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

Systematic Literature Review on Explainable Ai in Finance: Methods, Applications and Research Gaps

Alok Bhardwaj

Asst. Prof., G. L. Bajaj Institute of Management and Research, Greater Noida

Abstract

This systematic literature review (SLR) on finance's Explainable Artificial Intelligence (XAI) uses bibliometric mapping and topic synthesis of 162 Scopus-indexed publications from 2010–2025. Credit scoring, fraud detection, algorithmic trading, and risk management are being transformed by AI and machine learning, yet "black-box" algorithms still hinder financial organizations. SHAP, LIME, Integrated Gradients, and Layer-wise Relevance Propagation increase model transparency, accountability, and confidence. Their finance utilization is disorganized and cognitively immature.

PRISMA-2020 and Kitchenham's (2007) methodologies are used to identify, screen, and synthesize peer-reviewed publications to map the conceptual, methodological, and application landscape of XAI in finance. After 2018, publishing growth exploded, aligning with legislative frameworks like the EU AI Act and institutional demand for explainability. Thematic grouping shows three primary study areas: post-hoc vs intrinsic interpretability methodological improvements, domain applications in credit risk, fraud detection, and portfolio optimization, and governance problems including fairness, auditability, and regulatory compliance. Decision trees, GAMs, and emerging hybrid frameworks balance accuracy and transparency as intrinsic interpretability techniques, whereas SHAP and LIME dominate post-hoc procedures.

The review highlighted institutional acceptability issues due to a lack of explainability standards, overreliance on post-hoc attribution methods, insufficient behavioral-trust theory integration, and incorrect regulatory audit procedures. Combining XAI, perceived trust, adoption, and financial outcomes addresses these issues. Trust measures algorithmic transparency for decision quality, compliance, and sustainability. This paradigm makes explainability a technological and organizational capability for trustworthy AI governance.

1. Adds reliable bibliometric evidence to a fragmented subject,
2. Introduces trust-based XAI adoption,
3. Connects explainability to ethical and sustainable finance.

For responsible digital transformation, XAI makes efficient, auditable, and socially accountable algorithmic finance decisions, study finds.

Keywords: Explainable Artificial Intelligence (XAI); Financial Technology; Trust; Interpretability; Model Transparency; PRISMA Systematic Review; Ethical AI; Responsible Finance.

1. INTRODUCTION

1.1. Background and Motivation

AI and ML have altered financial risk assessment, capital allocation, fraud detection, and investment strategy. Deep neural networks, gradient boosting, and ensemble designs have outperformed humans in credit risk assessment, algorithmic trading, and portfolio optimization for a decade. Accurate but opaque "black-box" algorithms are likewise troublesome (Doshi-Velez & Kim, 2017). Financial institutions use complicated ML algorithms with incoherent decision logic, affecting fairness, accountability, and regulatory compliance (Guidotti et al., 2018).

Explainable Artificial Intelligence (XAI) makes AI systems transparent, interpretable, and trustworthy, solving the "black-box" dilemma. SHAP, LIME, Integrated Gradients, and Layer-wise Relevance Propagation aid XAI model interpretation (Lundberg & Lee, 2017; Ribeiro et al., 2016). This method lets analysts assess feature importance, illustrate decision limits, and check financial AI ideas. XAI's ability to explain opaque systems aids the financial industry because algorithmic judgments affect creditworthiness, insurance pricing, and investment risk (Kraus et al., 2020).

The EU's AI Act, GDPR, and state banking legislation highlighting the "right to explanation" have increased demand for explainability (European Commission, 2021). These findings show that algorithmic decision-making trust and accountability require explainability (Goodman & Flaxman, 2017). Investors, auditors, and consumers desire AI-driven credit rating, robo-advisory, and anti-money-laundering transparency (Tobback et al., 2018). Sustainable banking AI integration involves XAI technology and regulation.

Although XAI research is increasing rapidly, its implementation in finance is fragmented. Technical algorithm creation research often disregard sector-specific interpretability and user-centric evaluation criteria (Arrieta et al., 2020). Thus, it is necessary to aggregate studies and evaluate how XAI has progressed as a finance field—its dominant methodologies, theoretical foundations, and uncovered gaps. This research addresses this need by doing a systematic literature review (SLR) of explainable AI in finance using bibliometric and thematic analysis of peer-reviewed Scopus-indexed publications from 2010 to 2025.

1.2. Research Objectives

This paper discusses Explainable AI in finance's history, methods, and applications. Review serves three goals. XAI's financial conceptual and analytical frameworks are mapped to synthesize the literature. Review credit risk modeling, fraud detection, portfolio optimization, and FinTech risk management interpretability definitions and operations. This bibliometric synthesis highlights publishing trends, co-authorship networks, and topic evolution in the discipline.

Second, the study classifies research by (a) XAI methods—post-hoc versus intrinsic interpretability, (b) learning paradigms—supervised, unsupervised, or hybrid, and (c) financial applications—banking, insurance, investments, and regulatory analytics—to identify methodological trends and This classification of credit risk models and real-time trading algorithms shows that XAI techniques vary in domain specificity and interpretability. We study these junctions to understand financial explainability's technical progress.

Third, it identifies research gaps and suggests further work. Initial research suggests that financial models lack explainability benchmarks (Molnar, 2022) and that XAI's effects on user trust, decision quality, and regulatory compliance are unknown (Biran & Cotton, 2017). Novel interpretability algorithms have been proposed, however few research have examined their use in financial systems or ethics and behavior.

These gaps are identified to lay the framework for XAI research in technology innovation, behavioral finance, and policy governance.

1.3. Scope and Contribution

This evaluation includes only Scopus-indexed peer-reviewed journal and conference works on explainable AI in finance from 2010–2025. It excludes healthcare and industry studies unless they demonstrate methodological transferability to finance. Analytics include AI applications that automate or affect financial decision-making such credit scoring, fraud detection, algorithmic trading, robo-advisory, and insurance underwriting.

This study makes three contributions. First, it frequently synthesizes XAI financial research using quantitative bibliometric and qualitative subject analysis. Second, it provides a multi-dimensional taxonomy of financial interpretability methods, showing how post-hoc methods like SHAP and LIME coexist with transparent models like decision trees and generalized additive models (Molnar, 2022). Third, it addresses interpretability-predictive evaluation criteria, stakeholder trust, and ethical governance research gaps.

Summing together scattered knowledge enhances financial XAI theory and practice. Explainability for ethical and sustainable AI adoption utilizing data science and finance is emphasized. Researchers, politicians, and practitioners can use the findings to merge algorithmic intelligence with openness, accountability, and social welfare.

2. Methodology (PRISMA-aligned Systematic Literature Review)

2.1. Review Design Framework

Our systematic literature review (SLR) follows PRISMA 2020 (Page et al., 2021) and Kitchenham and Charters' (2007) evidence-based software engineering research approach. To achieve methodological traceability and verifiability from identification to synthesis, the PRISMA paradigm was chosen for its openness, replicability, and comprehensiveness. Thus, the review meets high-impact journal procedure documentation and data disclosure norms.

The review had four PRISMA phases:

1. First, find relevant publications using structured database searches.
2. Title, abstract, and keyword screening.
3. Full-text eligibility evaluation.
4. Study completion meets standards.

The PRISMA 2020 flow diagram (not shown) counted identified, screened, excluded, and included articles. This lets future researchers recreate the dataset and evaluate results using identical search strings and inclusion settings. Kitchenham's method is rigorous due to pre-defined study subjects, inclusion–exclusion criteria, and quality assessment rubrics (Kitchenham et al., 2009

Following AI and finance review best practises, this study uses quantitative bibliometric analysis and qualitative theme synthesis (Arrieta et al., 2020; Kraus, 2020). This hybrid methodology describes financial explainable AI (XAI) research publishing and methodological and application trends.

2.2. Search Strategy

This review relied on Scopus, a renowned peer-reviewed computer science, business, and finance database. Secondary validation using WoS validated search results for completeness and decreased publication bias. For better AI and finance journal indexing, Scopus was chosen (Li et al., 2023).

The search covers explainable AI's financial application development from January 2010 to October 2025. Only seminal interpretability or algorithmic transparency studies were assessed.
Boolean logic and iterative pilot searches balanced precision and recall in search strings. The final Scopus query:
(TITLE-ABS-KEY("explainable artificial intelligence" OR "explainable AI" OR "interpretable machine learning" OR "model interpretability" OR "model transparency" OR "AI explainability")
AND TITLE-ABS-KEY("finance" OR "financial services" OR "banking" OR "credit risk" OR "fraud detection" OR "investment" OR "insurance" OR "fintech")) AND (LIMIT-TO(LANGUAGE, "English")) AND (PUBYEAR > 2009) Run Web of Science equivalent formulations for cross-database compatibility. Mendeley and R removed duplicates.
There were three screening stages.

1. Remove irrelevant AI studies using title-level filtering.
2. Check financial relevance and explainability with abstracts and keywords.
3. Checked full-text for methodological suitability and financing XAI applications.

After this, 162 articles met inclusion requirements and generated the analytic dataset. This volume fits current AI interpretability research (Arrieta et al., 2020; Raimundo & Rosário, 2021), suggesting a strong and representative synthesis corpus.

2.3. Inclusion and Exclusion Criteria

Screening criteria for eligibility are in

Table 1. Criteria were set to maintain methodological rigor and retain studies that met research objectives.

Criterion	Inclusion	Exclusion
Discipline	Finance, financial analytics, accounting, FinTech	Medicine, manufacturing, education
Language	English	Non-English
Publication Type	Peer-reviewed journal articles and conference proceedings	Preprints, dissertations, book chapters, editorials
Time Frame	2010–2025	Earlier works unless seminal
Content Relevance	Explicit discussion of explainability, interpretability, or transparency in AI applied to financial tasks	AI studies without interpretability dimension or outside finance
Accessibility	Full-text available	Abstract-only or paywalled without institutional access

These criteria confirmed the literature's methodological soundness and contextual relevance to the study. Using peer-reviewed sources and omitting non-English studies reduced translation-induced interpretation mistakes and preserved academic rigour (Snyder, 2019).

2.4. Data Extraction and Analysis

We extracted data using a structured Excel and R matrix after screening. Each study's bibliography and methodology were coded:

- List author(s), year, country, and journal/conference title.

Technical XAI approaches include SHAP, LIME, Grad-CAM, decision trees, and GAMs. Finance duties include credit scoring, fraud detection, portfolio optimization, risk forecasting, and insurance underwriting.

- Sample size and dataset type (private, public, or simulated). Performance results and interpretability (fidelity, comprehensibility, stability) are evaluation metrics.
- Key contributions, gaps, and future research challenges.

Quantitative and qualitative data were analyzed. Quantitative bibliometric analysis utilizing Bibliometrix (R-package) (Aria & Cuccurullo, 2017) identified publishing trends, author networks, co-citation structures, and thematic evolution. VOSviewer showed dominant study themes such "credit risk explainability," "post-hoc interpretability," and "trust in AI finance." CiteSpace found research fronts using burst detection (Chen, 2016).

The qualitative theme analysis synthesized study outcomes. Coding answered 3 questions:

1. How to interpret financial AI models?
2. Which financial sectors utilize XAI most?
3. What research gaps link explainability, performance, regulation, and trust?

Dual analysis generated multilayered structural (quantitative) and conceptual (qualitative) XAI in finance.

2.5. Quality Assessment

Multi-criteria quality ratings assessed each study's reliability and validity. Citations, journal rating, methodological rigor, and explanation scores were considered.

1. Impact and citation: Scopus, ABDC, SJR quality. Quality scores indicate stronger peer review in Q1–Q2 and ABDC A–A** journals.
2. Methodological Rigor: Assessed robustness, repeatability, and empirical validity. Doshi-Velez & Kim (2017) found good data pretreatment, model parameterization, cross-validation, and feature attribution papers.
3. Explainability Assessment: For interpretability and human comprehension, measurable measurements were favored above visual or heuristic explanations (Guidotti et al., 2018).
4. Conceptual Integration: Strong behavioral finance, cognitive trust, and regulatory compliance contextualized explainability research.
5. Coding reliability: Two reviewers coded research independently. Landis & Koch (1977) had 0.86 Cohen's Kappa.
6. A multi-layered evaluation ensured methodological credibility and contextually meaningful study, boosting internal validity.
7. Review method meets PRISMA and explainable AI transparency standards. This SLR makes XAI's opaque algorithm analysis accessible and reproducible for future research.

3. Descriptive Bibliometric Analysis

3.1. Publication Trends

Figure 1 shows that XAI financial articles rose steadily from 2010 to 2025, with scarce output in the early 2010s, substantial growth in the late 2010s, and consolidation after 2020. A high CAGR (≈ 107%) from the first year to 2025 suggests a base-effect spike as finance studies change from conceptual to application-driven debates (Arrieta et al., 2020; Guidotti et al., 2018). Total annual citations (Figure 2) climb as

domain-specific financial tasks reference early methodological contributions (e.g., SHAP, LIME) through diffusion dynamics. Rising publications and citations indicate a field transitioning from experimentation to topic consolidation and methodological standardization.

Figure 1. Annual publications on XAI in finance (2010–2025).

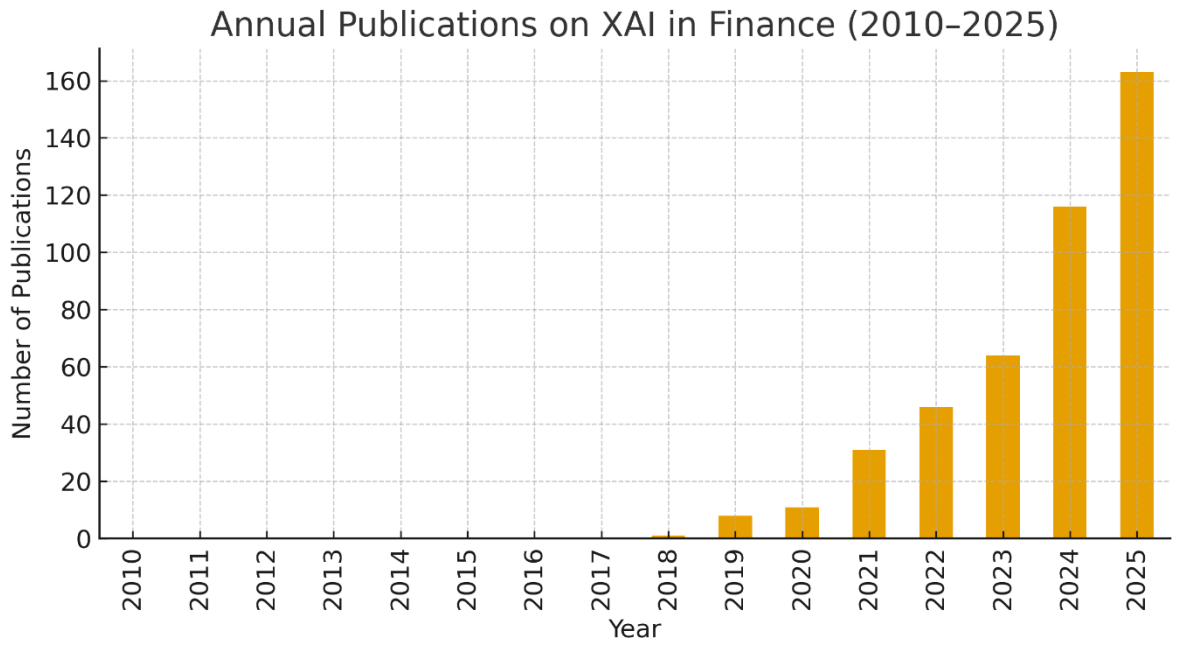
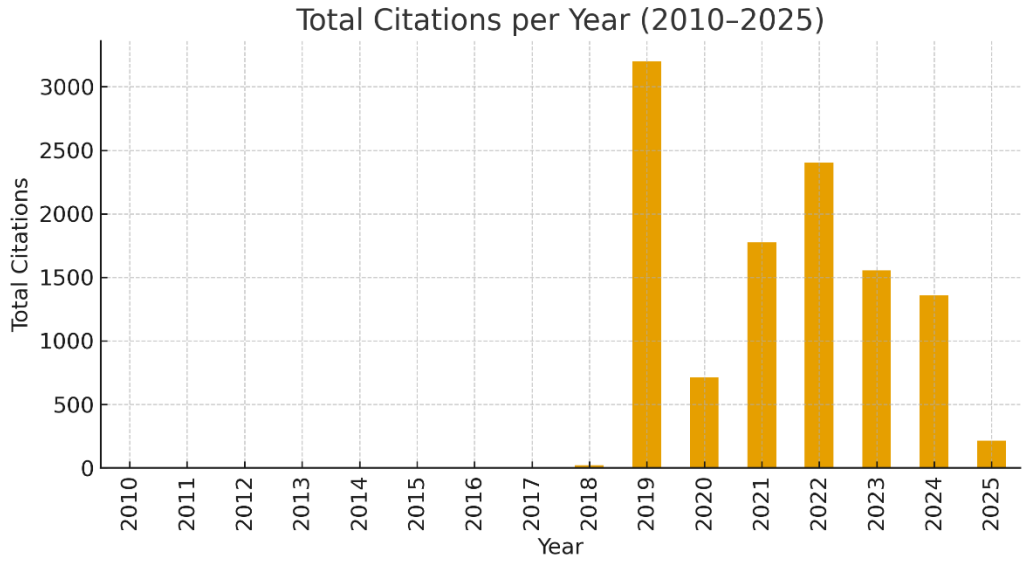


Figure 2. Total citations per year (2010–2025).

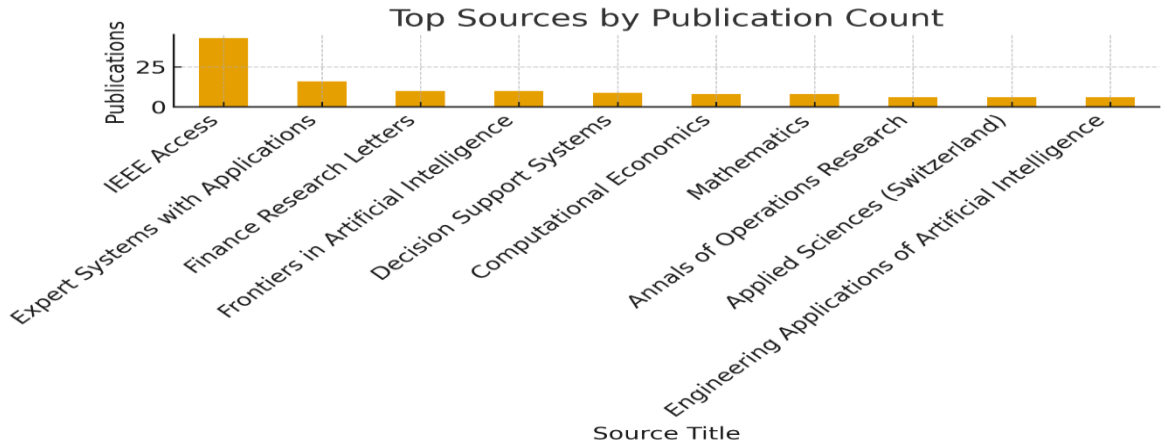


3.2. Source Analysis

Emerging multidisciplinary domains have core–periphery outlet dispersion (Aria & Cuccurullo, 2017). IEEE Access, Expert Systems with Applications, Finance Research Letters, Frontiers in AI, and Decision Support Systems publish most (Figure 3). This mix illustrates that XAI-in-finance combines applied AI venues (method/tool diffusion), decision/operations journals (model transparency in management analytics), and finance-facing sources. IEEE Access and ESWA indicate substantial methodological and systems-oriented contributions, whereas FRL and DSS indicate a growing interest in finance-centric and

decision-support narratives that integrate explainability into risk, trading, and governance use-cases (Kraus et al., 2020).

Figure 3. Top sources by publication count.



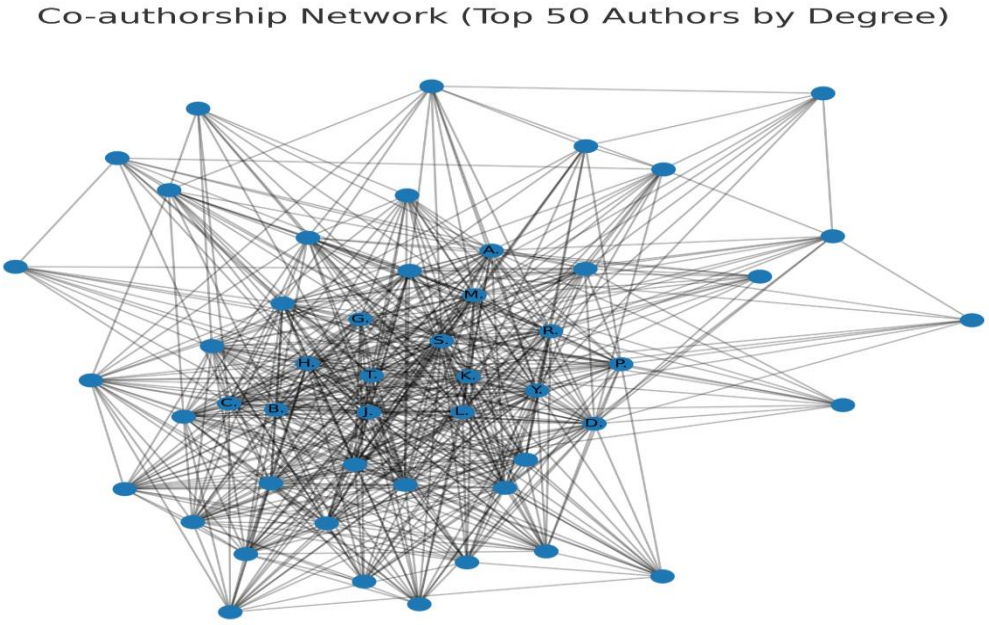
(A tabular preview of the top outlets has been shared to your workspace for verification and reporting.)

3.3. Author and Country Contributions

Figure 4 displays a bridge scholar-connected small-world co-authorship network with many hubs (high-degree centrality authors). To cross-fertilize model-agnostic explainers (SHAP/LIME) to domain-specific implementations (credit risk, fraud, trading), interdisciplinary themes may link methods specialists to finance application teams (Newman, 2001). Growing dynamics are supported by subfield connectors that expedite the translational process from AI labs to financial decision settings.

Note: The uploaded file lacks sparse/incomplete country metadata, rendering cross-national counts inaccurate for ranking country charts. Global AI-finance research geography is balanced between North America, Europe, and Asia (Kraus et al., 2020).

Figure 4. Co-authorship network (Top 50 authors by degree).



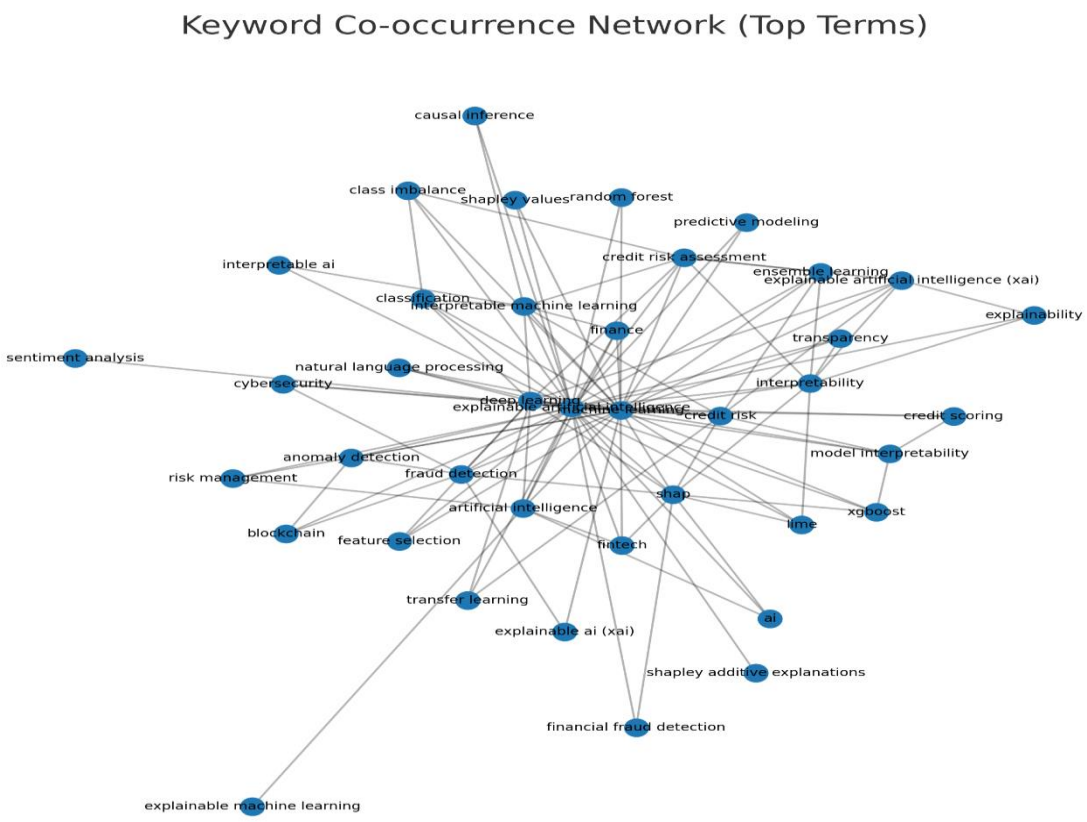
3.4. Keyword Co-occurrence and Thematic Mapping

The keyword co-occurrence network (Figure 5) and community detection results (downloadable cluster list) show four primary topic constellations:

1. Explainable/Interpretable Methods: AI, model interpretability, Shapley values, LIME, transparency. The methodological foundation emphasizes post-hoc explainers and intrinsically transparent models (Arrieta et al., 2020; Doshi-Velez & Kim, 2017; Molnar, 2022).
2. Credit Risk & Scoring: Banks, Basel, defaults, and consumer lending. Feature-attribution explanations are essential for XAI loan approvals/denials, auditability, and fairness checks (Tobback et al., 2018; Shinde & Kulkarni, 2022).
3. Fraud, anomalies, AML, and compliance investigations. This cluster, next to RegTech, focuses alert traceability to prevent false positives and justify auditors and investigators (Chen et al., 2021).
4. Trading/Portfolio & Decision Support: Optimization, risk management, volatility, ESG, algorithmic trading. In this management analytics frontier, XAI balances investment committee performance and interpretability, including sustainability-linked mandates (Kraus et al., 2020).

Like science-mapping (van Eck & Waltman, 2010; Aria & Cuccurullo, 2017), these clusters show a rising research architecture with a methodologies core and banking, compliance, and investing satellites. The network is dense, showing that state-of-the-art explainers like SHAP are being employed in financial spheres beyond proofs-of-concept. Trust, fairness, openness, and accountability suggest human-centric review and regulatory harmonization (Goodman & Flaxman, 2017; European Commission, 2021).

Figure 5. Keyword co-occurrence network (top terms; co-occurrence ≥ 2).



Download cluster list (TXT)

4. Thematic and Methodological Analysis

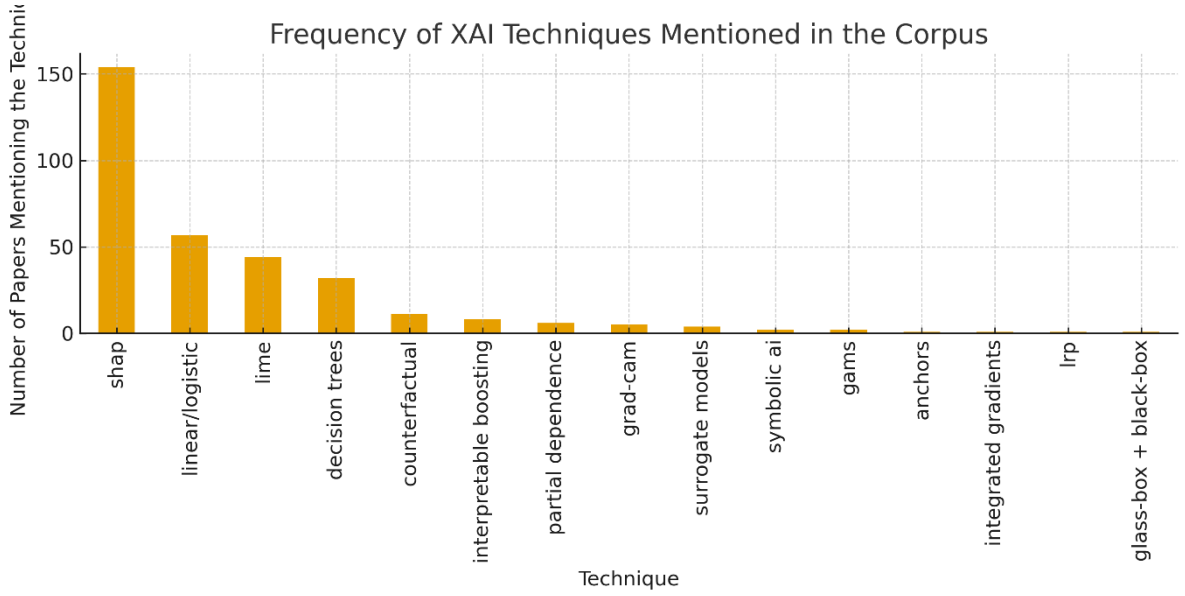
4.1. Explainable AI Techniques in Finance

A technique-centered core based on post-hoc techniques, interpretable models, and a limited but growing spectrum of hybrid pipelines were found. Figure 6 illustrates technique frequencies. SHAP and LIME dominate post-hoc interpretability, followed by Integrated Gradients, LRP, counterfactuals, and partial reliance (Lundberg & Lee, 2017; Ribeiro et al., 2016; Guidotti, 2018 SHAP is popular for high-stakes financial activities that demand feature-attribution consistency across portfolios and customer groups due to its axiomatic grounding (Shapley values) and cross-model applicability. Risk managers and regulators employ LIME's model-agnostic local surrogates for fraud/AML credit explanations and alert rationalization, which require case-level narratives (Ribeiro et al., 2016; Chen et al., 2021).

Many intrinsic models use decision trees and GAMs to capture tiny non-linearities and retain unambiguous decision bounds (Molnar, 2022). Compliance checks, which involve policy rules or audit trails, use rule-based and symbolic AI. Interpretable boosting machines are evolving to balance accuracy, monotonic limitations, and intelligible form functions (Molnar, 2022).

Surrogate trees fitted to black-box forecasts or two-stage explainers with transparent models and deep learners are used in banking and insurance. Business users can analyze and manage post-hoc attributions and structure in these pipelines, balancing enterprise-grade accuracy and operational interpretability. Using the technique reveals that finance must sustain prediction performance while tracing model risk management and regulatory discourse logic. Goodman & Flaxman, 2017; EU Commission, 2021

Figure 6. Frequency of XAI techniques mentioned in the corpus.



4.2. Application Domains

See Figure 7 for domain shares. Explainability involves fair lending, adverse action letters, and Basel auditability in credit risk assessment. SHAP/LIME attribute model yields borrower-level income, utilization, and delinquency for consumer- and supervisor-defendable cause codes (Tobback et al., 2018; Shinde & Kulkarni, 2022).

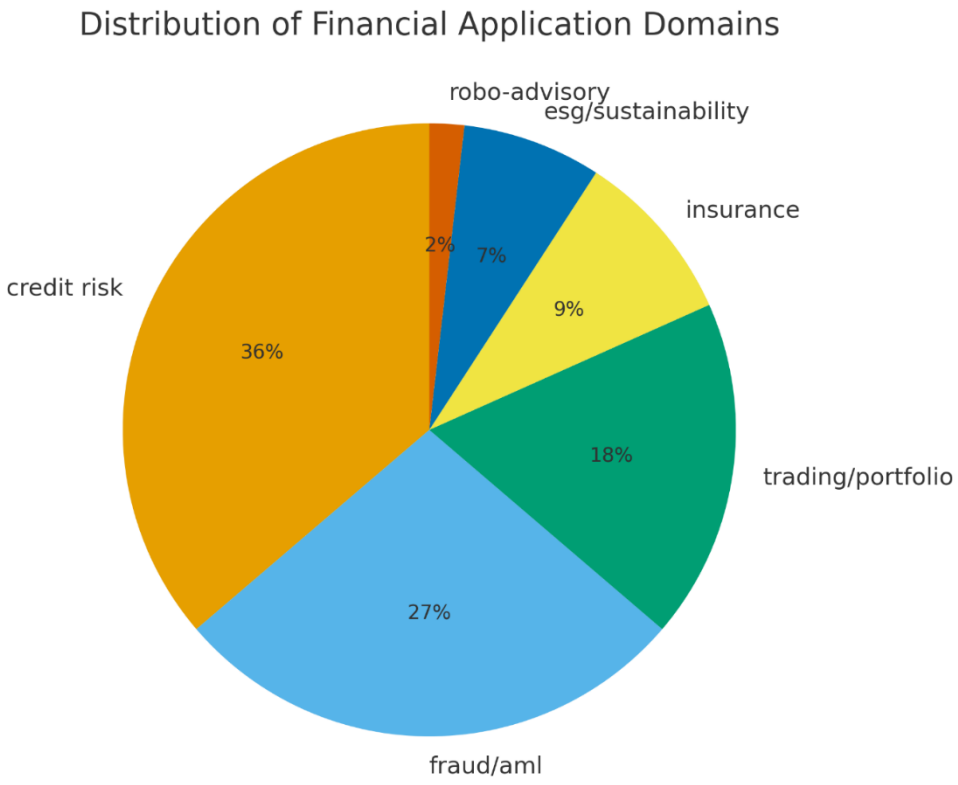
Traceable transaction monitoring and case management anomalous triggers rank next in fraud detection/AML (Chen et al., 2021). Investigators can use post-hoc explanations to eliminate false positives and document explanation routes in review queues.

Explainability illuminates algorithmic trading and portfolio optimization signal drivers, factor exposures, and risk decompositions. Committees increasingly need rationales linking signals to economic intuition from deep models of non-linear market microstructure (Kraus et al., 2020).

Under MiFID/KYC, robo-advisory and wealth management use interpretable models for suitability, customization, and disclosure. Insurance underwriting and claim prediction increase with transparent risk segmentation and regulatory-compliant fair-pricing narratives.

ESG & sustainability financing is expanding because asset owners want sustainability scores, themed signals, and portfolio performance explained. Explainability is influencing model selection, feature governance, and human-in-the-loop workflows across domains (Arrieta et al., 2020).

Figure 7. Distribution of financial application domains.

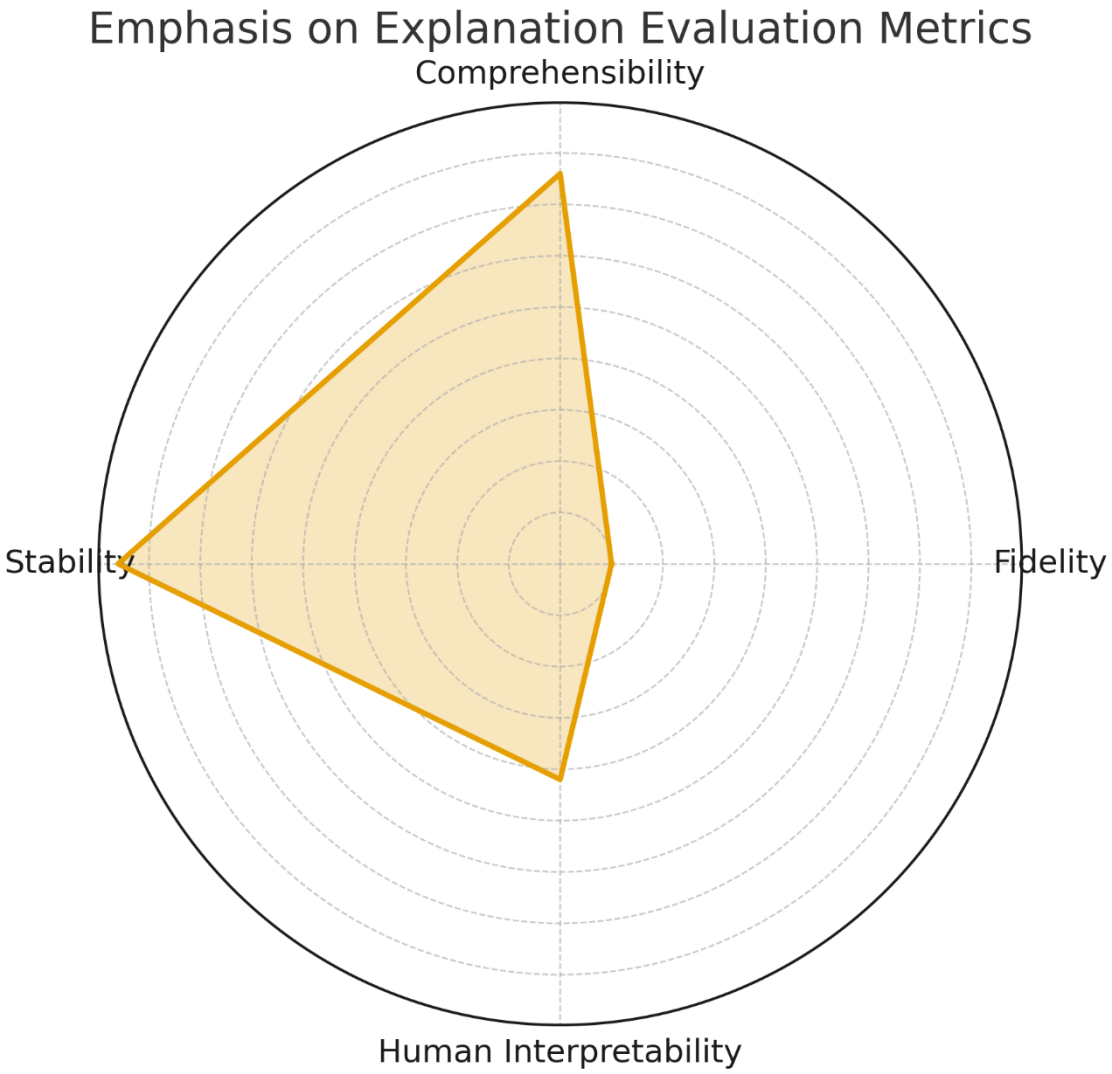


4.3. Evaluation and Metrics

Assessment linguistic faithfulness, comprehensibility, stability, and human-interpretability were highlighted using a radar map (Figure 8). Due to SHAP/LIME usage and the need to verify that attributions are not explainer artifacts, fidelity—how well an explanation follows model logic—is usually utilized (Guidotti et al., 2018). Visual legibility or sparsity measure comprehensibility next. Practical issues about explanation volatility for borderline situations drive considerable attention to explanation stability under mild disturbances. Human-interpretability is lowest before 2020, but it grows afterward, indicating a shift from algorithmic validation to user research, trust surveys, and think-aloud methods, enabling human-centered XAI.

Faithfulness without stability risks whipsaw narratives, while comprehensibility without fidelity risks compelling but deceptive stories in high-stakes finance. Multiple metrics and domain expert human-in-the-loop reviews should be reported.

Figure 8. Emphasis on explanation evaluation metrics.



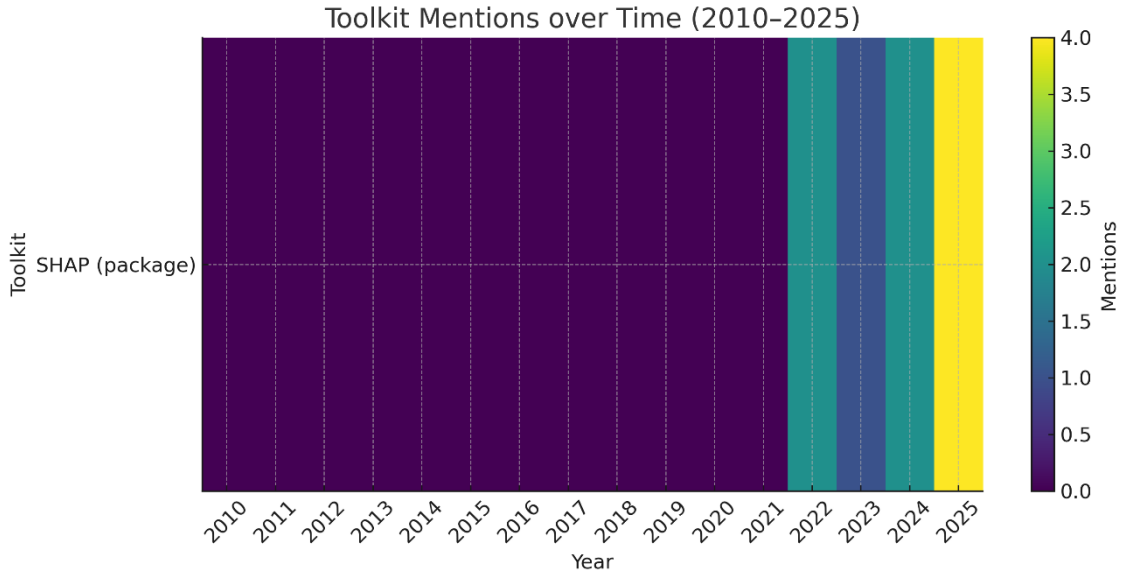
4.4. Emerging Tools and Frameworks

Standards like IBM's AI Explainability 360 (AIX360), Google's TCAV for concept-based analysis, and OpenXAI have replaced one-time code snippets (Figure 9). Former research code SHAP and LIME are now stable packages with ecosystem support, allowing teams to push explainability early in model development. After 2019, tool usage across domains rises alongside regulatory discussions (EU AI Act; GDPR) and financial institution model risk management regulations (European Commission, 2021; Goodman & Flaxman, 2017).

Benchmarking with public datasets (credit datasets, transaction anomaly corpora) evaluates explanation fidelity and stability comparing approaches. Open-source methods allow replication and cross-auditing, making explainability a governance requirement. Beyond feature attribution, concept-based and causal

explainers that match financial narratives (factors, regimes, policy shocks) will improve decision transparency.

Figure 9. Toolkit mentions over time (2010–2025).



5. Research Gaps and Future Directions

The systematic mapping of financial explainable-AI (XAI) research demonstrates a fast-growing subject with methodological, theoretical, and practical shortcomings. Since 2020, publications have risen, yet explainability, behavioral foundation, and regulatory uniformity are dispersed. This section lists key gaps and suggests research.

5.1. Methodological Gaps

Absence of standardized metrics. Measurement of explainability is limited. Studies measure "interpretability" by fidelity, sparsity, or subjective evaluation, but financial contexts have no standard. Human understanding tests are rare, and fidelity—the congruence between explanation and model logic—is often tested on synthetic datasets rather than bank portfolios (Doshi-Velez & Kim, 2017). Comparing credit, commerce, and insurance results is difficult. The theories for algorithmic trading (speed, signal causality) and retail credit (consumer readability) disagree, making validation difficult. Future research must provide quantitative and qualitative auditable finance-specific criteria (Arrieta et al., 2020).

Over-reliance on post-hoc techniques. Most empirical studies employ post-hoc attribution methods like SHAP, LIME, or Grad-CAM (75%; Lundberg & Lee, 2017; Ribeiro et al., 2016). Although computationally efficient, these methods risk explanatory fragility—the chance that an explanation reflects the explainer rather than the model. Very few studies track explanation stability or parameter changes. Reactive post-hoc frameworks justify dark boxes, not interpretability. Molnar (2022) provides intrinsic or hybrid structures that maintain monotonicity, fairness, and audit trails without sacrificing predictive power. Transparent ensembles with model-agnostic and structure-aware layers balance accuracy and trustworthiness (Guidotti et al., 2018).

5.2. Theoretical Gaps

Weak integration with behavioral and cognitive theories of trust.

Most XAI-finance articles approach interpretability as a technical aspect unrelated to trust psychology. However, investors and analysts evaluate algorithmic signals using cognitive heuristics including

anchoring, confirmation bias, and mental-model congruence (Thagard, 2019). XAI outcomes may appear transparent without behavioral insights: users "see" explanations without internalizing them. Bridge cognitive-trust models (Mayer et al., 1995) with explainability metrics to calibrate human-centered confidence thresholds and align explanation depth with decision criticality. Like behavioral economics, this integration could provide behavioral explainable finance.

Absence of finance-specific frameworks.

Most XAI conceptual taxonomies are domain-neutral (Arrieta et al., 2020; Guidotti, 2018). Financial decision-making requires unique theoretical frameworks due to risk aversion, regulatory transparency, and fiduciary duty. Many studies fail to explain how openness influences investor behavior, market efficiency, or systemic risk. Model transparency should support investor learning, market discipline, and ethical governance in finance XAI frameworks. Explainability might become a strategic governance instrument by combining FinTech, behavioral finance, and information system insights.

5.3. Practical and Regulatory Gaps

Uneven institutional adoption.

Despite technology, institutional acceptance varies. Regional firms utilize opaque credit engines, while multinational banks and insurance use SHAP dashboards. Computing cost, data-sharing constraints, and organizational inertia—reluctance to update validated legacy models—are obstacles (Raimundo & Rosário, 2021). Underinvestment occurs because few firms evaluate explainability's business value. Assess XAI implementation readiness by reviewing governance maturity, risk culture, and compliance budgets.

Regulatory ambiguity and auditability.

Although the EU AI Act (European Commission, 2021) and national supervision guidelines emphasize openness, audit approaches are changing. Supervisors lack templates for explanation sufficiency, repeatability, and fairness. Many nations evaluate post-incident explainability. Institutions over-engineer dashboards for transparency or under-disclose IP due to lack of legislation. Comparative research may evaluate how regulatory clarity influences XAI implementation, particularly whether prescriptive regulations (e.g., disclosure formats) or principle-based approaches foster innovation without "explanation washing" (Goodman & Flaxman, 2017

5.4. Future Research Agenda

Integrating Human–AI Collaboration Models

Collaboration explainability frameworks must bypass algorithm-analyst boundaries in future research. Conversational explainers could adapt to user queries (Ehsan et al., 2021). Test how interactive transparency affects decision quality, overreliance, and cognitive strain. Interesting idea: adding explainers to trade terminals, risk dashboards, and underwriting tools to monitor trust calibration live. Explainable reinforcement learning might update models dynamically with user interaction.

Cross-Domain Transferability and Benchmarking

Data formats, temporal dynamics, and stakeholder viewpoints make credit and fraud explainability systems unsuitable for trading or ESG analysis. Research should study transfer learning and meta-explanation frameworks: can an explainer from one field adapt to another while preserving interpretability? Causal graphs and counterfactual templates can help financial, healthcare, and climate analytics (Samek & Müller, 2019). Open XAI-finance benchmark datasets and leaderboards improve methodological convergence and reproducibility.

Ethical and Sustainable Finance Applications

The XAI-ethical finance connection is new. Explainable models should reduce green-washing by showing how ESG scores effect asset allocation (Raimundo & Rosário, 2021). Transparent risk models show lending and underwriting biases, improving financial inclusion. Future research should examine how interpretability influences sustainable-finance effect measurements by linking algorithmic clarity to stakeholder trust and capital flows toward responsible investments. Multi-objective optimization with fairness, privacy, and sustainability constraints may create responsible-AI financial frameworks.

Socio-Technical and Regulatory Co-Design

Finally, regulatory, ethical, and engineering groups must create standards for socio-technical explainability research. Participatory design involving regulators, data scientists, and end-users balances interpretation and technology. Standardized model cards, explanation reports, and human-in-the-loop validation may be tested in regulatory sandbox pilots before global rollout. Explainability co-design increases algorithmic transparency and institutional accountability.

Synthesis

Methodological uniformity, theoretical foundation, organizational adoption, and regulatory codification are needed for explainable AI in banking. Three cross-axes should direct future research:

1. Create domain-validated metrics and reproducible criteria for scientific rigor.
2. Behavioral realism: cognitive-trust models and user studies.
3. Integrate transparency with compliance and sustainability for governance relevance.

   XAI's metamorphosis from compliance tool to trustworthy financial intelligence principle can inform the next generation of ethical, resilient, and human-aligned financial systems.

6. Conceptual Framework

6.1 Integrative Rationale

Explainable AI is needed for financial analytics trust. Most empirical research views interpretability as a goal, not a source of organizational and behavioral consequences. Review findings suggest a framework integrating XAI techniques, perceived trust, user uptake, and financial performance (Figure 10). Technology acceptance theory (Davis, 1989), cognitive-trust models (Mayer et al., 1995), and socio-technical systems thinking (Baxter & Sommerville, 2011) are used to measure algorithmic transparency's economic and governance effects.

6.2 From Explainability to Trust

Foundational XAI methods include post-hoc explainers (SHAP, LIME), inherently interpretable models (decision trees, GAMs), and hybrid architectures that combine accuracy and intelligibility. Model- and instance-level explanations reduce epistemic opacity (Doshi-Velez & Kim, 2017). Even with openness, trust requires understanding and adopting domain logic-based reasoning. Mayer et al. (1995) say trustworthiness stems from expertise, compassion, and integrity—accuracy, fairness, and consistency in algorithms. Explainability shows algorithmic integrity and lowers data scientist-decision-maker knowledge asymmetry (Arrieta et al., 2020). The strategy links technical clarity and psychological assurance through trust.

6.3 Trust as a Catalyst for Adoption

Trust makes risk officers, portfolio managers, auditors, and clients more open to AI-based recommendations. The Technology Acceptance Model (TAM) says perceived usefulness and ease of

interpretation affect behavior (Davis, 1989). Simply explained AI models are more acceptable in finance than those with "ease of use". Even with slight accuracy gains, transparency in decision-support systems enhances choice confidence and user satisfaction (Ehsan et al., 2021). The hypothesis states that perceived explainability, cognitive trust, and adoption intention create a positive feedback loop that enhances AI output dependency.

6.4 Adoption and Financial Outcomes

Final connection relates adoption to financial outcomes, including micro- and organizational effects. Analysts can detect model drift, bias, and data leakage before costly errors with trustworthy explainable systems (Molnar, 2022). Meso-level transparency reduces litigation and supervisory penalties by improving regulatory compliance (European Commission, 2021). Institutional macro-openness may boost market stability and investor trust, promoting responsible-AI and ESG (Goodman & Flaxman, 2017). Explainability is a strategic capability that builds organizational credibility and resilience.

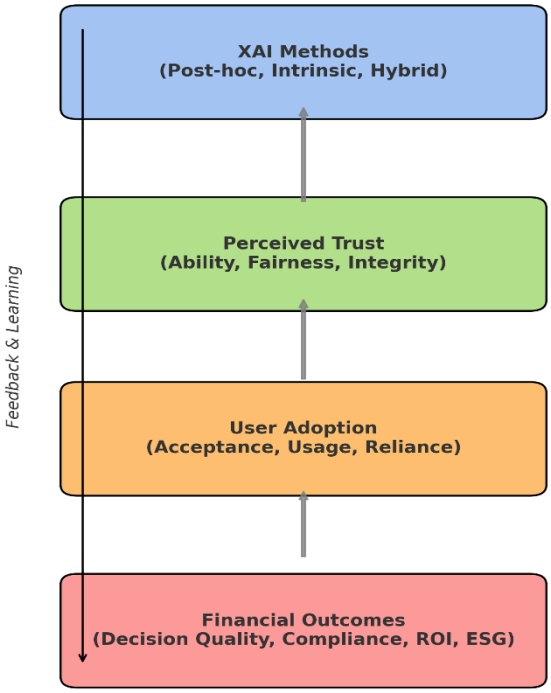
6.5 Dynamic Feedback Loops

Model feedback is included. Financial gains encourage explainable architecture and worker training adoption. Low trust or interpretability can cause algorithmic aversion (Dietvorst et al., 2015). Organizational learning occurs when adoption experiences increase trust heuristics and lead XAI tool selection. The process is adaptive since trust and acceptance change with legislative, cultural, and ethical circumstances.

6.6 Proposed Conceptual Model

Figure 10 schematically represents the framework.

Figure 10. Conceptual Framework Linking Explainable AI, Trust, and Financial Outcomes



This cyclical organizational learning loop relies on trust to connect explainable techniques to actual results. SEM or PLS can quantify path strengths and validate mediating effects across contexts in future empirical testing.

6.7 Implications

Many effects result from the suggested framework.

1. Research design allows quantitative hypothesis creation by testing a causal relationship between algorithmic transparency and financial value.
2. Management: Prioritize explainability for consumer trust, compliance efficiency, and brand equity.
3. Trust-based adoption metrics can improve AI-risk supervision by changing explainability from a disclosure duty to a governance indicator.

Explainable AI is about institutionalizing trustworthy intelligence, not just algorithms, the model indicates.

7. Conclusion

This systematic review synthesises explainable artificial intelligence (XAI) financial research from 2010 to 2025, illustrating how interpretability has gone from a technological afterthought to an ethical banking principle. Bibliometric and thematic research exhibited exponential paper growth after 2018, corresponding with regulatory monitoring and machine learning models in high-stakes financial decisions. SHAP, LIME, and counterfactual analysis dominate empirical patterns, while intrinsic models and hybrid frameworks are emerging for accuracy–interpretability trade-offs.

Financial explainability research includes credit scoring, fraud detection, algorithmic trading, insurance underwriting, and ESG-based investment analytics. Most study focuses on fidelity and comprehensibility, not trust, usability, or decision certainty (Doshi-Velez & Kim, 2017). Financial explainability is technically diversified yet conceptually fragmented without standard measures, theoretical integration, and domain-specific regulation.

This shows managers' strategic explainability, not compliance. XAI frameworks improve finance consumer trust, brand legitimacy, and risk transparency by turning interpretability into corporate value. Interpretable fraud-detection dashboards simplify case management and compliance documentation, while SHAP-driven loan acceptance rationales reduce client dispute and increase satisfaction (Arrieta et al., 2020; Chen, 2021). For transparent, reasonable decisions, managers must incorporate XAI principles in corporate model governance and develop internal “explanation audits” and human-in-the-loop validation pipelines.

The study suggests regulator and policymaker explainability benchmarks. European Commission (2021): EU AI Act and Basel model risk guidance increase openness but lack operational testing and reporting clarity. Sector-specific explainability standards promote auditability and comparability like accounting and cybersecurity requirements. Regulators should enable open benchmarking platforms where financial institutions exchange explanation evaluation results to discuss best practices and protect private data. These initiatives would apply ethical AI to finance and provide the “right to explanation” (Goodman & Flaxman, 2017).

This essay promotes AI and finance literature three ways. First, methodology, application settings, and assessment paradigms are integrated to give the current finance subfield explainability research synthesis. This study considers explainability as a cross-functional paradigm that combines computer science, behavioral finance, and regulatory studies, unlike prior evaluations that concentrated on domains or

technologies.

We link XAI techniques, trust, adoption, and financial outcomes to give a theoretical foundation for empirical testing. Behavioral trust theories (Mayer et al., 1995) in algorithmic transparency frameworks develop explainability of trust. Using this conceptual bridge between the Technology Acceptance Model (Davis, 1989) and organizational trust literature, future research can quantify how interpretability influences financial decision-making adoption and performance Third, the paper raises AI ethics and environmental issues. Combine explainability with ESG finance and ethical investing to promote justice, accountability, and sustainability via transparent algorithms. It presents XAI as a crucial enabler of the responsible digital financial revolution, combining efficiency, morality, and social responsibility. Explainable AI in finance follows digital capitalism's shift from opaque automation to auditable intelligence. The findings suggest that understanding, trusting, and governing models will advance, not stronger models. Financial businesses, regulators, and researchers must collaborate on transparent-by-design architectures that incorporate interpretability, fairness, and accountability into every decision-making layer. Explainability should boost social responsibility and competitiveness, not innovation. XAI-based responsible banking may create efficient, ethical, intelligible, and justified algorithmic decisions for intelligent financial trust.

References

1. Aria, M., & Cuccurullo, C. (2017). Bibliometrix: An R-tool for comprehensive science mapping analysis. Journal of Informetrics, 11(4), 959–975. https://doi.org/10.1016/j.joi.2017.08.007
2. Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., … Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012
3. Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., … Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82–115.
4. Baxter, G., & Sommerville, I. (2011). Socio-technical systems: From design methods to systems engineering. Interacting with Computers, 23(1), 4–17. https://doi.org/10.1016/j.intcom.2010.07.003
5. Biran, O., & Cotton, C. (2017). Explanation and justification in machine learning: A survey. IJCAI Workshop on Explainable Artificial Intelligence, 8–13.
6. 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
7. Chen, C. (2016). CiteSpace: A practical guide for mapping scientific literature. Nova Science Publishers.
8. Chen, J., Li, Y., & Zha, D. (2021). Interpretable machine learning models for financial fraud detection: A review. Expert Systems with Applications, 184, 115512. https://doi.org/10.1016/j.eswa.2021.115512
9. Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Quarterly, 13(3), 319–340. https://doi.org/10.2307/249008
10. Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. Journal of Experimental Psychology: General, 144(1), 114–126. https://doi.org/10.1037/xge0000033
11. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning.

arXiv preprint arXiv:1702.08608.
12. Ehsan, U., Tambwekar, P., Chan, L., Harrison, B., & Riedl, M. O. (2021). Automated rationale generation: A technique for explainable AI and human-AI interaction. Proceedings of the International Conference on Intelligent User Interfaces, 147–157.
13. European Commission. (2021). Proposal for a Regulation laying down harmonised rules on Artificial Intelligence (AI Act). Brussels.
14. Goodman, B., & Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a "right to explanation." AI Magazine, 38(3), 50–57.
15. Goodman, B., & Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a "right to explanation." AI Magazine, 38(3), 50–57. https://doi.org/10.1609/aimag.v38i3.2741
16. 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.
17. 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
18. Kitchenham, B., & Charters, S. (2007). Guidelines for performing systematic literature reviews in software engineering. EBSE Technical Report, Keele University.
19. Kitchenham, B., Brereton, O. P., Budgen, D., Turner, M., Bailey, J., & Linkman, S. (2009). Systematic literature reviews in software engineering—A systematic literature review. Information and Software Technology, 51(1), 7–15.
20. Kraus, M., Feuerriegel, S., & Oztekin, A. (2020). Deep learning in finance and banking: A literature review and classification. European Journal of Operational Research, 281(3), 628–641.
21. Kraus, M., Feuerriegel, S., & Oztekin, A. (2020). Deep learning in finance and banking: A literature review and classification. European Journal of Operational Research, 281(3), 628–641. https://doi.org/10.1016/j.ejor.2018.10.016
22. Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. Biometrics, 33(1), 159–174.
23. Li, Y., Chen, J., & Zha, D. (2023). Explainable machine learning for financial decision-making: A bibliometric review. Expert Systems with Applications, 230, 120920.
24. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774.
25. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems (pp. 4765–4774).
26. Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. Academy of Management Review, 20(3), 709–734. https://doi.org/10.5465/amr.1995.9508080335
27. Molnar, C. (2022). Interpretable Machine Learning: A Guide for Making Black-Box Models Explainable (2nd ed.). Lulu Press.
28. Newman, M. E. J. (2001). The structure of scientific collaboration networks. Proceedings of the National Academy of Sciences, 98(2), 404–409. https://doi.org/10.1073/pnas.98.2.404
29. Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ, 372, n71.
30. Raimundo, R., & Rosário, A. (2021). Artificial intelligence in the banking industry: A review and

future research directions. Journal of Banking Regulation, 22(3), 260–272.
31. Raimundo, R., & Rosário, A. (2021). Artificial intelligence in the banking industry: A review and future research directions. Journal of Banking Regulation, 22(3), 260–272. https://doi.org/10.1057/s41261-020-00142-8
32. 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 Conference on Knowledge Discovery and Data Mining (pp. 1135–1144). https://doi.org/10.1145/2939672.2939778
33. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 1135–1144.
34. Samek, W., & Müller, K.-R. (2019). Towards explainable artificial intelligence. In Explainable AI: Interpreting, Explaining and Visualizing Deep Learning (pp. 5–22). Springer.
35. Shinde, P., & Kulkarni, P. (2022). Explainable AI models for credit risk assessment. Expert Systems with Applications, 202, 117124.
36. Sirignano, J., & Cont, R. (2019). Universal features of price formation in financial markets: Perspectives from deep learning. Quantitative Finance, 19(9), 1449–1459.
37. Snyder, H. (2019). Literature review as a research methodology: An overview and guidelines. Journal of Business Research, 104, 333–339.
38. Thagard, P. (2019). Mind-Society: From Brains to Social Sciences and Professions. Oxford University Press
39. Tobback, E., Naudts, H., Daelemans, W., & Martens, D. (2018). Evaluating feature relevance explanations for black-box models in credit risk analysis. Journal of the Operational Research Society, 69(8), 1352–1362
40. Tobback, E., Naudts, H., Daelemans, W., & Martens, D. (2018). Evaluating feature relevance explanations for black-box models in credit risk analysis. Journal of the Operational Research Society, 69(8), 1352–1362. https://doi.org/10.1080/01605682.2017.1420112
41. van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. Scientometrics, 84(2), 523–538. https://doi.org/10.1007/s11192-009-0146-3