mainly because they do not estimate a common effect across differing datasets, class balances, and validation designs. They answer RQ1 by showing that the predictive performance is reported almost universally, whereas the explanation quality, as RQ2 shows, is usually not.

**Table 10** Predictive-performance metric reporting in the included corpus (non-exclusive mentions)

| Metric family | Studies | % of corpus |
|---|---|---|
| Accuracy | 142 | 47.8 |
| AUC/ROC | 91 | 30.6 |
| F1 score | 74 | 24.9 |
| Recall or sensitivity | 55 | 18.5 |
| Precision | 44 | 14.8 |
| RMSE | 29 | 9.8 |
| MAE | 17 | 5.7 |
| $R^2$ | 14 | 4.7 |
| MSE | 5 | 1.7 |

## 5.2 RQ2: Evaluation of explanation quality

We asked in our second research question whether studies actually evaluate their explanation quality. We found extremely low coverage. Only 27 of the 297 studies (9.1%; approximately 5.8–12.4%) report any evidence on explanation quality. Just 21 studies (7.1%; approximately 4.2–10.0%) conduct formal evaluations of their SHAP or LIME outputs. These formal evaluations include computed metrics like fidelity, faithfulness, stability, robustness, or structured human assessments. The remaining majority simply generate an explanation, interpret it, and assume it is correct. We already restricted the corpus to studies genuinely applying SHAP or LIME. This low rate is not an inclusion artefact. Even among researchers explicitly focused on generating explanations, formal evaluation remains rare. We confirmed this gap is robust through a coding-boundary sensitivity check. If we relax criteria to count basic

informal stability or feature-dropping checks as formal evaluation, the count moves from 21 to 28 studies (7.1% to 9.4%, 95% Wilson CI: 6.6%–13.3%). Over 90% of studies remain unevaluated under any reasonable boundary definition.

This evaluation gap extends beyond finance. Broader XAI evaluation research shows explanation quality is multi-dimensional. Researchers cannot establish it using illustrative examples alone [6]. The finance corpus mirrors this wider evaluation problem. However, finance is a high-stakes domain. Here, untested explanations carry much larger consequences.

Figure 12 maps explanation-quality reporting by domain. The shaded upper segment of each bar shows the fraction of studies reporting an evaluation. These shaded segments count the 27 studies reporting *any* explanation-quality evidence, not just the 21 with formal evaluations. The per-domain annotations sum to 27. We see a thin shaded fraction across every domain. Credit and lending has 10 of 140 (7%). Markets have 4 of 58 (7%). Fraud has 7 of 52 (13%). Financial distress has 0 of 20 (0%). Generic finance has 6 of 27 (22%). We report these figures to demonstrate widespread scarcity, not to rank domains. Cross-domain contrasts remain tentative because our search string omitted specific terms like fraud, AML, insurance, trading, or risk management. The retrieved fraud and insurance-like studies are likely non-random subsets of their subfields. We firmly conclude that explanation-quality evidence is rare across all domains, including credit. We do not claim one domain evaluates explanations more rigorously than others.

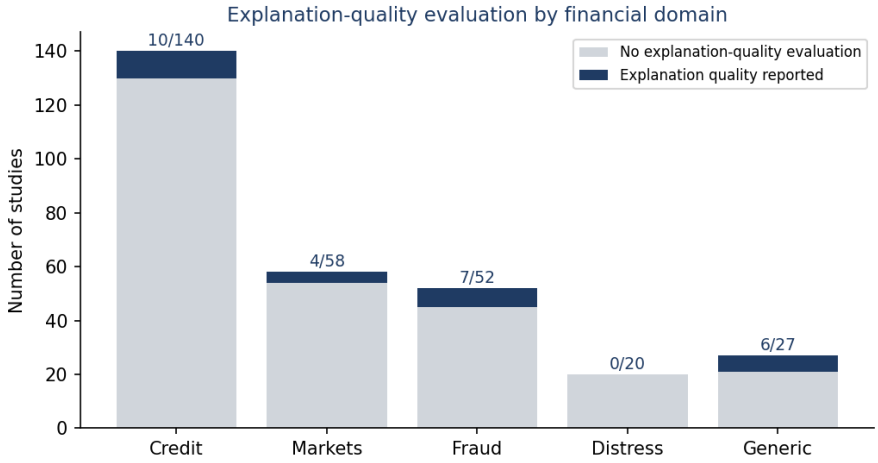



**Fig. 12** Explanation-quality evaluation by financial domain. For each domain the full bar is the number of included studies; the shaded upper segment (annotated evaluated/total) is the share reporting any evaluation of explanation quality.

We tested two obvious theories for this gap using cross-tabulations. First, proprietary-data studies do not evaluate more often than public-data studies. We found formal XAI evaluation in 3 of 38 proprietary-data studies (7.9%) and 17 of 241

public-data studies (7.1%) (Table 11). Second, we grouped venues exploratorily. Conference and proceedings papers report more formal evaluation than journal papers (19 of 196, 9.7%, versus 2 of 101, 2.0%). We interpret this venue split cautiously. Venue labels are heterogeneous. Some proceedings papers later appear in journal outlets. Regardless, neither cross-tabulation supports a reassuring narrative. Explanation evaluation is not routine, even in deployment-proximate proprietary studies or journal outlets.

**Table 11** Dataset type and venue type by explanation-quality evidence

| Grouping | Studies | Formal XAI eval. | % | Any quality signal |
|---|---|---|---|---|
| Public dataset | 241 | 17 | 7.1 | 23 |
| Proprietary dataset | 38 | 3 | 7.9 | 3 |
| Mixed public/proprietary dataset | 14 | 0 | 0.0 | 0 |
| Synthetic dataset | 4 | 1 | 25.0 | 1 |
| Conference/proceedings venue | 196 | 19 | 9.7 | 24 |
| Journal/other venue | 101 | 2 | 2.0 | 3 |

The nature of these evaluations matters as much as their rarity. We find functionally grounded metrics dominate the evaluated studies. Human-centred evaluation is almost non-existent. Authors most often report stability, calibration, fidelity, and faithfulness. We found only a handful of user studies or expert reviews. We distinguish *human-grounded* evaluation from *application-grounded* evaluation. In human-grounded studies, participants judge explanations in proxy tasks. In application-grounded studies, actual end users act on explanations during real deployed decisions. We see rare human-grounded studies. We found zero application-grounded evaluations with real end users making live financial decisions. Cost and convenience drive this ordering. Researchers can compute functional metrics automatically. Human-centred evaluation requires confidential data access and scarce domain experts. We highlight a few studies demonstrating stronger practice. Cristescu et al. [63] assess faithfulness using deletion curves. Aziz et al. [64] report fidelity, stability, and efficiency together. Kibriya et al. [65] incorporate a cognitive-load user study. Dikmen and Burns [66] execute a human-grounded trust and performance study in peer-to-peer lending. Tiukhova et al. [67] gather structured feedback from data-quality analysts. Some explicitly build evaluation into their design. Jo and Raj [68] target auditable rationales for credit scoring. Atef et al. [69] evaluate explanations of peer-to-peer default risk. Mazumder [70] constructs a reproducible SHAP framework with built-in fairness checks. Further studies similarly evaluate their explanation outputs [71, 72]. Our central quantitative finding is simply that these rigorous studies remain rare exceptions.

## 5.3 RQ3: Technical and data-related barriers to deployment

We use our third research question to identify obstacles blocking the path from published result to operational system. Out of the 297 studies, 215 (72.4%; approximately 67.3–77.5%) report at least one technical or data-related barrier. Authors discuss deployment barriers far more frequently than they evaluate explanation quality. We emphasize that these are *mention or discussion* counts. They are not counts of experimentally tested obstacles. The distribution of these mentions is highly skewed. Figure 13 shows class imbalance dominates. Authors name it in 117 studies. This represents over half the studies reporting any barrier. The corpus focuses heavily on credit-default and fraud problems, where the positive class is naturally rare. Some authors tackle this directly. They analyze how data balancing impacts classifier reliability [73]. They adopt imbalance-aware learning strategies [74–77]. Imbalance directly degrades explanation quality. Attributions estimated from few positive instances are consistently noisier and less stable.

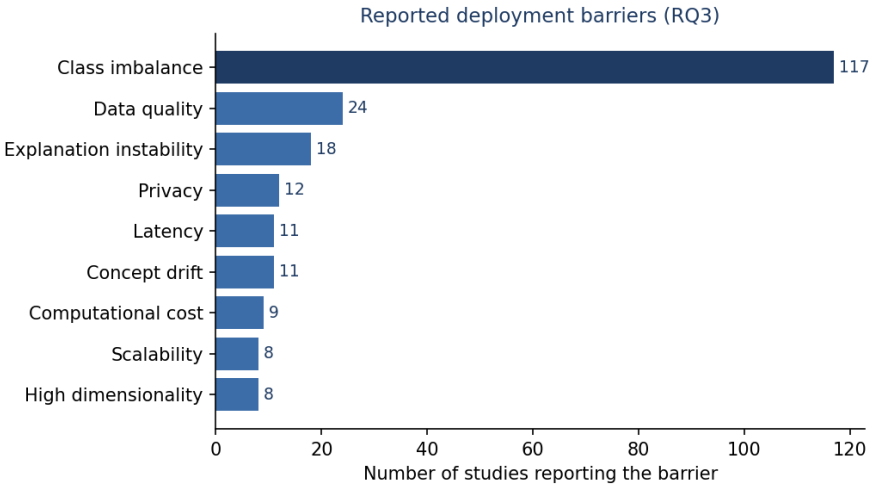



**Fig. 13** Deployment barriers mentioned or discussed across the corpus, ranked by number of reporting studies (studies may report more than one). Class imbalance dominates; the remaining barriers form a longer, thinner tail. Counts derive from free-text coding and should be read as ranked mention counts rather than precise estimates of empirically tested obstacles.

We find a long, thin tail of additional obstacles. Authors mention data quality (24 studies), explanation instability (18), privacy (12), latency (11), concept drift (11), computational cost (9), scalability (8), and high dimensionality (8). We open-coded this barrier field and did not compute formal inter-coder reliability statistics. Readers should interpret these tail counts as ranked bands of mention, not exact totals. The dominance of class imbalance is completely robust to recoding. We identify two important qualitative patterns. First, barriers track with domains. Fraud and anti-money-laundering face imbalance and concept drift because behaviour shifts over time. Real-time settings like transaction scoring struggle with latency, scalability, and

computational cost [78]. Domain shift pushes researchers toward adaptive ensemble approaches [79]. Second, authors rarely connect these barriers to the explanation-quality questions from RQ2. The interaction between class imbalance and explanation stability provides a perfect example. Imbalance is the top reported barrier. Attributions from rare positive instances are exactly the ones most prone to instability. Yet barely any studies test this interaction directly. Authors catalogue barriers but ignore their impact on explanation quality. This disconnect is a major finding of our review. It forms a natural bridge from RQ3 to the validation strategies in RQ4. Iqbal et al. [80] stand out as an exception. They handle imbalance by resampling before computing SHAP attributions. This emphasis on data-side barriers mirrors Kute et al. [29]. Their anti-money-laundering review concluded that lacking realistic transaction data is the main roadblock to deployable XAI.

We note this matches recent financial-services review evidence. Deployment barriers are organisational and regulatory, not just technical. XAI tools must integrate with model-risk workflows, audit trails, compliance review, and stakeholder-specific explanation needs [22, 23].

Broader credit-risk and banking studies outside our SHAP/LIME corpus highlight this exact tension. Institutions demand models that are accurate, transparent, auditable, and usable for credit assessment or credit-loss settings [4, 27, 28].

## 5.4 RQ4: Methodological approaches to validation

We ask in our fourth research question how studies build confidence in their explanations beyond relying on a single attribution method. We found 106 studies (35.7%; approximately 30.2–41.1%) reporting some validation-oriented practice. This is a weaker, more permissive threshold than the explanation-quality measurement from RQ2. One must not conflate the two. We did not compute formal inter-coder reliability statistics for validation-strategy coding. Readers should view these numeric RQ4 claims as consensus-adjudicated descriptive counts, not high-precision estimates. We grade each practice by the *relative assurance it provides about explanation faithfulness*. This means evaluating whether the explanation truly reflects the model's actual behaviour. We assign *low* assurance to practices merely showing that two explanations agree or that authors attempted validation. These include method agreement, hybrid/ensemble use, and weakly specified reporting. We assign *moderate* assurance to checks against unverified external references. These include domain-knowledge grounding and surrogate/rule comparisons. We grant *stronger* assurance to practices directly testing faithfulness or decision usefulness. These involve counterfactual and adversarial/robustness testing, and human-in-the-loop evaluation. The stronger label does not mean these designs guarantee SHAP/LIME faithfulness. It means they test a more relevant failure mode than simple method agreement. Figure 14 plots the 55 studies with a named strategy. Table 12 also includes the 51 weakly specified studies. The figure and table totals differ by design. We found weakly specified validation-oriented reporting is the most common class (51 studies). Here, authors discuss validation vaguely without naming a clear strategy. This residual category forms half of the 106 studies reporting any validation practice. This is a major finding. When authors

gesture toward validation, they rarely use a named, reproducible strategy. This perfectly matches the evaluation gap we identified under RQ2. Among named strategies, authors favour hybrid or ensemble XAI approaches (16 studies) and multi-method triangulation (13). They typically pair SHAP with LIME or other methods and treat their agreement as reassuring. A thin frontier of more rigorous methods follows. Causal or counterfactual analysis appears in 10 studies. Adversarial or robustness testing appears in 5. Human-in-the-loop validation has 5. Domain-knowledge grounding has 3. Surrogate or rule-based validation has 3.

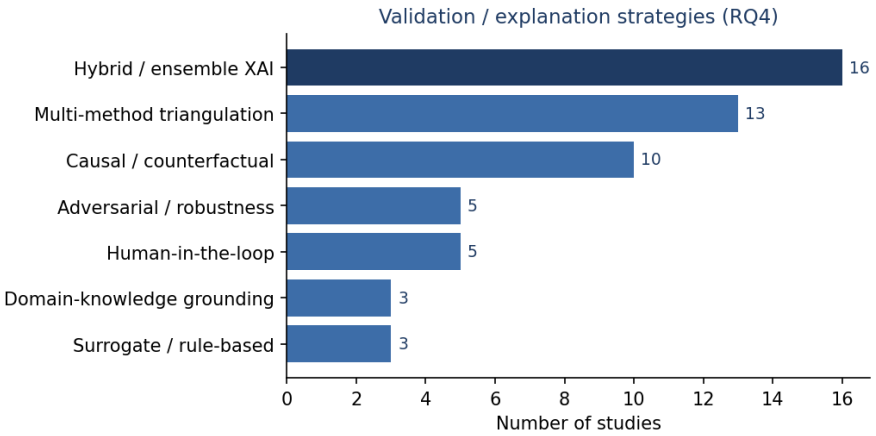



**Fig. 14** Named validation and explanation strategies reported under RQ4, ranked by number of studies. Low-assurance method-agreement strategies are most common among named strategies; stronger but still non-definitive causal, adversarial, and human-in-the-loop approaches remain rare. Counts derive from a free-text field and are indicative rather than exact.

The shape of this distribution reveals a lot. The most common strategies are the least demanding. Weak validation-oriented discussion and multi-method triangulation offer very limited assurance. Agreement between two attribution methods never proves that either is faithful to the model. We echo high-stakes interpretability arguments here. Researchers must not treat post-hoc explanations as substitutes for interpretable or independently validated decision models [5]. This distinction is critical for banking and fraud-detection. Operational explainability requires strong governance evidence. Institutions need auditability, repeatability, stakeholder usability, and monitoring under changing data conditions [22, 85]. The strategies that actually prove an explanation reflects the model and supports correct decisions remain marginal. We rarely see causal and counterfactual analysis, adversarial and robustness testing, or human-in-the-loop evaluation. The broader XAI literature has already established these stronger approaches. Their absence in finance indicates an adoption lag, not genuine unavailability. A few excellent studies prove they are usable. Vivek et al. [82] combine causal inference, counterfactual analysis, and SHAP for ATM-fraud detection. Roy et al. [83] propose a drift-aware framework uniting causal and counterfactual explanations for loan-default prediction. They directly tackle the concept-drift barrier from

**Table 12** Validation and explanation practices reported under RQ4 ($n = 106$ studies reporting a validation-oriented practice), showing the number of studies in each category, the relative assurance each practice can provide about explanation faithfulness, and a representative study where applicable. *Assurance* is graded by the rubric stated in Section 5.4 (low/moderate/stronger assurance about whether the explanation reflects the model). "Causal" here denotes causal-style analysis (e.g., counterfactual and drift-aware designs); consistent with Section 2.2, we do not treat SHAP/LIME attributions themselves as causal, and the "stronger" rating reflects the counterfactual and robustness testing these studies add rather than a causal interpretation of the attributions. Counts derive from a free-text field and are indicative rather than exact.

| **Validation strategy** | **Studies** | **Assurance** | **Representative study** |
|---|---|---|---|
| Weakly specified validation-oriented reporting | 51 | Low/unclear | — |
| Hybrid / ensemble XAI | 16 | Low | Kapoor et al. [81] |
| Multi-method triangulation | 13 | Low | Wu et al. [52] |
| Causal / counterfactual | 10 | Stronger | Vivek et al. [82]; Roy et al. [83] |
| Adversarial / robustness | 5 | Stronger | Straus et al. [84] |
| Human-in-the-loop | 5 | Stronger | Kibriya et al. [65] |
| Domain-knowledge grounding | 3 | Moderate | — |
| Surrogate / rule-based | 3 | Moderate | — |
| **Total studies reporting a validation-oriented practice** | **106** | | |

RQ3. Related work applies an explicitly causal perspective to fraud detection [86]. Other authors combine counterfactual explanations with a hybrid TabNet–LightGBM credit-risk model [81]. Additional studies pair SHAP or LIME with robustness testing, ensemble triangulation, or human-facing risk tiers [87–89].

## 5.5 Cross-cutting synthesis: from explanation to deployment readiness

We synthesize these four findings to describe the current state of financial XAI. The field generates explanations easily. It struggles to prove they are correct or deployable. These results are entirely interdependent. The field concentrates heavily on SHAP-based explanation of tree-ensemble classifiers on tabular data (RQ1). This focus explains both the dominant deployment barrier and the dominant validation strategy. Because researchers constantly apply SHAP to imbalanced tabular data, class imbalance surfaces as the top obstacle (RQ3). Because rigorous validation is expensive and complex, method agreement becomes the most common substitute (RQ4) for the missing explanation-quality measurement (RQ2). Ultimately, we find massive evidence of explanation production. We find very little evidence of explanation validation.

These weaknesses are interlocking, so improving any one of them in isolation would be insufficient. Cheaper or faster explanation would not make an unvalidated explanation trustworthy, and method-agreement checks do not resolve the class-imbalance and concept-drift problems that degrade explanation quality in the first place. The high-coverage subset ($n = 97$) serves as an upper-leaning sensitivity analysis on this picture. Its evaluation rate is higher than the full-corpus rate: 19 studies (19.6%) report a formal explanation evaluation and 20 (20.6%) report any explanation-quality signal. This increase shows that reporting completeness matters and that the 7.1% full-corpus estimate should not be treated as a context-free property of all strong papers. However, because the subset was constructed from multi-RQ coverage, it is inherently biased toward papers that report more about evaluation and validation. The conservative interpretation is consequently straightforward: even in this select, highly reported upper-bound third of the corpus, roughly four in five studies still do not formally evaluate their explanations. The deployment-readiness implication is therefore direct: the field has largely equated the use of an attribution method with being explainable, without the validation step that would allow an explanation to be relied upon in a regulated, high-stakes setting. Section 6 develops what this means for practitioners and regulators.

# 6 Discussion

The four research questions describe a literature that has advanced quickly in method adoption but more slowly in evidence discipline. We found a stark contrast in our review. Researchers have adopted explanation methods rapidly, yet they remain slow to build a strong base of evidence. SHAP and LIME appear routinely in financial machine learning studies. They are especially common in classification-heavy settings like credit, fraud, and market risk. Yet production outpaces validation. Researchers frequently attach post-hoc attribution methods to financial models. We found much weaker evidence showing these explanations are actually faithful, stable, or useful to decision makers. They are rarely ready for regulated deployment. We stress a central distinction here. The mere presence of an XAI method does not guarantee trustworthy explainability. The field must stop treating them as equivalent.

## 6.1 Relationship to broader finance-XAI reviews

Other review work suggests financial XAI is shifting. The field is moving past method selection to face implementation and governance questions. Banking-focused reviews point to socio-technical adoption barriers, auditability, and organisational readiness [22]. They also note that different stakeholders need different types of explanations. Broader financial-services reviews name regulatory uncertainty, standardisation gaps, and responsible-AI governance as stubborn challenges [23]. In fraud detection, bibliometric studies map rapid growth in AI research. Yet they also reveal poor integration between technical models, governance, and human-centred evaluation [85]. We build on this prior work. We quantify how often empirical finance papers using SHAP and LIME actually evaluate their explanations. We track how frequently they report deployment barriers and use validation strategies. Our contribution is narrower but

**Table 13** Compact summary of findings by research question

| Research question | Main empirical answer | Key evidence | Reviewer-facing implication |
|---|---|---|---|
| RQ1: Technique–task mapping | Within the SHAP/LIME-gated corpus, SHAP is much more common than standalone LIME, especially for classification and tabular modelling. Predictive performance is reported almost universally but is too heterogeneous to pool. | SHAP appears in 279 of 297 studies; 215 studies use SHAP alone, 64 use SHAP with LIME, and 18 use LIME alone; performance metrics (accuracy, AUC, F1) are reported in all studies. | The corpus supports a strong internal mapping claim, but not a prevalence claim about all financial XAI methods. |
| RQ2: Explanation-quality evaluation | Explicit evaluation of explanation quality is rare; most studies generate and interpret explanations without formally testing them. | Only 27 studies report any explanation-quality signal, and 21 studies report formal XAI evaluation. | The main evidence gap is not method adoption but explanation validation. |
| RQ3: Deployment barriers | Technical and data-related barriers are commonly reported, but often discussed rather than empirically tested. | 215 studies report at least one barrier; class imbalance is the dominant barrier, appearing in 117 studies. | Deployment readiness should not be inferred from reported predictive performance or explanation plots alone. |
| RQ4: Validation strategies | Broader validation practices appear in a minority of studies, with weakly specified validation discussion and low-assurance method agreement more common than stronger causal, adversarial, or human-centred designs. | 106 studies report some validation-oriented practice; Table 12 reconciles these studies across named and weakly specified categories. | The field needs stronger validation designs before explanations can support regulated financial decision-making. |

focuses strictly on auditing the evidence. We separate the basic adoption of an XAI method from actual proof of its quality.

## 6.2 From method adoption to explanation evidence

Our RQ1 findings map a narrow technical convergence. SHAP-centred attribution dominates the literature. Researchers typically use LIME only as a secondary or corroborating technique. We understand why this happens. Financial prediction problems frequently rely on tabular data and tree-based or ensemble models. SHAP is highly convenient here. Libraries like XGBoost and LightGBM implement it widely via TreeSHAP. It also produces both local and global feature-attribution summaries. Credit scoring and fraud detection naturally demand case-level explanations. Analysts and governance teams must know exactly why a specific applicant, transaction, account, or firm received a particular risk score. We must be precise about what this dominance means. It reflects tool availability and the popularity of tree models within our

specific SHAP/LIME-gated corpus. It does not describe every possible XAI approach across all of finance.

Such convergence brings clear benefits. It also restricts the field's imagination. When most studies rely on the same family of attribution tools, researchers start confusing standard practice with sufficient practice. SHAP summary plots, local force plots, and LIME outputs can make a model look inspectable. They do not automatically prove the explanation is faithful to the model. They do not guarantee stability under plausible perturbations or meaning for a financial expert. They might be completely inappropriate for the actual decision context. We are mostly seeing unevaluated explainability attached to well-evaluated predictive models. This is not validated explainability.

We must address this asymmetry. Predictive performance and explanation quality answer entirely different questions. A model might achieve excellent AUC, accuracy, F1-score, or error-rate performance. That same model can easily produce unstable, misleading, or overly abstract explanations. The reverse is also true. A genuinely useful explanation can expose serious flaws in a high-performing model. It might reveal dependence on proxy variables, fragile correlations, or feature effects that drift over time. Financial XAI researchers need to treat explanation outputs as first-class evidence. They are not merely visual supplements to predictive results.

## 6.3 Why the evaluation gap matters in financial settings

We consider the low rate of explanation-quality evaluation our most important practical finding. Explanations in finance drive decisions about credit access, fraud investigation, and insurance pricing. They influence trading risk, audit review, regulatory reporting, and customer communication. None of these are low-stakes environments. An explanation becomes part of the decision system the moment it is shown to a declined borrower or used by a fraud analyst to prioritize an investigation. It enters the system when cited in model-risk documentation. An unvalidated explanation generates operational, fairness, and compliance risks. This remains true even if the underlying predictive model performs perfectly.

We see two major risks. Lack of faithfulness poses the most immediate danger. An explanation might attribute a decision to variables that did not actually drive the model output. This misleads analysts, customers, and auditors. In credit and lending, inaccurate explanations distort adverse-decision reasoning. In fraud detection, they push analysts toward the wrong behavioural signals. For risk management, they severely weaken model challenge. Reviewers end up inspecting an inaccurate account of model behaviour. Lack of stability creates a second serious threat. Small changes in data preprocessing, sampling, background distribution, or local perturbation design can produce materially different explanations. These fragile explanations become nearly impossible to audit or reproduce. Financial institutions rely on repeatability and documentation for model governance. Unstable explanations break that governance.

Better practice is entirely possible. The minority of studies that actually evaluate explanation quality prove this. We found authors reporting faithfulness through deletion-style tests. Some combine fidelity, stability, and efficiency metrics. A few go

further and incorporate user studies, expert feedback, or task-performance measures [63–67]. The tools exist. We simply lack a routine reporting norm in financial XAI. Many functional tests cost very little compared to full human-subject evaluations. Applied studies have no real excuse to skip them. Authors must report at least one explanation-quality test that matches how they intend the explanation to be used.

## 6.4 Deployment readiness and the barrier evidence

Our RQ3 findings reveal a split focus. The literature pays far more attention to deployment barriers than to explanation evaluation. Class imbalance tops the list of barriers. Data quality, explanation instability, privacy, latency, concept drift, computational cost, scalability, and high dimensionality follow behind. This mix makes sense. Our corpus leans heavily into credit-default and fraud-detection settings. These areas constantly deal with rare events, delayed labels, shifting behaviour, and noisy operational data.

We noticed a problem with how authors report these barriers. They tend to describe them rather than diagnose them. A paper might name class imbalance, drift, privacy, or scalability as a limitation. Very few actually test how these factors degrade explanation quality. The connection between barriers and reliability is critical. Class imbalance frequently makes local attributions much noisier for rare default or fraud cases. Concept drift causes feature-importance patterns to decay. Data-quality problems warp explanations. They end up reflecting missingness, leakage, or institutional quirks instead of real financial signals. Latency and computational cost force engineering teams to compromise. They simplify explanation pipelines, cache outputs, or rely on approximations. Those approximations then require their own fidelity checks.

We draw a direct practical lesson here. Nobody should infer deployment readiness just because a paper shows good model performance and a SHAP plot. True readiness requires connecting the explanation method directly to the operational constraints of the problem. If a paper claims to explain fraud detection, we need to see it evaluated under severe class imbalance and shifting transaction behaviour. Credit explanations demand tests for stability, proxy behaviour, and consistency across similar applicants. Market-forecasting explanations must survive temporal non-stationarity and regime change. If an explanation is meant for audit or model risk, it has to be reproducible and traceable. A reviewer who never touched the original code must be able to challenge it.

## 6.5 Weak triangulation and stronger validation strategies

Broader validation strategies exist, as our RQ4 results show, but their use remains spotty. The most popular forms offer surprisingly weak assurance. Researchers frequently use multi-method triangulation or hybrid XAI approaches. They will compare SHAP outputs with LIME results, see that they agree, and treat that agreement as a success. This works fine as a basic sanity check. It completely fails to prove that either method is actually faithful to the underlying model. Two post-hoc methods often agree simply because they share the same blind spots. They might rely on similar assumptions, track the exact same correlated features, or take the same shortcuts when simplifying complex model behaviour.

We found stronger validation strategies in the corpus, though they appear rarely. Causal and counterfactual analyses test whether explanations map to plausible interventions instead of just reporting correlational feature importance [81–83]. Adversarial and robustness tests check explanation stability when inputs, sampling, or attack conditions shift [84]. Human-in-the-loop evaluation asks hard questions. Does the explanation actually improve expert reasoning? Does it calibrate trust correctly? Does it lead to better decisions, or just look plausible to the engineers who built it [65–67]? We know these approaches cost more than simply checking if two methods agree. They are necessary because they generate evidence that matches how explanations work in real financial practice. We do not treat them as perfect guarantees. Counterfactual, adversarial, and user-study evidence simply reduces specific risks. They cannot eliminate every possible failure of faithfulness or stability.

We argue for a much clearer validation hierarchy. The evaluation taxonomy proposed by Doshi-Velez and Kim [17] offers an excellent framework. It connects directly to the human-grounded and application-grounded distinctions we used in Section 5.2. The baseline is mere transparent generation. A study reports its method, implementation, background distribution, model type, and scope. One step up is *functionally grounded* evaluation. These are automated tests for faithfulness, stability, robustness, or fidelity that run without human participants. The next tier demands checks that probe reasoning. These include domain-knowledge grounding, counterfactual tests, causal analysis, and adversarial evaluation. *Human-grounded* evaluations sit higher still. Here, participants judge or act on explanations in proxy tasks. Finally, *application-grounded* evaluations sit at the apex. Intended users interact with the explanations during live, deployed decisions. Our corpus bunches almost entirely at the bottom. The numbers thin out rapidly as we move up. We found zero studies conducting a true application-grounded evaluation. We recognize that not every paper requires a live deployment test. However, researchers must stop making high-stakes claims based purely on baseline generation. Figure 15 maps this hierarchy and plots where our corpus currently sits.

## 6.6 Implications for researchers, practitioners, and regulators

Researchers need to stop cataloguing methods and start reporting evidence. A rigorous study specifies exactly what the explanation is supposed to do. It names the intended user. It defines which quality metric is being tested and declares what a failure would look like. We must separate predictive performance from explanation evaluation. You cannot build a strong classifier, print a SHAP plot, and claim you have solved explainability. Every paper should attempt at least one functionally grounded explanation test. Authors must then honestly report where that test falls short.

Industry teams need to abandon "SHAP-and-stop" workflows immediately. We acknowledge SHAP works well for tree-based financial models. TreeSHAP provides real value for tabular credit, fraud, and risk problems. Yet choosing SHAP does not excuse you from testing it. Teams must verify explanation stability. They need to check sensitivity to background data and probe feature-dependence assumptions. They must prove the explanation actually helps the final stakeholder. The same rules apply to LIME, where sloppy local perturbation choices easily warp the output. We argue that

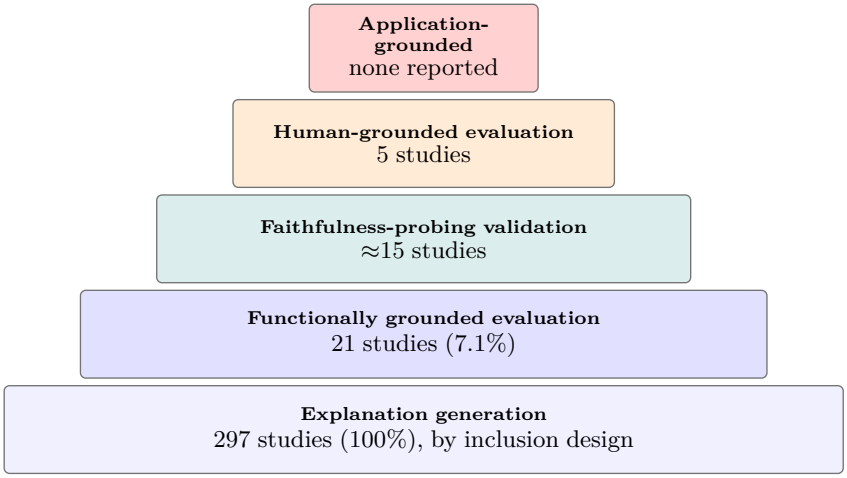



**Fig. 15** Evaluation and validation hierarchy for the corpus, mapped onto the functionally-grounded, human-grounded, and application-grounded taxonomy of Doshi-Velez and Kim [17]. Each tier is annotated with the number of included studies that report evidence at that tier: all 297 generate an explanation by inclusion design, but only 21 (7.1%) report a formal functionally-grounded evaluation (Section 5.2); faithfulness-probing validation (counterfactual, causal, and adversarial/robustness testing) and human-in-the-loop evaluation are reported by roughly 15 and 5 studies respectively (Table 12), and no study reports a full application-grounded evaluation with real users in a live financial decision. Counts derive from free-text coding and are indicative rather than exact, and the tiers are not strictly nested subsets.

operational systems must treat explanation pipelines exactly like model pipelines. Version them. Monitor them. Run them through formal model-risk review. They are not optional visual accessories.

Regulators and internal risk teams should update their documentation demands. They must ask for explanation evidence, not just explanation presence. The core question is not whether a team used an XAI method. The questions are whether they evaluated it, for whom, under what data constraints, and against what failure criteria. Existing reviews correctly push for auditability, organizational readiness, and stakeholder-specific governance [22, 23]. Our audit adds a blunt reality check. The current crop of empirical SHAP and LIME studies cannot meet those governance expectations. They validate their models aggressively. They leave their explanations completely untested.

## 6.7 Toward minimum reporting standards for financial XAI

We propose a straightforward minimum reporting standard for future financial XAI work. We are being explicitly normative here. Table 14 presents a recommendation built to fix the gaps we found, not a taxonomy we passively observed. Every study needs to define its financial decision context and name the intended user. Authors must specify the model family, data type, and any deployment-relevant data conditions. They must detail the exact XAI method and its implementation. They should state whether the explanation is local or global. Crucially, they must report at least one specific explanation-quality test. They should document observed deployment

barriers. Finally, they must disclose if their validation strategy goes beyond simple single-method attribution.

**Table 14** Minimum reporting standard recommended for financial SHAP/LIME studies

| Reporting item | Minimum information to provide |
|---|---|
| Decision context | Financial domain, task, decision setting, and whether the explanation is for research, internal validation, analyst support, customer communication, or regulatory review. |
| Model and data setting | Model family, dataset type, class balance, temporal structure, preprocessing, and any deployment-relevant data constraints. |
| Explanation method | SHAP/LIME variant, local/global scope, background or perturbation design, software implementation, and explanation output type. |
| Explanation evaluation | At least one explicit test or assessment, such as faithfulness, fidelity, stability, robustness, expert review, user study, or decision-support evaluation. |
| Deployment barriers | Reported barriers such as imbalance, drift, privacy, latency, scalability, data quality, computational cost, and governance fit. |
| Validation strategy | Whether evidence relies on single-method interpretation, method agreement, robustness testing, counterfactual or causal analysis, domain-knowledge checks, or human-in-the-loop validation. |

A minimum standard will not fix every flaw in the field. It will, however, force the evidence base to become cumulative. It forces weaknesses out into the open. A team could absolutely still publish an exploratory SHAP or LIME analysis. The standard just requires them to explicitly label their explanations as unevaluated if they skipped quality testing. Fast-growing fields desperately need this transparency. We drew one inescapable conclusion from Section 5. Financial XAI excels at generating explanations but mostly fails at validating them.

We consider our core findings secure, even if their scope is narrow. SHAP and LIME adoption has surged. Researchers report predictive performance obsessively. They almost never evaluate explanation quality. They list deployment barriers without actually diagnosing them. We treat other findings with more caution. The cross-domain differences, the precise ranking of minor barriers, and the exact distribution of RQ4 validation strategies are all tentative. These rely on specific search-string coverage and smaller denominators. Open coding always introduces some variance.

## 6.8 Future research directions

Our audit demands a specific research agenda. We are not just making a generic call for more studies. We see four urgent priorities.

Explanation evaluation has to become the default behavior. Functionally grounded tests for fidelity, faithfulness, and stability cost very little compared to running human-subject studies. Yet we found over 90% of the corpus skipped evaluation entirely (Section 5.2). We blame bad practice, not missing methods. Publishing an attribution without a quality test should be a rare exception. Today it is the rule.

Researchers must push past the current narrow focus. The literature heavily favors SHAP explanations of tree ensembles on tabular credit data (Sections 4.2–4.4). We need work on tasks that actively resist simple explanation. Time-series forecasting, anomaly detection, and reinforcement-learning agents all demand attention. We also need more focus on insurance, anti-money-laundering, and financial-distress prediction. These areas desperately need strong validation evidence. We acknowledge our search strings might have missed some domain-specific work (Section 3.2.2). Even so, the gap is glaring.

We need joint studies of deployment barriers and explanation quality. Consider class imbalance. It was the most frequently reported barrier (Section 5.3). It also aggressively destabilizes local attributions for rare fraud or default cases. Almost nobody actually tests that interaction. Finance runs on operational stressors. We deal with imbalance, concept drift, missing data, and tight latency requirements. We need researchers to evaluate their explanations under exactly these hostile conditions.

Finally, we need serious, comparable validation methods. Causal analysis, counterfactuals, adversarial tests, and human-in-the-loop evaluations barely register in our corpus (Section 5.4). These methods thrive in the broader XAI literature. Their absence here indicates cultural lag, not technical impossibility. We need shared benchmarks. We need standard reporting conventions for faithfulness, stability, and human utility. We should be comparing models on how well they explain themselves, not just on their raw predictive scores. We now turn to the limitations of our own evidence base in Section 7.

# 7 Threats to Validity

We built our conclusions from a messy, heterogeneous body of applied studies. We did not run a controlled experiment. This exposes our findings to specific validity threats. We follow standard systematic-review guidance to address them [37]. We group these threats into construct, internal, external, and conclusion validity categories. We detail our mitigations for each.

## 7.1 Construct validity

Construct validity asks if we actually measured what we intended to measure. Two design choices matter heavily here. First, our search strings defined "financial XAI" using a fixed vocabulary (Appendix A). We deployed broad terms like finance, banking, credit risk, portfolio optimisation, financial distress, and financial market. We deliberately did not try to list every niche term. We skipped specific phrases like fraud, money laundering, AML, anomaly detection, insurance, accounting, audit, trading, or risk management. We caught studies in those areas if they matched the broad terms. We inevitably missed papers indexed purely under the narrow labels. We treat

our domain counts as patterns within the retrieved SHAP/LIME corpus. They are not definitive market-share estimates for all financial XAI. Our fraud and insurance counts represent a strict lower bound. Second, we used SHAP and LIME as a strict inclusion gate. This choice made the corpus internally comparable. It also means we are measuring a very specific subpopulation. Papers using intrinsically interpretable models, attention mechanisms, or counterfactuals without SHAP or LIME fall outside our scope. When we report prevalence rates, like the rate of explanation evaluation, we are talking exclusively about this SHAP/LIME group (Sections 4 and 5). We are not estimating behavior across all financial explainability methods. Finally, we created the coding categories for evaluation, barriers, and validation ourselves. We piloted the codebook and defined fields tightly to control this risk (Section 3.5). We strictly separated *formal* evaluation from weaker signals. We refused to inflate the headline evaluation rate with permissive coding.

## 7.2 Internal validity

Internal validity focuses on extraction accuracy. Did we faithfully pull data from the primary studies? Screening and extraction always require judgement. We faced tough calls on borderline SHAP/LIME use, asserted versus tested barriers, and vague validation strategies. We applied three major mitigations. Two reviewers independently screened all titles and abstracts. We then assessed full texts against a strict written guide, using a third reviewer to break ties (Section 3.2.4). We double-coded a stratified random sample of 50 studies ($\approx$ 17% of the corpus) to prove quantitative agreement. The inter-coder reliability scores were consistently high ($P_o \geq 92.0\%$, Cohen's $\kappa \geq 0.78$, Gwet's $\text{AC}_1 \geq 0.86$). The formal evaluation tier performed exceptionally well ($\kappa = 0.83$, $\text{AC}_1 = 0.95$; Section 3.5). We sharpened definitions mid-stream when validation strategies proved too ambiguous. We forced consensus coding for that specific field. We also ran a dedicated verification pass across technique, task, domain, evaluation, barrier, and research-question data. We hunted for three exact failure modes. We checked for false-positive SHAP/LIME inclusions. We checked for predictive-performance stats mislabeled as explanation evaluation. We checked for purely theoretical barriers coded as empirical tests. We know some residual extraction error remains. The RQ4 codes work perfectly well for qualitative synthesis. We just treat their exact numerical distribution as slightly less rock-solid than the technique or domain counts.

## 7.3 External validity

External validity limits how far we can generalise. Four specific factors hem us in. We only searched for English-language, peer-reviewed journal and conference papers in Scopus, Web of Science, and IEEE Xplore. We ignored grey literature and non-English work. We missed papers in unindexed venues or those using vocabulary our strings missed. We also failed to retrieve thirteen eligible reports. The corpus skews extremely recent. Roughly two-thirds of the papers appeared in 2025 or 2026 (Section 4.1). These studies are simply too new to have generated replication data or sustained operational history. Any conclusions we draw about deployment readiness require inference. The

dataset problem creates our biggest constraint. Four out of five studies rely entirely on public benchmark datasets (Section 4.5). Public data makes research reproducible. It also completely hides the drift, missingness, delayed labels, and messy feature engineering that define live financial systems. An explanation that looks perfectly stable on public data might shatter in production. Our findings hit hardest for high-volume domains like credit, markets, and fraud. They remain tentative for under-represented areas and real operational deployments.

### 7.4 Conclusion validity

Conclusion validity dictates whether our inferences actually hold up. The primary studies use wildly different datasets, class balances, target definitions, and metrics. A quantitative meta-analysis was impossible. We built a descriptive, narrative synthesis instead (Section 3.7). We refused to pool predictive-performance figures. We do not report effect sizes or certainty-of-evidence estimates. We do report approximate intervals for key binomial proportions to maintain transparency. We found 21 formal evaluations out of 297 studies, yielding 7.1% (approximately 4.2–10.0%). We found 27 weak quality signals, or 9.1% (approximately 5.8–12.4%). We logged 215 barrier reports, hitting 72.4% (approximately 67.3–77.5%). We tracked 106 validation practices, reaching 35.7% (approximately 30.2–41.1%). We built the barrier and validation tallies from open-coded free text. Readers should treat these as ranked mention counts, not absolute totals. Still, the massive patterns hold up perfectly well to recoding. Class imbalance dominates the barriers. Formal evaluation barely exists. We kept our protocol internal rather than prospectively registering it on PROSPERO or OSF (Section 3.1.1). We changed zero eligibility criteria after screening started, but we acknowledge the transparency hit. We included the full extraction workbook as supplementary material. Anyone can audit our decisions, coding, and verification checks.

## 8 Conclusion

We designed this review to audit the empirical literature on SHAP and LIME in finance, not just catalogue it. We ran a strict protocol across Scopus, Web of Science, and IEEE Xplore. We screened 3,010 records. We extracted data from the 297 studies that actually apply these methods to financial machine learning. We built our synthesis around four core questions. We mapped techniques to tasks, tracked explanation evaluation, logged deployment barriers, and analysed validation strategies.

Our evidence points to a massive gap between production and proof. The field has settled into a comfortable groove. Researchers overwhelmingly use SHAP to explain tree-ensemble classifiers running on tabular credit, fraud, and market data. They produce explanations easily. They almost never prove those explanations are trustworthy. Quality evidence remains scarce even when we look only at the best-reported papers. The evaluation rate bumps up from 7.1% in the full corpus to 19.6% in the high-coverage sensitivity sample. Four out of five of those top-tier papers still fail to formally evaluate their explanations. Authors talk about deployment barriers constantly. They rarely test them. They almost never measure how those barriers wreck

explanation reliability. We found broader validation strategies, but they offer mostly illusion. Teams rely heavily on weak method-agreement checks. They ignore the causal, adversarial, and human-in-the-loop designs that could actually prove faithfulness.

These failures feed into each other. The obsession with applying SHAP to imbalanced tabular data drives the deployment barriers we see. It explains why teams lean so heavily on cheap method-agreement checks instead of real evaluation. We cannot fix these problems one at a time. Speeding up an explanation pipeline does not make an unvalidated explanation trustworthy. Checking if SHAP and LIME agree does nothing to solve the data imbalance and drift that quietly ruin the output.

We set out to drive a wedge between adopting a method and proving it works. Financial XAI currently treats them as identical. We must break that habit. The stakes demand it. An unvalidated explanation becomes dangerous the moment a bank shows it to a declined borrower. It corrupts the process when a fraud analyst trusts it or a model-risk reviewer signs off on it. We proposed a minimum reporting standard in Section 6.7 (Table 14) to fix this. It costs nothing to implement. Authors simply need to state the decision context and name the intended user. They must run at least one quality test. They must disclose if they actually validated the output beyond just printing the attribution. This forces the evidence base to become cumulative. It forces unvalidated explanations to label themselves.

Financial explainability has won the adoption war. It has completely failed to demonstrate its own reliability. We have to change how we build these systems. The next phase of research must treat the explanation itself as a first-class object. Researchers must evaluate it rigorously. They must validate it under brutal, realistic operational conditions. They must report their results to a common standard. Only then can we safely use these tools in regulated, high-stakes financial decisions.

## Declarations


- **Funding**: The authors declare that no funding was received during this work.
- **Competing interests**: The authors have no relevant financial or non-financial interests to disclose.
- **Ethics approval**: Not applicable. This review does not involve human participants, their data, or animals.
- **Use of AI Tools**: Generative AI tools were used to assist with language editing and reference organisation. The authors directed the work, verified the technical content and citations, and take full responsibility for the manuscript. No AI system is listed as an author.
- **Data Availability Statement**: The data supporting the findings of this study are available within the article and its supplementary materials. The full list of 297 included studies is provided in Supplementary File [Extraction Sheet.xlsx]. All other extracted data are available from the corresponding author upon reasonable request.
- **CRediT author statement**:
  Manpreet Singh: Conceptualization, Formal analysis, Investigation, Methodology, Writing – original draft

Akshatha Srikantha: Conceptualization, Formal analysis, Investigation, Methodology, Writing – original draft
Keshav Kaushik: Data curation, Validation, Visualization, Writing – review and editing
Prajwal Bajpai: Formal analysis, Visualization, Writing – review and editing

# References


[1] Makridakis, S. (2017). The forthcoming Artificial Intelligence (AI) revolution: Its impact on society and firms. Futures, 90, 46–60. https://doi.org/10.1016/j.futures.2017.03.006

[2] Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2019). 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

[3] 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

[4] Bücker, M., Szepannek, G., Gosiewska, A., & Biecek, P. (2022). Transparency, auditability, and explainability of machine learning models in credit scoring. Journal of the Operational Research Society, 73(1), 70–90. https://doi.org/10.1080/01605682.2021.1922098

[5] 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

[6] Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., Schlötterer, J., van Keulen, M., & Seifert, C. (2023). From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI. ACM Computing Surveys, 55(13s), Article 295. https://doi.org/10.1145/3583558

[7] Černevičienė, J., & Kabašinskas, A. (2024). Explainable artificial intelligence (XAI) in finance: a systematic literature review. Artificial Intelligence Review, 57(8). https://doi.org/10.1007/s10462-024-10854-8

[8] Weber, P., Carl, K. V., & Hinz, O. (2023). Applications of Explainable Artificial Intelligence in Finance—a systematic review of Finance, Information Systems, and Computer Science literature. Management Review Quarterly, 74(2), 867–907. https://doi.org/10.1007/s11301-023-00320-0

[9] Yeo, W. J., Van Der Heever, W., Mao, R., Cambria, E., Satapathy, R., & Mengaldo, G. (2025). A comprehensive review on financial explainable AI. Artificial Intelligence Review, 58(6). https://doi.org/10.1007/s10462-024-11077-7

[10] Khan, F., Mazhar, S., Mazhar, K., AlSaleh, D. A., & Mazhar, A. (2025). Model-agnostic explainable artificial intelligence methods in finance: A systematic

review, recent developments, limitations, challenges and future directions. Artificial Intelligence Review, 58, 232. https://doi.org/10.1007/s10462-025-11215-9

[11] Martins, T., de Almeida, A. M., Cardoso, E., & Nunes, L. (2024). Explainable artificial intelligence (XAI): A systematic literature review on taxonomies and applications in finance. IEEE Access, 12, 618–629. https://doi.org/10.1109/access.2023.3347028

[12] Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: A survey on Explainable Artificial Intelligence (XAI). IEEE Access, 6, 52138–52160. https://doi.org/10.1109/access.2018.2870052

[13] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. ACM Computing Surveys, 51(5), 1–42. https://doi.org/10.1145/3236009

[14] 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 (KDD '16) (pp. 1135–1144). Association for Computing Machinery. https://doi.org/10.1145/2939672.2939778

[15] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems 30 (NIPS 2017) (pp. 4765–4774). Curran Associates. https://doi.org/10.48550/arXiv.1705.07874

[16] Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence, 2(1), 56–67. https://doi.org/10.1038/s42256-019-0138-9

[17] Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608. https://doi.org/10.48550/arXiv.1702.08608

[18] Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N., & Herrera, F. (2023). Explainable Artificial Intelligence (XAI): What we know and what is left to attain trustworthy artificial intelligence. Information Fusion, 99, 101805. https://doi.org/10.1016/j.inffus.2023.101805

[19] Hassija, V., Chamola, V., Mahapatra, A., Singal, A., Goel, D., Huang, K., Scardapane, S., Spinelli, I., Mahmud, M., & Hussain, A. (2023). Interpreting black-box models: A review on explainable artificial intelligence. Cognitive Computation, 16, 45–74. https://doi.org/10.1007/s12559-023-10179-8

[20] Gramegna, A., & Giudici, P. (2021). SHAP and LIME: An evaluation of discriminative power in credit risk. Frontiers in Artificial Intelligence, 4, 752558. https://doi.org/10.3389/frai.2021.752558

[21] Famà, A., Myftiu, J., Pagnottoni, P., & Spelta, A. (2024). Explainable machine learning for financial risk management: Two practical use cases. Statistics, 58(5), 1267–1282. https://doi.org/10.1080/02331888.2024.2401078

[22] Tuge, N. B., & Msweli, N. T. (2026). Explainable artificial intelligence in the banking sector: A systematic literature review. Applied Artificial Intelligence, 40(1). https://doi.org/10.1080/08839514.2026.2676353

[23] Vuković, D. B., Dekpo-Adza, S., & Matović, S. (2025). AI integration in financial services: A systematic review of trends and regulatory challenges. Humanities and Social Sciences Communications, 12, 562. https://doi.org/10.1057/s41599-025-04850-8

[24] Heng, Y. S., & Subramanian, P. (2022). A systematic review of machine learning and explainable artificial intelligence (XAI) in credit risk modeling. In Proceedings of the Future Technologies Conference (FTC) (pp. 596–614). Springer. https://doi.org/10.1007/978-3-031-18461-1_39

[25] Babaei, G., Giudici, P., & Raffinetti, E. (2023). Explainable FinTech lending. Journal of Economics and Business, 125–126, 106126. https://doi.org/10.1016/j.jeconbus.2023.106126

[26] Nallakaruppan, M. K., Chaturvedi, H., Grover, V., Balusamy, B., Jaraut, P., Bahadur, J., Meena, V., & Hameed, I. A. (2024). Credit risk assessment and financial decision support using explainable artificial intelligence. Risks, 12(10), 164. https://doi.org/10.3390/risks12100164

[27] de Lange, P. E., Melsom, B., Vennerød, C. B., & Westgaard, S. (2022). Explainable AI for credit assessment in banks. Journal of Risk and Financial Management, 15(12), 556. https://doi.org/10.3390/jrfm15120556

[28] Bastos, J. A., & Matos, S. M. (2022). Explainable models of credit losses. European Journal of Operational Research, 301(1), 386–394. https://doi.org/10.1016/j.ejor.2021.11.009

[29] Kute, D. V., Pradhan, B., Shukla, N., & Alamri, A. (2021). Deep learning and explainable artificial intelligence techniques applied for detecting money laundering — A critical review. IEEE Access, 9, 82300–82317. https://doi.org/10.1109/access.2021.3086230

[30] Zhou, Y., Li, H., Xiao, Z., & Qiu, J. (2023). A user-centered explainable artificial intelligence approach for financial fraud detection. Finance Research Letters, 58, 104309. https://doi.org/10.1016/j.frl.2023.104309

[31] Rizinski, M., Peshov, H., Mishev, K., Chitkushev, L. T., Vodenska, I., & Trajanov, D. (2022). Ethically responsible machine learning in FinTech. IEEE Access, 10, 97531–97554. https://doi.org/10.1109/access.2022.3202889

[32] Ridzuan, N. N., Masri, M., Anshari, M., Fitriyani, N. L., & Syafrudin, M. (2024). AI in the financial sector: The line between innovation, regulation, and ethical responsibility. Information, 15(8), 432. https://doi.org/10.3390/info15080432

[33] Sabharwal, R., Miah, S. J., Wamba, S. F., & Cook, P. (2024). Extending the application of explainable artificial intelligence for managers in financial organizations. Annals of Operations Research. https://doi.org/10.1007/s10479-024-05825-9

[34] Chen, X., Ma, C., Ren, Y., Lei, Y., Huynh, N. Q. A., & Narayan, S. (2023). Explainable artificial intelligence in finance: A bibliometric review. Finance Research Letters, 56, 104145. https://doi.org/10.1016/j.frl.2023.104145

[35] Bahoo, S., Cucculelli, M., Goga, X., & Mondolo, J. (2024). Artificial intelligence in finance: A comprehensive review through bibliometric and content analysis. SN Business & Economics, 4, 23. https://doi.org/10.1007/s43546-023-00618-x

[36] Arsenault, P., Wang, S., & Patenaude, J.-M. (2024). A survey of explainable artificial intelligence (XAI) in financial time series forecasting. ACM Computing Surveys. https://doi.org/10.1145/3729531

[37] Kitchenham, B., & Charters, S. (2007). Guidelines for performing systematic literature reviews in software engineering. EBSE Technical Report EBSE-2007-01, Keele University and Durham University.

[38] Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., McGuinness, L. A., Stewart, L. A., Thomas, J., Tricco, A. C., Welch, V. A., Whiting, P., & Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ, 372, n71. https://doi.org/10.1136/bmj.n71

[39] Gode, S., Dhote, A., Giri, A., Patel, M., Reddy, P., & Thamke, N. (2026). Federated graph neural networks with explainable AI for privacy-preserving credit scoring and loan risk assessment. In 2026 IEEE 3rd International Conference on Emerging Trends in Engineering and Medical Sciences (ICETEMS 2025). IEEE. https://doi.org/10.1109/ICETEMS66917.2026.11469633

[40] Sun, X., Liu, J., & Zhang, Y. (2025). Enhancing credit risk prediction through an ensemble of explainable model. Journal of Systems Science and Systems Engineering. https://doi.org/10.1007/s11518-025-5663-y

[41] Meduri, M. N. B., Siddi, V., Pokala, A., Pallothu, V. S. K., Rajasekhar, B., Munirathnam, M., & Moturi, S. (2026). From data cleaning to deep learning: A full-stack approach to loan default risk prediction in banking applications. In 2026 2nd International Conference on Cognitive Computing in Engineering, Communications, Sciences and Biomedical Health Informatics (IC3ECSBHI 2026). IEEE. https://doi.org/10.1109/IC3ECSBHI67834.2026.11469019

[42] Wang, L., Yu, Z., Ma, J., Chen, X., & Wu, C. (2025). A two-stage interpretable model to explain classifier in credit risk prediction. Journal of Forecasting. https://doi.org/10.1002/for.3288

[43] Li, L.-H., Sharma, A. K., & Cheng, S.-T. (2025). Explainable AI-based LightGBM prediction model to predict default borrower in social lending platform. Intelligent Systems with Applications. https://doi.org/10.1016/j.iswa.2025.200514

[44] Wu, H., Zhou, F., & Pan, X. (2026). Decoupling household financial vulnerability in China: A heterogeneous stacking ensemble approach with interpretability-driven risk attribution. Expert Systems with Applications. https://doi.org/10.1016/j.eswa.2026.132819

[45] Praveena, K., Pavithra, K., Padmavathy, S., Ismankhan, Y. M., & Monisha, A. (2026). Transparent AI for financial security: Explainable deep learning for fraudulent transaction detection. In Proceedings of the International Conference on Computing, Communication, Control and Cyber-Physical Systems (I5CPS 2026). IEEE. https://doi.org/10.1109/I5CPS67958.2026.11452626

[46] Choi, S., Lim, S., Kwanggeun, K., & Seo, J. (2026). Hierarchical group-based Shapley explanation for interpretable financial fraud detection. In Proceedings of the IEEE International Conference on Big Data and Smart Computing (BIGCOMP). IEEE. https://doi.org/10.1109/BigComp68355.2026.00056

[47] Nayomi, B. D. D., Habeeb, S., Raheman, S. A., Mohammad Fahad, S. M., & Ajas, S. M. (2026). An explainable ensemble learning framework for robust and transparent credit card fraud detection. In Proceedings of the 8th International Conference on Intelligent Sustainable Systems (ICISS 2026). IEEE. https://doi.org/10.1109/ICISS67859.2026.11453962

[48] Karnavou, E., Cascavilla, G., Marcelino, G., & Geradts, Z. (2025). I know you're a fraud: Uncovering illicit activity in a Greek bank transactions with unsupervised learning. Expert Systems with Applications. https://doi.org/10.1016/j.eswa.2025.128148

[49] P. S. N., Saini, D. K. J. B., Shelke, N., Pimpalkar, A., Kumar, G. H., & P. S. H. R. (2025). Trustworthy and interpretable AI for robust fraud detection in financial transactions. In 2025 6th International Conference for Emerging Technology (INCET). IEEE. https://doi.org/10.1109/INCET64471.2025.11140975

[50] Wang, Y. (2026). Intelligent financial risk assessment and management: Ensemble learning and SHAP analysis for credit card fraud evaluation. In Proceedings of the 2nd International Conference on Digital Society, Information Science and Risk Management (ICDIR 2026). ACM. https://doi.org/10.1145/3804504.3804591

[51] Moustati, I., & Gherabi, N. (2025). Bridging behavioral insights and automated trading: An Internet of Behaviors approach for enhanced financial decision-making. Information, 16(5), 338. https://doi.org/10.3390/info16050338

[52] Wu, B., Gao, R., Wang, X., & Zeng, H. (2025). Agricultural futures prices combination forecasting with whole-process interpretable method. Applied Soft Computing, 114049. https://doi.org/10.1016/j.asoc.2025.114049

[53] Ma, T., Wang, W., & Chen, Y. (2023). Attention is all you need: An interpretable transformer-based asset allocation approach. International Review of Financial Analysis. https://doi.org/10.1016/j.irfa.2023.102876

[54] Kim, J., & Park, S. (2025). IKNet: Interpretable stock price prediction via keyword-guided integration of news and technical indicators. In Proceedings of ICAIF 2025: 6th ACM International Conference on AI in Finance. ACM. https://doi.org/10.1145/3768292.3770343

[55] Siyu, Z., & Kumar, B. V. D. (2025). Intelligent stock market forecasting: A multifactor decision-making approach combining ARIMA-LSTM models with sentiment analysis. In 2025 IEEE 3rd International Conference on Sensors, Electronics and Computer Engineering (ICSECE 2025). IEEE. https://doi.org/10.1109/ICSECE65727.2025.11257039

[56] Xiao, X., Yang, B., & Yang, B. (2026). Temporal dependencies and global context fusion: A hybrid transformer-LSTM framework for adaptive option pricing in dynamic financial markets. Information Sciences. https://doi.org/10.1016/j.ins.2026.123295

[57] Nandi, A., Ghorai, S., Banerjee, S., Ghosh, A., Das, A. K., & Dutta, S. (2026). Explainable hybrid recurrent models for stock price prediction: Integrating attention for transparency. Computational Economics. https://doi.org/10.1007/s10614-025-11285-5

[58] Wang, J. W., & Wang, X. R. (2025). Intelligent stock price forecasting integrating adaptive VMD and feature selection with BiLSTM. Computational Economics. https://doi.org/10.1007/s10614-025-11148-z

[59] Liu, W., Han, Y., Lan, X., Yu, Z., Xia, M., Lin, S., Pang, C., & Chen, N. (2026). Progressive gradient boosted trees for imbalanced financial distress prediction. Expert Systems with Applications. https://doi.org/10.1016/j.eswa.2026.132187

[60] Lin, Y.-C., Padliansyah, R., Lu, Y.-H., & Liu, W.-R. (2025). Bankruptcy prediction: Integration of convolutional neural networks and explainable artificial intelligence techniques. International Journal of Accounting Information Systems. https://doi.org/10.1016/j.accinf.2025.100744

[61] Wang, P., & Borjigin, S. (2026). Bank systemic risk prediction based on text mining and explainable machine learning. North American Journal of Economics and Finance. https://doi.org/10.1016/j.najef.2025.102577

[62] Nandini, D., Shamine Carina, B., Rekhaa, P., Yogeshwari, R., Kishore, S. B., & Hemapriya, K. (2026). An explainable neural modeling approach for financial risk governance and transparency-driven managerial decision support. In Proceedings of ICSES 2026: 5th International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems. IEEE. https://doi.org/10.1109/ICSES66558.2026.11478973

[63] Cristescu, M. P., Brândăş, C., Mara, D. A., & Ioana, P. (2025). Fine-tuning and explaining FinBERT for sector-specific financial news: A reproducible workflow. Electronics, 14(23), 4680. https://doi.org/10.3390/electronics14234680

[64] Aziz, M., Danish, M. U., Grolinger, K., Rehman, U., & Iqbal, F. (2025). Finite-difference approximation for explaining neural network predictions in financial credit evaluation. In Proceedings of the 2025 International Conference on Advanced Machine Learning and Data Science (AMLDS). IEEE. https://doi.org/10.1109/AMLDS63918.2025.11159472

[65] Kibriya, M. G., Masud, N., Al Bassam, A., Sourav, M. T. I., & Sobhan, A. (2025). Explainable AI (XAI) transparency for financial fraud detection. In 2025 IEEE International Conference on Quantum Photonics, Artificial Intelligence, and Networking (QPAIN). IEEE. https://doi.org/10.1109/QPAIN66474.2025.11172108

[66] Dikmen, M., & Burns, C. (2022). The effects of domain knowledge on trust in explainable AI and task performance: A case of peer-to-peer lending. International Journal of Human-Computer Studies, 162, 102792. https://doi.org/10.1016/j.ijhcs.2022.102792

[67] Tiukhova, E., Salcuni, A., Oguz, C., Niglio, M., Storti, G., Forte, F., Baesens, B., & Snoeck, M. (2025). Boosting credit risk data quality using machine learning and explainable AI techniques. In Communications in Computer and Information Science. Springer. https://doi.org/10.1007/978-3-031-74643-7_30

[68] Jo, A. A., & Raj, E. D. (2026). Auditable credit scoring: Explainable models with stable rationales on tabular data. In Proceedings of the 2026 International Conference on AI-Driven Smart Systems and Ubiquitous Computing (ICAUC 2026). IEEE. https://doi.org/10.1109/ICAUC68182.2026.11441016

[69] Atef, M., Ouf, S., Seoud, W., & Gabr, M. I. (2025). A novel approach using explainable prediction of default risk in peer-to-peer lending based on machine learning models. Neural Computing and Applications. https://doi.org/10.1007/s00521-025-11489-8

[70] Mazumder, P. T. (2026). Explainable and fair anti-money laundering models using a reproducible SHAP framework for financial institutions. Discover Artificial Intelligence. https://doi.org/10.1007/s44163-026-00944-7

[71] Ghosh, S., Baitalik, S., Datta, R., Mukherjee, R., Sarkar, D., & Chaudhuri, A. (2026). Explanation-first agentic forecaster for stock market. In 2026 9th International Conference on Electronics, Materials Engineering and Nano-Technology (IEMENTech 2026). IEEE. https://doi.org/10.1109/IEMENTech202669403.2026.11434302

[72] Ali, A. A. M., Furaijl, H. B., Kadhim, M. A., Mahmood, I. M., Al-Anezi, W., Howaidi, O., Ali, R. J., Muqdad, H. H., & Faead, A. (2025). Predictive modeling of bankruptcy risk in SMEs using explainable AI with temporal and financial health indicators. In Proceedings of ICCR 2025: 3rd International Conference on Cyber Resilience. IEEE. https://doi.org/10.1109/ICCR67387.2025.11292014

[73] Aruleba, I., & Sun, Y. (2025). Enhanced credit risk prediction using deep learning and SMOTE-ENN resampling. Machine Learning with Applications. https://doi.org/10.1016/j.mlwa.2025.100692

[74] Weng, F. T., Zhu, M., Buckle, M., Hajek, P., & Abedin, M. Z. (2025). Class imbalance Bayesian model averaging for consumer loan default prediction: The role of soft credit information. Research in International Business and Finance. https://doi.org/10.1016/j.ribaf.2024.102722

[75] Syambabu, J., Doppalapudi, K., Rao, T. S., Srinivasarao, P., Munirathnam, M., & Medabalimi, G. (2026). Explainable ensemble framework for imbalanced credit card fraud detection with feature selection. In Proceedings of the 5th International Conference on Communication, Computing and Electronics Systems (ICCCES 2026). IEEE. https://doi.org/10.1109/ICCCES62661.2026.11437241

[76] Okafor, K., Ilyas, M., & Althebyan, Q. (2025). Tackling class imbalance and model transparency in interpretable machine learning for term deposit prediction. In International Conference on Social Networks Analysis, Management and Security (SNAMS). IEEE. https://doi.org/10.1109/SNAMS67467.2025.11391106

[77] Abbas Jasim Al-Hchaimi, A., Khalifa, M. A., & El-Shafai, W. (2026). Explainable AI with imbalanced learning strategies for blockchain transaction fraud detection. Engineering Reports. https://doi.org/10.1002/eng2.70545

[78] Sri, G. S., Lalitha, R., Viswanth, C. S., Sravya, V. L., & Sulthana, S. R. (2026). Scalable credit-card fraud detection using a RAAE and XGBoost. In 2026 2nd

International Conference on Cognitive Computing in Engineering, Communications, Sciences and Biomedical Health Informatics (IC3ECSBHI 2026). IEEE. https://doi.org/10.1109/IC3ECSBHI67834.2026.11469130

[79] Kar, V., & Alam, K. M. (2025). Adaptive ensemble technique-based robust loan risk prediction for financial systems. In 2025 International Conference on Computing Technologies and Data Communication (ICCTDC 2025). IEEE. https://doi.org/10.1109/ICCTDC64446.2025.11158985

[80] Iqbal, S., Awan, K. M., Kamal, S., & Rehman, Z. U. (2025). Interpretable ensemble learning models for credit card fraud detection. Applied Sciences, 15(22), 12073. https://doi.org/10.3390/app152212073

[81] Kapoor, S., Jain, S., Gupta, P. K., & Bhardwaj, S. (2026). Integrating explainability and fairness in credit risk prediction: A hybrid approach using TabNet, LightGBM, SHAP, and counterfactual explanations on the FICO dataset. In 2026 2nd International Conference on Computing, Sciences and Communications (ICCSC 2026). IEEE. https://doi.org/10.1109/ICCSC67078.2026.11468663

[82] Vivek, Y., Ravi, V., Mane, A., & Naidu, L. R. (2024). Explainable artificial intelligence and causal inference based ATM fraud detection. In 2024 IEEE Symposium on Computational Intelligence for Financial Engineering and Economics (CIFEr). IEEE. https://doi.org/10.1109/CIFER62890.2024.10772906

[83] Roy, N. S., Ketan, S. H. N., & Mohebbanaaz. (2026). A hybrid drift-aware causal and counterfactual explainability framework for transparent loan default prediction. In 2026 IEEE 15th International Conference on Communication Systems and Network Technologies (CSNT). IEEE. https://doi.org/10.1109/CSNT69054.2026.11502344

[84] Straus, J. V. D., Andreev, V. K., Koeleman, T., Sepanosian, T., & Machado, M. R. (2026). Explaining to defend: SHAP-based defense mechanism against adversarial attacks in P2P credit risk. Expert Systems with Applications. https://doi.org/10.1016/j.eswa.2025.130432

[85] Moura, L., Barcaui, A., & Payer, R. (2025). AI and financial fraud prevention: Mapping the trends and challenges through a bibliometric lens. Journal of Risk and Financial Management, 18(6), 323. https://doi.org/10.3390/jrfm18060323

[86] van Veen, K., Ahmed, F., & van Keulen, M. (2026). XAI in fraud detection: A causal perspective. Communications in Computer and Information Science. https://doi.org/10.1007/978-3-032-08330-2_15

[87] Vakrani, D. S., Padhye, P. S., Gupta, S. K., Roy, J. K., Nerlekar, V. S., & Parashar, N. (2026). Evaluating AI-driven credit scoring models versus traditional statistical techniques. Discover Artificial Intelligence. https://doi.org/10.1007/s44163-025-00772-1

[88] Thomas, R., & Savitha, K. K. (2026). Pre-transaction fraud risk prediction in DeFi using explainable AI. In 2026 6th International Conference on Advances in Electrical, Computing, Communications and Sustainable Technologies (ICAECT 2026). IEEE. https://doi.org/10.1109/ICAECT68478.2026.11426021

[89] Thar, K. W., & Wai, T. T. (2025). Enhancing transaction fraud detection in digital banking with explainable AI. In 2025 6th International Conference on Advanced Information Technologies (ICAIT 2025). IEEE. https://doi.org/10.1109/ICAIT68809.2025.11236776

# A Appendix A

## A.1 Search Strings

The following search queries were used on Scopus, Web of Science, and IEEE Xplore.

```
Scopus:
( TITLE-ABS-KEY ( "Asset evaluation" OR "banking" OR "credit risk"
OR "finance" OR "financial performance" OR "portfolio optimi*"
OR "portfolio selection" OR "financial planning" OR "bankruptcy"
OR "credit scoring" OR "financial distress" OR "loan default"
OR "Financial Advice" OR "Financial Market" OR "Option pricing" ) )
AND ( TITLE-ABS-KEY ( "explainable artificial intelligence" OR "xai"
OR "interpretable machine learning" OR "explainable ai"
OR "explainable machine learning" OR "interpretability"
OR "ai transparency" OR "SHAP" OR "LIME" ) )
AND ( LIMIT-TO ( PUBSTAGE , "final" ) )
AND ( LIMIT-TO ( SUBJAREA , "COMP" ) OR LIMIT-TO ( SUBJAREA , "MATH" )
OR LIMIT-TO ( SUBJAREA , "ENGI" ) OR LIMIT-TO ( SUBJAREA , "DECI" )
OR LIMIT-TO ( SUBJAREA , "BUSI" ) OR LIMIT-TO ( SUBJAREA , "ECON" ) )
AND ( EXCLUDE ( DOCTYPE , "re" ) )
AND ( LIMIT-TO ( LANGUAGE , "English" ) )

Web of Science:
(TS=(("Asset evaluation" OR "banking" OR "credit risk" OR "finance"
OR "financial performance" OR "portfolio optimi*" OR "portfolio selection"
OR "financial planning" OR "bankruptcy" OR "credit scoring"
OR "financial distress" OR "loan default" OR "Financial Advice"
OR "Financial Market" OR "Option pricing")
AND
("explainable artificial intelligence" OR "xai"
OR "interpretable machine learning" OR "explainable ai"
OR "explainable machine learning" OR "interpretability"
OR "ai transparency" OR "SHAP" OR "LIME"))
AND WC=(Computer Science OR Mathematics OR Engineering
OR Operations Research Management Science OR Business OR Economics)
AND LA=English) NOT (DT==("REVIEW"))
```

```
IEEE Xplore:
(("Asset evaluation" OR "banking" OR "credit risk" OR "finance"
OR "financial performance" OR "portfolio optimi*" OR "portfolio selection"
OR "financial planning" OR "bankruptcy" OR "credit scoring"
OR "financial distress" OR "loan default" OR "Financial Advice"
OR "Financial Market" OR "Option pricing")
AND
("explainable artificial intelligence" OR "xai"
OR "interpretable machine learning" OR "explainable ai"
OR "explainable machine learning" OR "interpretability"
OR "ai transparency" OR "SHAP" OR "LIME"))
```

# B Appendix B

## B.1 Included-study register and extraction workbook

The full included-study register is maintained in the supplementary workbook *Extraction Sheet.xlsx*, which is submitted as supplementary material with the manuscript. The workbook contains 297 non-empty study records in the sheet named *Data*. This sheet is the authoritative included-study appendix for the review.

The workbook also contains supporting sheets named *README*, *CODEBOOK*, *EXTRACTION*, *PRISMA_Log*, and *Verification*. These sheets document the extraction conventions, PRISMA screening trace, and verification checks used to support the manuscript. Table 15 summarises the fields in the *Data* sheet so that readers can interpret the supplementary appendix without relying on undocumented spreadsheet columns.

For article-length reasons, the 297 rows are not reproduced in the printed manuscript. Table 16 lists the 97 high-coverage studies that form the basis of the sensitivity analysis described in Section 3.4. The supplementary workbook should be cited wherever detailed per-study traceability is required, including eligibility decisions, extraction evidence, and research-question coverage.

## B.2 High-coverage sample used for the sensitivity analysis

The high-coverage sample contains the 97 studies that provide evidence for at least three of the four review dimensions. These studies form the basis of the sensitivity analysis reported in Section 5.5, which tests whether the low corpus-wide rate of explanation-quality evaluation persists among the most fully reported studies.

**Table 16**: High-coverage studies used for the sensitivity analysis ($n = 97$)

| ID | Year | Domain | Study title |
|---|---|---|---|
| P001 | 2026 | Fraud Detection | An Intelligent Optimization-Based Deep Belief Network for Fraud Detection in Financial Transaction Systems |
| P002 | 2026 | Credit Risk Assessment | KredAI: Federated Learning Simulation for Credit Risk Assessment Using SHAP Explainability |
| P004 | 2026 | Fraud Detection | An Explainable Ensemble Learning Framework for Robust and Transparent Credit Card Fraud Detection |
| P005 | 2025 | Credit Risk Assessment | Explainable Domain Adaptation Learning Framework for Credit Scoring in Internet Finance Through Adversarial Transfer Learning and Ensemble Fusion Model |

Continued on next page

**Table 16**: High-coverage studies used for the sensitivity analysis ($n = 97$) (continued)

| ID | Year | Domain | Study title |
|---|---|---|---|
| P006 | 2026 | Credit Risk Assessment | From Data Cleaning to Deep Learning: A Full-Stack Approach to Loan Default Risk Prediction in Banking Applications |
| P008 | 2025 | Credit Risk Assessment | Enhancing Credit Risk Prediction through an Ensemble of Explainable Model |
| P010 | 2025 | Credit Risk Assessment | A novel approach using explainable prediction of default risk in peer-to-peer lending based on machine learning models |
| P011 | 2026 | Fraud Detection | Explainable and fair anti-money laundering models using a reproducible SHAP framework for financial institutions |
| P018 | 2025 | Stock | Bridging Behavioral Insights and Automated Trading: An Internet of Behaviors Approach for Enhanced Financial Decision-Making |
| P019 | 2026 | Credit Risk Assessment | Integrating Explainability and Fairness in Credit Risk Prediction: A Hybrid Approach Using Tabnet, LightGBM, SHAP, and Counterfactual Explanations on the FICO Dataset |
| P020 | 2026 | Fraud Detection | Hierarchical Group-Based Shapley Explanation for Interpretable Financial Fraud Detection |
| P022 | 2026 | Credit Risk Assessment | Ensemble Learning and SHAP-based Interpretability for AI-Driven Credit Risk Assessment in Financial Default Prediction |
| P025 | 2026 | Fraud Detection | XAI In Fraud Detection: A Causal Perspective |
| P026 | 2026 | Credit Risk Assessment | Auditable Credit Scoring: Explainable Models with Stable Rationales on Tabular Data |
| P027 | 2026 | Fraud Detection | Transparent AI for Financial Security: Explainable Deep Learning for Fraudulent Transaction Detection |
| P028 | 2025 | Stock | Fine-Tuning and Explaining FinBERT for Sector-Specific Financial News: A Reproducible Workflow |
| P029 | 2026 | Stock | Capturing Short- and Long-Term Temporal Dependencies Using Bahdanau-Enhanced Fused Attention Model for Financial Data—An Explainable AI Approach |
| P030 | 2025 | Credit Risk Assessment | A Two-Stage Interpretable Model to Explain Classifier in Credit Risk Prediction |

Continued on next page

**Table 16**: High-coverage studies used for the sensitivity analysis ($n = 97$) (continued)

| ID | Year | Domain | Study title |
|---|---|---|---|
| P032 | 2025 | Stock | Agricultural futures prices combination forecasting with whole-process interpretable method |
| P033 | 2025 | Credit Risk Assessment | Explainable AI based LightGBM prediction model to predict default borrower in social lending platform |
| P034 | 2026 | Financial Distress | Bank systemic risk prediction based on text mining and explainable machine learning |
| P038 | 2026 | Credit Risk Assessment | Explaining to defend: SHAP-based defense mechanism against adversarial attacks in P2P credit risk |
| P040 | 2026 | Credit Risk Assessment | Federated Graph Neural Networks with Explainable AI for Privacy-Preserving Credit Scoring and Loan Risk Assessment |
| P042 | 2026 | Credit Risk Assessment | Advancing Loan Approval Prediction With SHAP-Guided Feature Selection and LIME-Based Model Interpretability in a Multiclassifier Context Through a Web-Based Application Development Approach |
| P045 | 2026 | Fraud Detection | A Privacy-Preserving and Interpretable Framework for Financial Fraud Detection Using AI |
| P048 | 2026 | Credit Risk Assessment | Decoupling household financial vulnerability in China: A heterogeneous stacking ensemble approach with interpretability-driven risk attribution |
| P050 | 2025 | Financial Distress | Bankruptcy prediction: Integration of convolutional neural networks and explainable artificial intelligence techniques |
| P051 | 2025 | Credit Risk Assessment | Credit Behavior Scorecards Based on TabTransformer and Model Interpretability Exploring |
| P052 | 2026 | Fraud Detection | Integrating Model Explainability and Uncertainty Quantification for Trustworthy Fraud Detection |
| P053 | 2026 | Stock | Hybrid Deep Learning Model for Accurate and Explainable Stock Market Prediction |
| P054 | 2025 | Stock,Revisit | How well do machine learning models in finance work? |
| P055 | 2026 | Credit Risk Assessment | Evaluating AI-driven credit scoring models versus traditional statistical techniques |

Continued on next page

**Table 16**: High-coverage studies used for the sensitivity analysis ($n = 97$) (continued)

| ID | Year | Domain | Study title |
|---|---|---|---|
| P057 | 2025 | Stock | IKNet: Interpretable Stock Price Prediction via Keyword-Guided Integration of News and Technical Indicators |
| P058 | 2025 | Fraud Detection | Interpretable Ensemble Learning Models for Credit Card Fraud Detection |
| P059 | 2025 | Stock | Intraday and Post-Market Investor Sentiment for Stock Price Prediction: A Deep Learning Framework with Explainability and Quantitative Trading Strategy |
| P060 | 2025 | Stock | SHAP-Based Recursive Feature Elimination and Hyperparameter Optimization for Enhanced Financial Stock Forecasting |
| P062 | 2026 | Stock | Explainable Stock Recommendation System Using Hybrid Signal Modeling and Reinforcement Learning |
| P063 | 2026 | Stock | Explanation-First Agentic Forecaster for Stock Market |
| P064 | 2026 | Stock | Explainable Ensemble Learning for Predicting Stock Market Crises: Calibration, Threshold Optimization, and Robustness Analysis |
| P065 | 2026 | Fraud Detection | Pre-Transaction Fraud Risk Prediction in DeFi Using Explainable AI |
| P067 | 2026 | Financial Distress | Frequency-aware gradient modulated boosted trees for interpretable financial distress prediction |
| P068 | 2026 | Generic Finance | Bridging Data Silos in Corporate Governance: A Hierarchical Stacking Ensemble With Federated Dynamic Aggregation |
| P069 | 2025 | Credit Risk Assessment | JointLIME: An interpretation method for machine learning survival models with endogenous time-varying covariates in credit scoring |
| P070 | 2025 | Generic Finance | Accuracy, precision, recall, f1-score, or MCC? empirical evidence from advanced statistics, ML, and XAI for evaluating business predictive models |
| P073 | 2026 | Credit Risk Assessment | A fuzzy logistic-based ensemble framework with heterogeneous weighting effects for credit risk evaluation |
| P074 | 2026 | Credit Risk Assessment | Peer-To-Peer Lending: Machine Learning Insights for Enhanced Understanding |

Continued on next page

**Table 16**: High-coverage studies used for the sensitivity analysis ($n = 97$) (continued)

| ID | Year | Domain | Study title |
|---|---|---|---|
| P076 | 2026 | Credit Risk Assessment | Federated and Interpretable AI Framework for Secure and Transparent Loan Default Prediction in Financial Institutions |
| P079 | 2026 | Credit Risk Assessment | From Black Box to Clarity: Machine Learning and Agnostic Techniques in Credit Risk Management |
| P080 | 2026 | Credit Risk Assessment | Interpretable AI Framework for Credit Risk Assessment and Fraud Detection |
| P081 | 2026 | Financial Distress | An Explainable Neural Modeling Approach for Financial Risk Governance and Transparency-Driven Managerial Decision Support |
| P082 | 2025 | Credit Risk Assessment | Multilayer Credit Card Fraud Detection Using a Hybrid Autoencoder-BiLSTM-XGBoost Framework with XAI |
| P083 | 2026 | Credit Risk Assessment | Modeling Bias and Transparency in Credit Scoring Algorithms: A Differentially Private SHAP-Based Framework |
| P084 | 2026 | Fraud Detection | Intelligent Financial Risk Assessment and Management: Ensemble Learning and SHAP Analysis for Credit Card Fraud Evaluation |
| P086 | 2026 | Fraud Detection | Towards Explainable AI: A Context-Aware Hybrid Framework for Fraud Detection in Financial Market |
| P088 | 2026 | Stock | Temporal dependencies and global context fusion: A hybrid transformer-LSTM framework for adaptive option pricing in dynamic financial markets |
| P090 | 2024 | Stock | Unlocking ETF price forecasting: Exploring the interconnections with statistical dependence-based graphs and xAI techniques |
| P091 | 2025 | Credit Risk Assessment | Interpretable Credit Card Fraud Detection Using LightGBM and XAI Techniques |
| P092 | 2025 | Credit Risk Assessment | Predictive Modeling of Bankruptcy Risk in SMEs Using Explainable AI with Temporal and Financial Health Indicators |
| P093 | 2025 | Credit Risk Assessment | Multi-Layer Perceptron with Dropout Regularization for Credit Scoring on Imbalanced Financial Datasets |

Continued on next page

**Table 16**: High-coverage studies used for the sensitivity analysis ($n = 97$) (continued)

| ID | Year | Domain | Study title |
|---|---|---|---|
| P094 | 2025 | Fraud Detection | A Hybrid Approach to Fraud Detection: Integrating Explainable Deep Learning with Blockchain-Based Identity Verification |
| P096 | 2025 | Fraud Detection | Privacy-Preserving and Explainable AI in Credit Card Fraud Detection: Balancing Accuracy, Transparency and Security |
| P097 | 2025 | Stock | Integrating Machine Learning with Financial Risk Modeling for Portfolio Management |
| P099 | 2025 | Fraud Detection | Explainable AI (XAI) Transparency for Financial Fraud Detection |
| P100 | 2025 | Credit Risk Assessment | Finite-Difference Approximation for Explaining Neural Network Predictions in Financial Credit Evaluation |
| P101 | 2025 | Stock | An Interpretable Framework for Stock Market Forecasting Using Long Short-Term Memory (LSTM) Networks with SHAP - Explainable AI (XAI) Method |
| P102 | 2024 | Credit Risk Assessment | Explainable Deep Learning Approaches to Credit Risk Evaluation |
| P110 | 2025 | Credit Risk Assessment | Enhancing Credit Risk Decision Making in Supply Chain Finance with an Interpretable Random Forest Model Optimized by SVM |
| P120 | 2025 | Fraud Detection,Credit Risk Assessment | Credit Card Fraud Detection using Machine Learning Models and Increasing Explainability using Explainable AI Methods |
| P124 | 2025 | Credit Risk Assessment | Active Learning and Uncertainty-Driven Labeling for Improving Financial Risk Models in B2B Corporate Client |
| P129 | 2025 | Credit Risk Assessment | Explainable Gradient Boosting Models for Fair Loan Approval Decisions |
| P131 | 2024 | Credit Risk Assessment | Credit risk prediction based on causal machine learning: Bayesian network learning, default inference, and interpretation |
| P152 | 2024 | Fraud Detection | Explainable Artificial Intelligence and Causal Inference based ATM Fraud Detection |
| P158 | 2025 | Fraud Detection | A Hybrid Framework for Privacy-Preserving and Explainable Credit Card Fraud Detection Using XGBoost and Homomorphic Encryption |
| P159 | 2025 | Fraud Detection | Enhancing Transaction Fraud Detection in Digital Banking with Explainable AI |

Continued on next page

**Table 16**: High-coverage studies used for the sensitivity analysis ($n = 97$) (continued)

| ID | Year | Domain | Study title |
|---|---|---|---|
| P160 | 2024 | Credit Risk Assessment | Unlocking Transparency in Credit Scoring: Leveraging XGBoost with XAI for Informed Business Decision-Making |
| P162 | 2025 | Stock | Advanced Financial Sentiment Analysis of Global Stock Market News Using Transformer-Based Deep Language Understanding Models |
| P163 | 2025 | Credit Risk Assessment | Explainable AI (XAI) Using SHAP and LIME for Financial Fraud Detection and Credit Scoring |
| P166 | 2025 | Credit Risk Assessment | Boosting Credit Risk Data Quality Using Machine Learning and eXplainable AI Techniques |
| P173 | 2020 | Generic Finance | Towards Responsible AI for Financial Transactions |
| P183 | 2022 | Revisit,Generic Finance | The effects of domain knowledge on trust in explainable AI and task performance: A case of peer-to-peer lending |
| P188 | 2023 | Stock | Evaluating Transparency: A Cross-Model Exploration of Explainable AI in Financial Forecasting |
| P196 | 2024 | Credit Risk Assessment | Linear Discriminant Analysis in Credit Scoring: A Transparent Hybrid Model Approach |
| P197 | 2021 | Credit Risk Assessment | Fairness by Explicability and Adversarial SHAP Learning |
| P201 | 2022 | Fraud Detection | RESHAPE: Explaining Accounting Anomalies in Financial Statement Audits by enhancing SHapley Additive exPlanations |
| P223 | 2026 | Credit Risk Assessment | Explainable credit risk modeling with hybrid tabular deep learning and adaptive feature routing |
| P231 | 2021 | Fraud Detection | Towards Design Principles for User-Centric Explainable AI in Fraud Detection |
| P232 | 2026 | Generic Finance | Machine Learning XAI for Early Loan Default Prediction |
| P234 | 2023 | Financial Distress | Interpreting the prediction results of the tree-based gradient boosting models for financial distress prediction with an explainable machine learning approach |

Continued on next page

**Table 16**: High-coverage studies used for the sensitivity analysis ($n = 97$) (continued)

| ID | Year | Domain | Study title |
|---|---|---|---|
| P246 | 2024 | Financial Distress | Financial Distress Prediction Using an Artificial Neural Network Integrating a Two- Stage Feature Selection Method |
| P275 | 2026 | Credit Risk Assessment | A Hybrid Drift-Aware Causal and Counterfactual Explainability Framework for Transparent Loan Default Prediction |
| P279 | 2026 | Stock | Explainable AI Techniques for Transparent Algorithmic Trading Decisions |
| P283 | 2026 | Fraud Detection | User Centric Explainable Machine Learning for Financial Fraud Detection |
| P288 | 2026 | Credit Risk Assessment | Hierarchical Uncertainty Aware Explainable Hybrid Deep Learning Framework for Credit Default Risk Prediction |
| P289 | 2025 | Credit Risk Assessment | A Continual Learning, Explainable AI Model for Credit Risk with Neuro-Symbolic Integration in a Fuzzy Framework |
| P301 | 2025 | Credit Risk Assessment | Distributional Feature Separability for Explainability of Neural Networks for Financial Credit Evaluation |
| P309 | 2025 | Fraud Detection | Non-Linear Influence Analysis for Adaptive Fraud Prevention in FinTech Using Hybrid Learning Models |
| P323 | 2026 | Fraud Detection | TGNN-Powered GST Fraud Detection and Monitoring System with XAI - 2026 Fourth International Conference on Augmented Intelligence and Sustainable Systems (ICAISS) |

**Table 15** Fields in the included-study register, sheet *Data* ($n = 297$)

| Field group | Workbook columns |
|---|---|
| Study identifiers | Paper_ID; Rayyan_ID; Authors; Year; Title; Venue; DOI |
| Retrieval and eligibility | FullText_Retrieved; FullText_Eligible; FT_Exclusion_Reason |
| Financial context | Finance_Domain; Finance_Subapplication; Country_Region |
| Data characteristics | Dataset; Dataset_Type |
| Model and task | ML_Model; ML_Task_Type; Predictive_Performance |
| Explainability method | XAI_Technique; XAI_CoPresent_Methods; Explanation_Quality_Reported; XAI_Evaluation_Done; XAI_Eval_Quant_Metric; XAI_Eval_Qual_Method |
| Deployment barriers | Barriers_Technical; Barriers_Data; Barriers_DomainSpecific |
| Validation and synthesis | Method_Approach_RQ4; RQ_Coverage; Evidence_Notes_and_Pages |

## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ExtractionSheet.xlsx