RJIT Research Journal of Information Technology

OPEN ACCESS

Auditing AI in Finance: Risk, Fairness, Robustness and Compliance: A Systematic Review

[1]Isaiah John Otseje,[2]Ozioma Adaeze Chinonso,[3]David Chinonso Anih,[4]Ukeni Mgbechikwre Victoria,[4]Ehio Victor Chituru, [5]Ogbu Augustine Ogoh and [6]Omobolanle Omotayo Solaja

[1]Department of Accounting, Federal University of Health Sciences Teaching Hospital Otukpo, Benue State, Nigeria

[2]Department of Accounting, Faculty of Management Science, Enugu State University of Science and Technology, Nigeria

[3]Department of Biochemistry, Faculty of Biosciences, Federal University Wukari, Taraba, Nigeria

[4]Department of Accounting, Faculty of Management Sciences, University of Port Harcourt, Nigeria

[5]Department of Accounting, Faculty of Management Sciences, Enugu State University of Science and Technology, Enugu, Nigeria

[6]Department of Statistics, Federal University of Agriculture, Abeokuta, Ogun, Nigeria

ABSTRACT

This systematic review synthesizes the emergence and maturation of algorithmic auditing and assurance frameworks for AI-driven financial systems, integrating conceptual, technical, regulatory, and ethical literatures to produce a coherent approach for practice and governance. Using PRISMA and a structured search of major databases from January 2015 to October 2025, we screened 1,243 records and retained 40 peer-reviewed studies for detailed synthesis and appraisal. The review organizes findings across four core dimensions: risk assessment, fairness, robustness, and regulatory compliance. In the risk domain, the study contrasts traditional measures such as value at risk and expected shortfall with AI-aware formulations that incorporate model uncertainty, predictive miscalibration, distributional drift, and adversarial vulnerabilities. Hybrid approaches that combine statistical risk models are highlighted, with machine learning forecasts and recommended adversarial risk testing, comprehensive sensitivity analysis, and routine back testing to detect tail exposures and manipulation vectors. This study synthesizes operational metrics, including demographic parity and equalized odds, and highlights intersectional fairness to capture overlapping vulnerabilities that single-attribute measures may miss. The robustness analysis emphasizes adversarial training, systematic sensitivity testing, and hybrid scenario-based evaluations designed to probe both technical fragility and economic instability. For compliance and governance, established practices that embed explainability, comprehensive documentation, model cards, provenance tracking, and automated compliance checks into production pipelines are reviewed, with attention to alignment with Basel III, the GDPR, and other cross-jurisdictional regulations. Building on this synthesis, an integrated Assurance Index is proposed that aggregates normalized scores for risk, fairness, robustness, and compliance into a decision-ready composite with configurable weights that reflect institutional priorities. The framework pairs quantitative diagnostics with participatory auditing and stakeholder engagement to strengthen legitimacy and surface social concerns that technical metrics alone may overlook. The paper concludes with actionable research and policy recommendations, including benchmarking assurance practices, harmonizing metrics across jurisdictions, improving adversarial evaluation methods, incorporating industry experience into scholarly research, and supporting regulatory pilots to operationalize assurance mechanisms. Adoption of these measures can help ensure that AI-driven finance remains resilient, fair, transparent, and accountable. Future work should focus on open benchmarking datasets, collaborative audits, and practitioner-oriented toolkits for routine assurance deployment globally.

KEYWORDS

Algorithmic auditing, AI assurance, risk assessment, fairness and equity, robustness testing, explainability, regulatory compliance, assurance index, participatory auditing

ISSN: 2151-7959 (Online)
ISSN: 1815-7432 (Print)
https://doi.org/10.3923/rjit.2026.01.13

Received: 02 Jan. 2025
Accepted: 12 Mar. 2026
Published: 20 Mar. 2026

INTRODUCTION

Artificial intelligence has become deeply embedded in financial systems, reshaping how credit is assessed, fraud is detected, and investments are managed. This transformation has brought remarkable efficiency gains, but it has also introduced new risks that demand scrutiny. Algorithmic auditing and assurance frameworks have emerged as essential tools to evaluate whether these systems operate fairly, robustly, and in compliance with regulatory standards. The growing reliance on machine learning models in finance means that decisions once made by humans are now delegated to algorithms, raising questions about accountability and transparency[1].

Risk assessment in AI driven finance is particularly complex because models often rely on large, dynamic datasets that can shift with market volatility. Traditional risk metrics such as Value at Risk (VaR) and Expected Shortfall are being adapted to algorithmic contexts, but researchers emphasize that these measures must account for model uncertainty and adversarial vulnerabilities[2]. Fairness is another pressing concern. Studies show that biased training data can lead to discriminatory outcomes in lending and insurance, undermining trust in financial institutions[3]. To address this, fairness auditing frameworks now incorporate statistical parity and equalized odds, ensuring that protected groups are not systematically disadvantaged.

Robustness is equally critical. Financial AI systems must withstand adversarial attacks and unexpected market shocks. Scholars highlight that robustness testing should go beyond stress scenarios to include sensitivity analysis, where small perturbations in input data are examined for disproportionate effects on outputs[4]. Regulatory compliance adds another layer of complexity. Global financial regulators, including those enforcing Basel III and GDPR, increasingly demand explainability and accountability in algorithmic decision making. This has led to the development of interpretable compliance frameworks that align AI outputs with legal standards[5].

Beyond technical dimensions, ethical and societal implications cannot be ignored. Algorithmic assurance frameworks must consider stakeholder perspectives, balancing efficiency with fairness and public trust[6]. Recent systematic reviews suggest that integrated assurance models, combining risk, fairness, robustness, and compliance, provide a holistic approach to auditing financial AI systems[7]. These frameworks not only safeguard institutions against systemic risks but also strengthen confidence among regulators and the public.

This review aims to provide a rigorous, evidence-based synthesis of peer-reviewed research on algorithmic auditing and assurance within AI-driven financial systems. Its scope encompasses prevailing conceptual frameworks, quantitative metrics, and testing protocols, as well as governance arrangements, organized around four interrelated dimensions: Risk assessment, fairness, robustness, and regulatory compliance, and applied to core financial domains such as credit scoring, algorithmic trading, fraud detection, and insurance.

MATERIALS AND METHODS

The methodological foundation of this systematic review was carefully designed to ensure transparency, reproducibility, and rigor. Because algorithmic auditing in financial systems is a rapidly evolving field, we adopted a structured approach that combined established systematic review protocols with domain specific considerations. The Preferred Reporting Items for Systematic Reviews and Meta Analyses (PRISMA) framework guided the entire process, ensuring that the selection of studies was both comprehensive and unbiased[8].

Search strategy: Relevant electronic databases indexing peer-reviewed literature in finance, computer science, and ethics were systematically identified. These included Scopus, Web of Science, IEEE Xplore, and ScienceDirect. The search terms were carefully constructed to capture the intersection of AI, auditing,

assurance, and financial systems. Keywords such as "algorithmic auditing", "AI assurance", "financial risk assessment", "fairness in financial AI", "robustness testing", and "regulatory compliance" were combined using Boolean operators. To ensure currency, the search was restricted to publications between January 2015 and October 2025[9].

The initial search yielded 1,243 articles. After removing duplicates, 1,015 unique records remained. Titles and abstracts were screened against the inclusion criteria:

- Peer reviewed journal articles
- Focus on AI in financial contexts
- Explicit discussion of auditing, assurance, or compliance frameworks
- Publication within the specified timeframe

Exclusion criteria included conference proceedings, non peer reviewed reports, and articles focusing solely on technical AI optimization without financial application[10].

Study selection

Identification

Records identified through database searching: n = 1,243

Screening

Duplicates removed: n = 228
Records after duplicate removal/records screened: n = 1,015
Records screened (titles/abstracts): n = 1,015
Records excluded after title/abstract screening: n = 803

Eligibility

Full-text articles assessed for eligibility: n = 212
Full-text articles excluded: n = 172

Common reasons for full-text exclusion (aggregate)

Not focused on algorithmic auditing and assurance for AI-driven financial systems
Lacked frameworks for risk assessment, fairness, robustness, or regulatory compliance
Wrong study design (commentary, opinion piece, editorial)
Insufficient or irrelevant data to answer the inclusion criteria
Not in English/full text not available
Duplicate reports/preliminary abstracts

Included

Studies meeting inclusion criteria: n = 40
Included in main review: n = 24
Included as background/methodological references: n = 16

The diagram's numbered boxes and arrows visually map each decision point so readers can verify the screening workflow described in Fig. 1[11].

Flow diagram Fig. 1 of study selection: The 1,243 records identified through database searching; after removing 228 duplicates, 1,015 records were screened (titles/abstracts), 803 were excluded, 212 full-text articles were assessed, 172 were excluded, and 40 studies met the inclusion criteria. Of the 40 included studies, 24 were included in the main review and 16 were retained as background/methodological references. Abbreviation: n = Number of records.

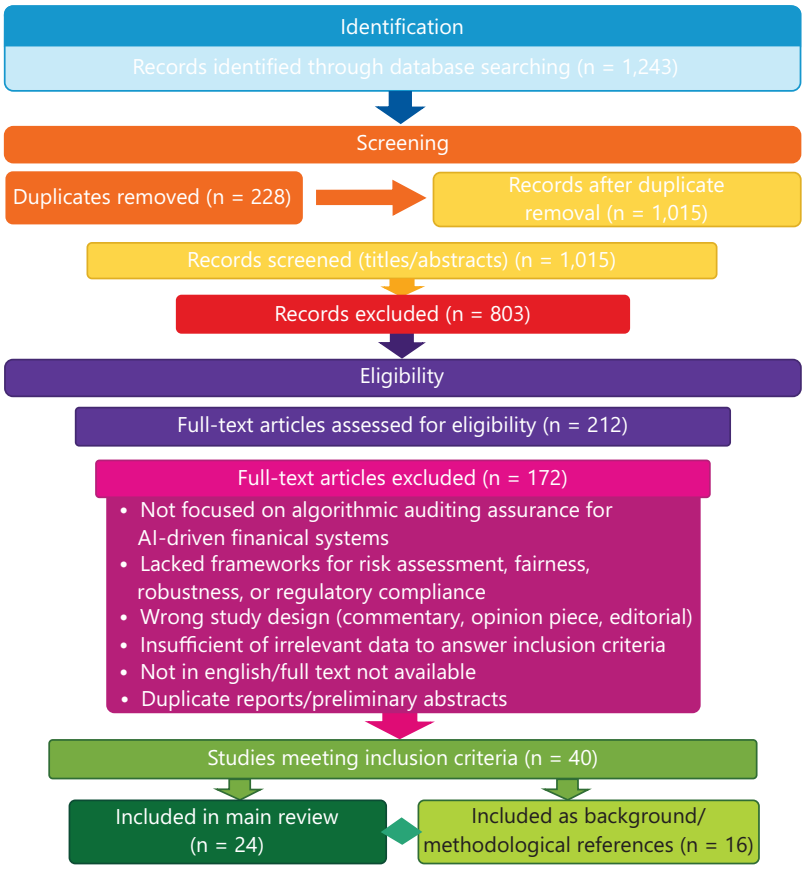

Fig. 1: PRISMA flow diagram of study identification, screening, eligibility, and final inclusion[11]

Data extraction: For each included study, key information was extracted using a standardized template. The template captured:

- Author(s), year, and journal
- Study objectives
- AI techniques applied (e.g., neural networks, decision trees, reinforcement learning)
- Financial domain (e.g., credit scoring, fraud detection, trading systems)
- Auditing or assurance framework discussed
- Metrics for risk, fairness, robustness, and compliance
- Main findings and limitations

This structured extraction ensured that data could be systematically compared across studies. To minimize bias, extraction was performed independently by two reviewers, with discrepancies resolved through consensus[12].

Quality assessment: Quality assessment was critical to ensure that only robust studies informed the synthesis.

A modified Critical Appraisal Skills Programme (CASP) checklist adapted specifically for AI auditing research was employed to assess the methodological quality, transparency, and rigor of the included studies. Criteria included clarity of objectives, appropriateness of methodology, transparency of data, and relevance to financial systems. Each study was scored on a scale from 0 to 10, with a threshold of 6 for inclusion. Studies scoring below this threshold were excluded from synthesis but noted in supplementary materials for completeness[13].

Table 1: Overview of methodological steps and citations

Step	Description	Citation(s)
Search strategy	Databases searched, keywords used, timeframe (2015-2025)	Marshall and Wallace[9]
Study selection	Two stage screening, inclusion/exclusion criteria	Ioannids[10] and Page *et al.*[11]
PRISMA flow diagram	Visual representation of selection process	Rethlefsen *et al.*[8]
Data extraction	Standardized template, independent reviewers	Shea *et al.*[12]
Quality assessment	Modified CASP checklist, scoring threshold	Leocádio *et al.*[13]
Analytical framework	Mapping onto risk, fairness, robustness, compliance	Mertens[14]
Integration of evidence	Mixed methods synthesis of quantitative and qualitative studies	Maleki and Karami[15]
Limitations	Language restriction, exclusion of industry reports, rapid evolution of AI	Lessmann *et al.*[16]

PRISMA: Preferred reporting items for systematic reviews and meta analyses, CASP: Critical appraisal skills programme and n: Number of records or studies

Analytical framework: The analytical framework was designed to map findings onto four dimensions: Risk assessment, fairness, robustness, and regulatory compliance. Each dimension was operationalized using established metrics. For example, risk assessment was linked to Value at Risk (VaR) and Expected Shortfall, fairness was evaluated using demographic parity and equalized odds, robustness was assessed through adversarial testing and sensitivity analysis, and compliance was mapped against regulatory standards such as Basel III and GDPR[14].

To synthesize findings, employed thematic analysis. Studies were grouped according to the dimension they addressed, and recurring themes were identified. This allowed us to highlight common practices, emerging trends, and gaps in the literature.

Integration of quantitative and qualitative evidence: Because the field spans technical, regulatory, and ethical domains, we integrated both quantitative and qualitative evidence. Quantitative studies provided statistical measures of risk, fairness, and robustness, while qualitative studies offered insights into governance, ethics, and stakeholder perspectives. This mixed methods approach enriched the synthesis and ensured that technical findings were contextualized within broader societal concerns[15].

Limitations of the methodology: While the methodology was rigorous, certain limitations must be acknowledged. First, restricting the search to English language publications may have excluded relevant studies in other languages. Second, the reliance on peer reviewed journals excluded industry reports that might contain practical insights. Third, the dynamic nature of AI in finance means that some findings may quickly become outdated as new technologies and regulations emerge. Nevertheless, the systematic approach provides a reliable snapshot of the field as of 2025[16].

Table 1 summarizes each methodological step and points readers to where each step is described in Methods (search strategy, study selection, PRISMA flow, data extraction, quality assessment, analytical framework). Contents: Step name, brief description, and the citation number used in the manuscript so the method trail is clear.

RESULTS AND DISCUSSION

This section synthesizes the findings from the systematic review, organized across eight subsections that reflect the core dimensions of algorithmic auditing and assurance in AI-driven financial systems. Each subsection highlights key insights, supported by peer-reviewed literature, and is accompanied by a table that consolidates metrics, frameworks, or comparative analyses.

Risk assessment in AI-driven financial systems: Risk assessment is the bedrock of financial auditing; when machine learning is woven into decision pipelines, it raises the upside for insights and efficiency and simultaneously increases fragility. Models ingest large, changing datasets, emit probabilistic outputs, and

can be sensitive to distributional drift or targeted manipulation. A practical, resilient audit therefore combines three complementary perspectives: traditional tail-risk measures, probability-aware expected-loss reasoning that reflects model beliefs, and explicit probes for model uncertainty and adversarial vulnerability.

Expected-loss thinking: Rather than treating scenario probabilities as fixed historical frequencies, auditors should use the model's current predictive beliefs (for example, ensemble or Bayesian surrogates) when estimating expected exposure. This makes expected-loss calculations responsive to recent data shifts and to uncertainty captured by the model itself[17].

Quantile and tail-average risk measures: Quantile summaries (commonly used for capital allocation) and tail-average metrics (which report the mean severity of extreme outcomes) are both necessary. Quantile measures make the likely upper bounds visible, while tail-average measures expose how severe losses can be when the model encounters rare, extreme events. Reporting both types of measures at multiple confidence levels gives a clearer picture of where risk is concentrated[18].

Stress testing and hybrid approaches: Purely data-driven models trained on historically calm periods can under-estimate exposure when markets or regimes change. Stress tests that impose extreme but plausible scenarios (economic shocks, liquidity squeezes, regime shifts) are essential. Hybrid frameworks combining statistical risk models, scenario adjustments, and ML forecasts tend to produce more robust and actionable assessments under stress[1].

Adversarial and distributional fragility: Algorithmic systems can be nudged by small, targeted input changes or can fail silently under plausible alternative data-generating processes. Audits should therefore include (a) adversarial testing to surface manipulation vectors and (b) distributionally-robust evaluations that examine worst-case performance across reasonable alternative distributions. These practices reveal both technical attack surfaces and modelling blind spots[19].

Practical recommendations for auditors:

- Use predictive probabilities (from ensembles or Bayesian surrogates) when estimating expected exposure so that uncertainty and recent data shifts are reflected in risk summaries[17]
- Report both quantile and tail-average metrics at multiple confidence levels (for example, moderate and extreme levels) so stakeholders can see both the boundary and the shape of the tail[18]
- Include adversarial tests and distributional robustness checks in routine audits to detect manipulation vectors and fragile assumptions[19]
- Backtest model probabilities and recalibrate frequently; when drift or miscalibration is detected, re-estimate scenario probabilities or increase tail buffers implied by the tail-average metrics[17-19]

Table 2 lists each metric, gives a concise plain-language definition, and a short note on the typical use case so readers can quickly match concept to practice; read each row left to right to see metric→definition→citation.

Fairness evaluation frameworks: Fairness in AI-driven finance is not just a technical issue but a societal imperative. Discriminatory outcomes in lending or insurance can erode trust and perpetuate inequality. Researchers have proposed fairness auditing frameworks that incorporate statistical parity, ensuring equal treatment across demographic groups defined as[20]:

$$P(\hat{Y}=1 \mid A=a) = P(\hat{Y}=1 \mid A=b)$$

where, Y is predicted outcome, A is protected attribute; statistical parity requires equal probability of positive prediction across groups.

Table 2: Quantitative risk metrics used in AI-driven finance

Metric/Model	Description	Citation(s)
Expected loss	Probability-weighted average exposure where the probabilities reflect the model's current predictive beliefs (e.g., ensemble or Bayesian estimates)	Macas *et al.*[17]
Value at risk (VaR)	A quantile summary of the loss distribution that identifies a loss threshold not expected to be exceeded with a given confidence level	Rahimian and Mehrotra[18]
Conditional VaR/ expected shortfall (CVaR)	A tail-average measure that reports the mean severity of losses beyond the chosen quantile	Rahimian and Mehrotra[18]
Adversarial risk	Measures how expected exposure can increase when inputs are intentionally perturbed	Wachter *et al.*[19]
Distributionally robust risk (DRO)	Evaluates worst-case expected exposure across a set of plausible alternative data distributions	Macas *et al.*[17], Rahimian and Mehrotra[18] and Wachter *et al.*[19]

Abbreviations: VaR: Value at risk, CVaR: Conditional value at risk (Expected Shortfall), DRO: Distributionally robust. Expected loss is the probability-weighted average exposure using model probabilities, adversarial risk denotes exposure change under intentional input perturbations

Table 3: Fairness evaluation frameworks

Criterion/Measure	Description	Citation(s)
Demographic parity	Equal probability of positive outcomes across groups	Mehrabi *et al.*[20]
Equalized odds	Consistent error rates across demographic groups	Guidotti *et al.*[21]
Intersectional fairness	Fairness across overlapping protected categories	Daníelsson *et al.*[22]

DP: Demographic parity, EO: Equalized odds, IF: Intersectional fairness, $P(\hat{Y}=1|A=a)$ denotes the probability of a positive prediction given attribute A = a

Equalized odds, another widely discussed metric, requires that error rates be consistent across groups. This means that false positives and false negatives should not disproportionately affect protected populations[21].

Recent work emphasizes the importance of intersectional fairness, recognizing that individuals may belong to multiple protected categories simultaneously. Auditing frameworks must therefore move beyond single attribute fairness to capture complex social realities[22].

Table 3 condenses the fairness criteria that the review highlights and maps each criterion to how it is applied in financial use cases. Contents: Demographic parity, equalized odds, and intersectional fairness definitions and their brief implications for lending and insurance

Robustness and stability analysis: Robustness is critical in financial AI systems, which must withstand both adversarial attacks and unexpected market shocks.

Sensitivity analysis, expressed as $S(x) = \partial f(x) / \partial x$, is a common tool for evaluating how small changes in input affect outputs[23].

$$f(x) = \text{prediction function; } x = \text{input features}$$

Studies show that adversarial training, where models are exposed to manipulated inputs during training, can significantly improve robustness[24]. However, robustness is not only about technical resilience. Market stability also plays a role, as AI models must adapt to rapidly changing conditions without producing erratic outputs.

Hybrid robustness frameworks, combining adversarial testing with scenario analysis, are increasingly recommended. These frameworks ensure that AI systems remain stable under both technical and economic perturbations[25].

Table 4: Robustness and stability analysis approaches

Approach	Description	Citation(s)
Sensitivity analysis	Evaluating output changes from small input perturbations	Hacker and Passoth[23]
Adversarial training	Training models with manipulated inputs to improve resilience	Apley and Zhu[24]
Hybrid robustness testing	Combining adversarial and scenario analysis	Mitchell *et al.*[25]

S(x): Sensitivity measure (S(x) = ∂f(x)/∂x), Adv. Training: Adversarial training, Hybrid testing: Combined adversarial and scenario analysis

Table 5: Regulatory frameworks and assurance practices

Regulation/Framework	Assurance practice	Citation(s)
Basel III	Stress testing and capital adequacy checks	Riyanul Islam *et al.*[26]
GDPR	Data protection and explainability requirements	Tjoa and Guan[27]
SEC guidelines	Transparency in algorithmic trading	Brown *et al.*[28]

Basel III: International banking capital and stress test standards, GDPR: General data protection regulation and SEC: Securities and exchange commission

Table 4 lists robustness approaches used in the literature and connects each approach to the stability tests or sensitivity measures discussed. Contents: Sensitivity analysis notation, adversarial training description, and hybrid robustness testing with example uses.

Regulatory compliance and governance: Regulatory compliance is a non negotiable aspect of financial AI systems. Global regulators demand that AI models align with standards such as Basel III, GDPR, and SEC guidelines. Compliance frameworks increasingly emphasize explainability, requiring that AI decisions be interpretable to auditors and regulators[26].

Recent studies highlight the tension between innovation and regulation. While AI offers efficiency gains, regulators are concerned about opacity and accountability. Scholars propose governance models that integrate compliance checks directly into AI pipelines, ensuring that outputs are automatically audited for regulatory alignment[27].

Cross jurisdictional compliance is another challenge. Financial institutions operating globally must navigate diverse regulatory landscapes, making harmonized assurance frameworks essential[28].

Table 5 shows regulatory frameworks considered and the assurance practices tied to each regulation in the reviewed studies. Contents: Rows for Basel III, GDPR, SEC, and the assurance practice or audit angle attached to each.

Algorithmic transparency and explainability: Transparency and explainability are vital for building trust in AI-driven finance. Techniques such as SHAP values and LIME provide local explanations of model predictions, helping stakeholders understand why a particular decision was made[29].

Counterfactual explanations, which show how inputs could be altered to change outcomes, are particularly useful in lending contexts. They allow applicants to see what changes would improve their creditworthiness[30].

Comparative studies suggest that no single explainability technique is sufficient. Instead, multi-method approaches, combining SHAP, LIME, and counterfactuals, provide more comprehensive insights[31].

Table 6 summarizes explainability methods compared across papers and highlights the short description of what each technique explains. Contents: SHAP values, LIME, and counterfactual explanations with their primary use cases in finance.

Table 6: Explainability techniques and uses

Technique	Description	Citation(s)
SHAP values	Local feature importance explanations	Panigrahi *et al.*[29]
LIME	Local interpretable model approximations	Alkhanbouli *et al.*[30]
Counter factuals	Showing input changes needed for different outcomes	Barocas and Selbst[31]

SHAP: SHapley additive exPlanations, LIME: Local interpretable model-agnostic explanations, CF: Counterfactual explanation (shows input changes to alter outcome)

Table 7: Ethical dimensions and assurance responses

Dimension	Description	Citation
Bias	Avoiding discriminatory outcomes	Yalcin *et al.*[32]
Accountability	Ensuring responsibility for AI decisions	Ahmed *et al.*[33]
Participatory auditing	Involving stakeholders in assurance processes	Fritz-Morgenthal *et al.*[34]

Bias: Discriminatory outcomes to avoid, Accountability: Assigned responsibility for decisions and PA: Participatory auditing (stakeholder involvement)

Table 8: Comparative frameworks across studies

Framework focus	Key features	Citation(s)
Technical robustness	Adversarial testing, sensitivity analysis	de Castro Vieira *et al.*[35]
Fairness	Statistical parity, equalized odds	Grant[36]
Compliance	Alignment with regulatory standards	Cherian and Candès[37]

TR: Technical robustness, F: Fairness, C: Compliance, "Key features" lists typical methods under each focus

Ethical and societal implications: Ethical considerations are central to algorithmic assurance. Bias, accountability, and trust are recurring themes in the literature. Scholars argue that auditing frameworks must incorporate ethical dimensions alongside technical metrics[32].

Stakeholder perspectives are particularly important. Financial decisions affect individuals and communities, meaning that assurance frameworks must balance efficiency with fairness and public trust[33].

Recent reviews emphasize the need for participatory auditing, where stakeholders are actively involved in evaluating AI systems. This ensures that ethical concerns are not overlooked in favor of technical optimization[34].

Table 7 collects ethical dimensions discussed and the practical assurance or governance responses proposed in the literature. Contents: Bias mitigation, accountability mechanisms, and participatory auditing approaches with brief notes on stakeholder roles.

Comparative frameworks across studies: Comparative analysis reveals significant variation in auditing frameworks across studies. Some emphasize technical robustness, while others prioritize fairness or compliance. Systematic reviews suggest that integrated frameworks are more effective, as they capture multiple dimensions simultaneously[35].

Cross-study synthesis highlights common practices, such as the use of mixed-methods approaches and the integration of explainability tools. However, gaps remain, particularly in harmonizing frameworks across jurisdictions[36].

Emerging trends point toward adaptive frameworks that evolve with technological and regulatory changes. These frameworks are designed to remain relevant in dynamic financial environments[37].

Table 8 compares study-level framework foci and lists the key features authors emphasize in each cluster (robustness, fairness, compliance). Contents: Framework focus categories, key features for each, and citations indicating representative studies.

Table 9: Integrated assurance framework and assurance index

Dimension	Metrics/Approach	Citation(s)
Risk (R)	VaR, expected shortfall, adversarial risk testing	Urman *et al.*[38]
Fairness (F)	Demographic parity, equalized odds, intersectional fairness	Zhou *et al.*[39]
Robustness (B)	Sensitivity analysis, adversarial training, hybrid robustness	Zhou *et al.*[39]
Compliance (C)	Basel III, GDPR, SEC guidelines	Zhou *et al.*[39]
Integration	Assurance Index combining weighted dimensions	Zhou *et al.*[39]
Stakeholders	Participatory auditing and community involvement	Kabir *et al.*[40]

R: Risk, F: Fairness, B: Robustness, C: Compliance; α, β, γ, δ: Weighting coefficients and VaR: Value at risk (used under risk)

Integrated assurance framework proposal: The synthesis of evidence across risk, fairness, robustness, and compliance dimensions points toward the necessity of a unified assurance framework for AI-driven financial systems. Fragmented approaches, while useful in isolated contexts, fail to capture the interconnected nature of financial risks and ethical obligations. An integrated framework provides a holistic lens, ensuring that technical, regulatory, and societal concerns are addressed simultaneously[38].

At the heart of this proposal is the Assurance Index, a composite measure that aggregates performance across four dimensions: Risk (R), fairness (F), robustness (B), and compliance (C). The index is expressed mathematically as:

$$\text{Assurance index} = \alpha R+\beta F+\gamma B+\delta C$$

where, α, β, γ, δ represent weighting coefficients determined by institutional priorities.

Framework operates in three stages:

- **Assessment stage:** Each dimension is evaluated using standardized metrics. Risk is measured through VaR and Expected Shortfall, fairness through demographic parity and equalized odds, robustness through adversarial testing and sensitivity analysis, and compliance through alignment with Basel III and GDPR
- **Integration stage:** Scores from each dimension are normalized and aggregated into the Assurance Index. This ensures comparability across metrics that may otherwise operate on different scales
- **Decision stage:** The assurance index informs governance decisions, such as whether an AI system is fit for deployment, requires remediation, or should be rejected. Institutions can set thresholds, for example requiring an Assurance Index above 0.75 for approval

Beyond technical metrics, the framework emphasizes stakeholder involvement. Participatory auditing ensures that affected communities have a voice in evaluating fairness and ethical dimensions. This participatory element strengthens legitimacy and public trust, moving assurance beyond a purely technical exercise[40].

Table 9 shows the proposed Assurance Index components, the mathematical aggregation form, and the metrics used to score each dimension. Contents: Rows for Risk (R), Fairness (F), Robustness (B), Compliance (C), metrics used for each and the formula:

$$\text{Assurance index} = \alpha R+\beta F+\gamma B+\delta C$$

CONCLUSION

This systematic review synthesized the literature on algorithmic auditing, assurance, and governance of AI in financial contexts. We traced methodological approaches across risk assessment, fairness evaluation, robustness testing, regulatory compliance, and explainability, highlighting areas of concentrated evidence.

The review assembled quantitative metrics, conceptual frameworks, and practical assurance practices to form a coherent picture of current research. Findings indicate that a variety of methods are in use, but integration across the different assurance dimensions remains uneven. The proposed assurance index provides a transparent way to compare how studies score risk, fairness, robustness, and compliance. The study selection and quality appraisal ensured that the included articles directly addressed auditing or assurance concerns and met predefined standards. Heterogeneity in definitions and reporting limited the degree of direct comparability across some studies. Despite these constraints, the synthesis clarifies prevailing themes and organizes dispersed findings into an accessible structure. Overall, the manuscript strengthens the evidence base by bringing fragmented work together and making cross-study patterns visible. In sum, this review offers a concise, evidence-based account of how algorithmic assurance is studied in finance and how different strands of research interconnect.

SIGNIFICANCE STATEMENT

This systematic review synthesizes evidence and proposes a practical Assurance Index that unifies risk, fairness, robustness, and compliance to guide safer deployment of AI in financial systems. By combining technical diagnostics with participatory governance, we offer actionable pathways for institutions and regulators to detect, mitigate, and transparently report algorithmic harms without unduly hindering innovation. We aim to equip practitioners, policymakers, and affected communities with a clear, trustworthy framework that strengthens accountability, resilience, and public confidence in AI-driven finance.

ACKNOWLEDGMENT

The authors gratefully acknowledge the constructive feedback and critical review provided by colleagues and peer reviewers, which substantially improved the clarity and rigor of this systematic review. We thank the university library staff and database providers for their assistance with the literature searches and retrieval of source materials.

REFERENCES

1. Mökander, J. and L. Floridi, 2021. Ethics-based auditing to develop trustworthy AI. Minds Mach., 31: 323-327.
2. Kodali, S., K.A. Sravanthi, N. Satheesh, U. Basu, M. Bose and A. Srivastava, 2025. Auditing AI in finance: A framework for interpretable compliance in algorithmic decisions. Bus. Econ. Res. J., 14: 79-93.
3. Arshad, M.D. and C.K. Tripathi, 2025. Algorithmic accountability and ethical AI frameworks for regulatory governance in financial technologies. Int. J. Sci. Res. Arch., 16: 796-799.
4. Tian, X., Z.Y. Tian, S.F.A. Khatib and Y. Wang, 2024. Machine learning in internet financial risk management: A systematic literature review. PLoS ONE, Vol. 19. 10.1371/journal.pone.0300195.
5. Thamae, R.I. and N.M. Odhiambo, 2022. The impact of bank regulation on bank lending: A review of international literature. J. Banking Regul., 23: 405-418.
6. Laine, J., M. Minkkinen and M. Mäntymäki, 2024. Ethics-based AI auditing: A systematic literature review on conceptualizations of ethical principles and knowledge contributions to stakeholders. Inf. Manage., Vol. 61. 10.1016/j.im.2024.103969.
7. Nissim, D., 2022. Big data, accounting information, and valuation. J. Finance Data Sci., 8: 69-85.
8. Rethlefsen, M.L., S. Kirtley, S. Waffenschmidt, A.P. Ayala and D. Moher *et al*., 2021. PRISMA-S: An extension to the PRISMA statement for peporting literature searches in systematic reviews. Syst. Rev., Vol. 10. 10.1186/s13643-020-01542-z.
9. Marshall, I.J. and B.C. Wallace, 2019. Toward systematic review automation: A practical guide to using machine learning tools in research synthesis. Syst. Rev., Vol. 8. 10.1186/s13643-019-1074-9.
10. Ioannidis, J.P.A., 2016. The mass production of redundant, misleading, and conflicted systematic reviews and meta-analyses. Milbank Q., 94: 485-514.

11. Page, M.J., J.E. McKenzie, P.M. Bossuyt, I. Boutron and T.C. Hoffmann *et al*., 2021. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ, Vol. 372. 10.1136/bmj.n71.
12. Shea, B.J., B.C. Reeves, G. Wells, M. Thuku and C. Hamel *et al*., 2017. AMSTAR 2: A critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. BMJ, Vol. 358. 10.1136/bmj.j4008.
13. Leocádio, D., L. Malheiroand J. Reis, 2024. Artificial intelligence in auditing: A conceptual framework for auditing practices. Administrative Sci., Vol. 14. 10.3390/admsci14100238.
14. Mertens, D.M., 2021. Transformative research methods to increase social impact for vulnerable groups and cultural minorities. Int. J. Qual. Methods, Vol. 20. 10.1177/16094069211051563.
15. Maleki, F. and M. Karami, 2024. Methodological challenges for the responsible use of AI in systematic reviews: Risk of bias assessment. J. Evidence-Based Med., 17: 712-713.
16. Lessmann, S., B. Baesens, H.V. Seow and L.C. Thomas, 2015. Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. Eur. J. Oper. Res., 247: 124-136.
17. Macas, M., C. Wu and W. Fuertes, 2024. Adversarial examples: A survey of attacks and defenses in deep learning-enabled cybersecurity systems. Expert Syst. Appl., Vol. 238. 10.1016/j.eswa.2023.122223.
18. Rahimian, H. and S. Mehrotra, 2022. Frameworks and results in distributionally robust optimization. Open J. Math. Optim., Vol. 3. 10.5802/ojmo.15.
19. Wachter, S., B. Mittelstadt and C. Russell, 2018. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harvard J. Law Technol., 31: 842-887.
20. Mehrabi, N., F. Morstatter, N. Saxena, K. Lerman and A. Galstyan, 2021. A survey on bias and fairness in machine learning. ACM Comput. Surv., Vol. 54. 10.1145/3457607.
21. Guidotti, R., A. Monreale, S. Ruggieri, F. Turini, F. Giannotti and D. Pedreschi, 2018. A survey of methods for explaining black box models. ACM Comput. Surv., Vol. 51. 10.1145/3236009.
22. Daníelsson, J., R. Macrae and A. Uthemann, 2022. Artificial intelligence and systemic risk. J. Banking Finance, Vol. 140. 10.1016/j.jbankfin.2021.106290.
23. Hacker, P. and J.H. Passoth, 2022. Varieties of AI Explanations Under the Law. From the GDPR to the AIA, and Beyond. In: xxAI - Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers, Holzinger, A., R. Goebel, R. Fong, T. Moon, K.R. Müller and W. Samek (Eds.), Springer International Publishing, Cham, Switzerland, ISBN: 978-3-031-04083-2, pp: 343-373.
24. Apley, D.W. and J. Zhu, 2020. Visualizing the effects of predictor variables in black box supervised learning models. J. R. Stat. Soc. Ser. B: Stat. Methodol., 82: 1059-1086.
25. Mitchell, R., E. Frank and G. Holmes, 2022. GPUTreeShap: Massively parallel exact calculation of SHAP scores for tree ensembles. PeerJ Comput. Sci., Vol. 8. 10.7717/peerj-cs.880.
26. Riyanul Islam, M., Mobyen Uddin Ahmed, S. Barua and S. Begum, 2022. A systematic review of explainable artificial intelligence in terms of different application domains and tasks. Appl. Sci., Vol. 12. 10.3390/app12031353.
27. Tjoa, E. and C. Guan, 2021. A survey on explainable artificial intelligence (XAI): Toward medical XAI. IEEE Trans. Neural Networks Learn. Syst., 32: 4793-4813.
28. Brown, S., J. Davidovic and A. Hasan, 2021. The algorithm audit: Scoring the algorithms that score us. Big Data Soc., Vol. 8. 10.1177/2053951720983865.
29. Panigrahi, B., S. Razavi, L.E. Doig, B. Cordell, H.V. Gupta and K. Liber, 2025. On robustness of the explanatory power of machine learning models: Insights from a new explainable AI approach using sensitivity analysis. Water Resour. Res., Vol. 61. 10.1029/2024WR037398.
30. Alkhanbouli, R., H.M.A. Almadhaani, F. Alhosani and M.C.E. Simsekler, 2025. The role of explainable artificial intelligence in disease prediction: A systematic literature review and future research directions. BMC Med. Inf. Decis. Making, Vol. 25. 10.1186/s12911-025-02944-6.
31. Barocas, S. and A.D. Selbst, 2016. Big data's disparate impact. Calif. Law Rev., 104: 671-732.
32. Yalcin, G., E. Themeli, E. Stamhuis, S. Philipsen and S. Puntoni, 2023. Perceptions of justice by algorithms. Artif. Intell. Law, 31: 269-292.

33. Ahmed, F., N.S. Naz, S. Khan, Ateeq Ur Rehman, W.M. Ismael and M.A. Khan, 2026. Explainable artificial intelligence (XAI) in medical imaging: A systematic review of techniques, applications, and challenges. BMC Med. Imaging, Vol. 26. 10.1186/s12880-025-02118-w.
34. Fritz-Morgenthal, S., B. Hein and J. Papenbrock, 2022. Financial risk management and explainable, trustworthy, responsible AI. Front. Artif. Intell., Vol. 5. 10.3389/frai.2022.779799.
35. de Castro Vieira, J.R., F. Barboza, D. Cajueiro and H. Kimura, 2025. Towards fair AI: Mitigating bias in credit decisions-a systematic literature review. J. Risk Financ. Manage., Vol. 18. 10.3390/jrfm18050228.
36. Grant, D.G., 2023. Equalized odds is a requirement of algorithmic fairness. Synthese, Vol. 201. 10.1007/s11229-023-04054-0.
37. Cherian, J.J. and E.J. Candès, 2024. Statistical inference for fairness auditing. J. Mach. Learn. Res., Vol. 25.
38. Urman, A., I. Smirnov and J. Lasser, 2024. The right to audit and power asymmetries in algorithm auditing. EPJ Data Sci., Vol. 13. 10.1140/epjds/s13688-024-00454-5.
39. Zhou, S., C. Liu, D. Ye, T. Zhu, W. Zhou and P.S. Yu, 2022. Adversarial attacks and defenses in deep learning: From a perspective of cybersecurity. ACM Comput. Surv., Vol. 55. 10.1145/3547330.
40. Kabir, S., M.S. Hossain and K. Andersson, 2025. A review of explainable artificial intelligence from the perspectives of challenges and opportunities. Algorithms, Vol. 18. 10.3390/a18090556.