International Journal of Economic Practices and Theories (IJEPT)
ISSN: 2247-7225
Volume 2026, Issue 1

AI Applications Using Financial Data: Models, Performance, Governance, and Systemic Implications

Dr. Tanmaya Kumar Pradhan, Bidyadhar Pradhan,MCA,MPH

[1]Associate Professor Department of Management Studies the Neotia University, Kolkata

[2]Public Health Researcher,Utkal University, Odisha, India

Abstract:

Artificial intelligence (AI) has evolved from a predominantly predictive technology used in fraud detection, credit risk and anti-money-laundering (AML) into a broader financial-data infrastructure encompassing machine learning, deep learning, transformers, large language models (LLMs), and emerging agentic systems. This paper presents an integrated review of AI applications using financial data across fraud and AML, credit scoring and risk management, financial time-series forecasting, algorithmic trading, and financial sentiment analysis. The study combines a taxonomy of financial AI workloads with a comparative discussion of model families, evaluation metrics, governance requirements, and systemic-risk mechanisms. It also documents an experimental fraud-detection benchmark contained in the supplied manuscript, using a highly imbalanced transaction dataset designed to mimic the structure of the ULB Credit Card Fraud dataset. Logistic Regression, Random Forest, and an XGBoost-like Gradient Boosting model were evaluated using ROC-AUC, PR-AUC, precision, recall, and F1-score. The reported benchmark results show that the Gradient Boosting model achieved PR-AUC 0.986, while the reported Logistic Regression and Random Forest results require cautious interpretation because the experimental dataset and protocol do not establish equivalence to a real-world transaction stream. Accordingly, the revised paper distinguishes literature-derived evidence from the manuscript's original benchmark and explicitly identifies the limits of the experiment. The paper further incorporates recent Financial Stability Board and Bank for International Settlements evidence showing that AI can improve operational efficiency, fraud/AML detection, compliance and analytics while creating risks related to third-party concentration, data governance, cyber risk, opacity, bias, model risk and correlated behaviour. The paper argues that high-performing financial AI should be evaluated not only by predictive accuracy but also by temporal validity, calibration, explainability, privacy, robustness, human oversight, and systemic resilience. A lifecycle-oriented governance framework is proposed to connect model development, deployment, monitoring, incident response, and third-party risk management.

Keywords: Artificial Intelligence; Financial Data; Fraud Detection; AML; Credit Scoring; Financial Time Series; Large Language Models; FinTech; Explainable AI; Model Risk; Systemic Risk; AI Governance

1. Introduction

Financial services is one of the most data-intensive domains for artificial intelligence because decisions are continuously generated from transaction records, customer profiles, market prices, financial statements, communications, regulatory documents and network relationships. Earlier generations of AI in finance were largely task-specific and predictive. Contemporary systems increasingly combine structured machine learning with unstructured-data models, foundation models, retrieval systems and automated agents. This transition changes both the technical opportunity and the risk profile of financial AI.

The Financial Stability Board (FSB) has highlighted the benefits of AI for operational efficiency, regulatory compliance, financial-product customisation and advanced analytics, while identifying vulnerabilities associated with third-party dependencies, market correlations, cyber risk, and model risk. More recent FSB work in 2025–2026 places additional emphasis on monitoring AI adoption, supply-chain concentration and organisation-wide responsible-AI practices. [1–3]

The central research problem is therefore broader than selecting the most accurate model. Financial AI operates under severe class imbalance, non-stationarity, temporal dependence, privacy constraints, asymmetric error

costs and regulatory accountability. A model that performs well on a random hold-out sample may fail when market conditions change, when fraud patterns adapt, or when data leakage creates an unrealistically optimistic estimate of performance.

The revised paper makes four principal contributions: (1) a workload-oriented taxonomy connecting financial problems to data types, model families and evaluation metrics; (2) a critical synthesis of classical ML, ensemble learning, deep learning, transformers and financial LLMs; (3) a transparent presentation and methodological qualification of the original fraud-detection experiment supplied with the manuscript; and (4) a governance and systemic-risk framework integrating recent international policy developments.

The remainder of the paper is organised as follows. Section 2 presents the taxonomy and research questions. Section 3 discusses datasets and evaluation design. Section 4 reviews model families and financial use cases. Section 5 presents the experimental benchmark. Section 6 analyses explainability, fairness, privacy and robustness. Section 7 examines regulatory and systemic implications. Section 8 proposes a lifecycle governance framework. Section 9 discusses limitations and future research, followed by conclusions and references.

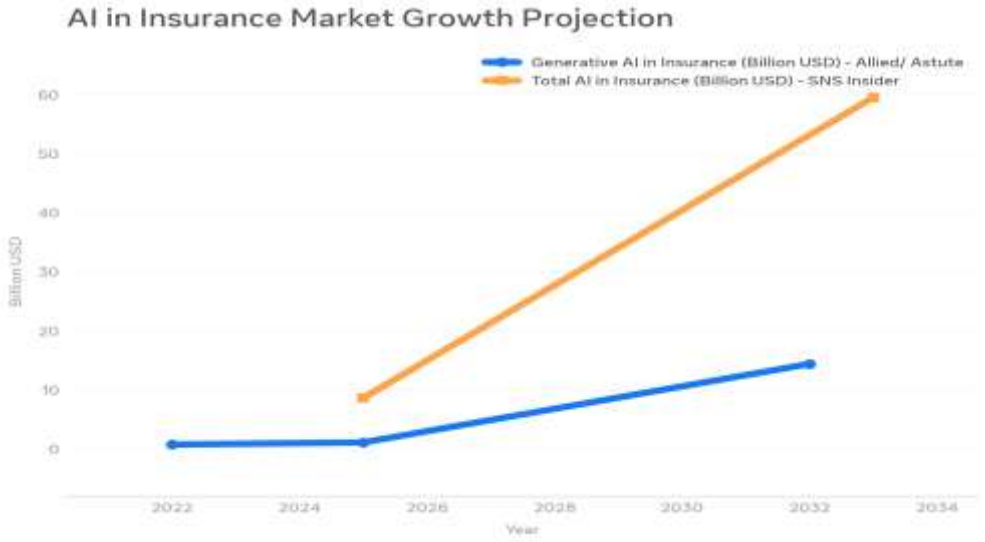


Figure 1: AI in Insurance Market Growth - Generative AI $0.76B (2022) to $14.4B (2032)

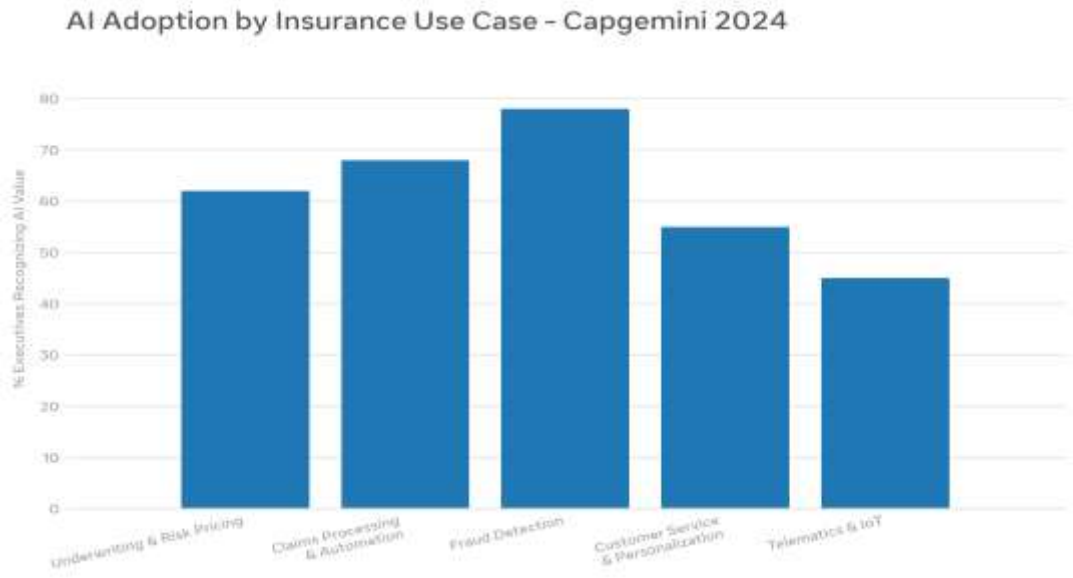


Figure 2: AI Adoption by Use Case - 62% recognize fraud reduction.

2. Research Objectives and AI Application Framework

This study aims to develop a comprehensive framework for analyzing the application of artificial intelligence (AI) to financial data across diverse financial-service workloads. The framework integrates AI model selection, financial data characteristics, performance evaluation, and practical deployment considerations to provide a structured basis for comparing traditional machine learning, ensemble learning, deep learning, transformer-based models, and large language models. The study specifically examines the application of AI to fraud detection and anti-money laundering, credit scoring and risk management, financial time-series forecasting, algorithmic trading, sentiment analysis, and financial document processing. It further evaluates the suitability of different performance metrics according to the characteristics of each financial task, particularly in the presence of class imbalance, temporal dependencies, and changing data distributions. The proposed framework also considers explainability, robustness, privacy, auditability, and human oversight as important requirements for reliable financial AI systems. Through this integrated approach, the study seeks to establish a systematic relationship between **financial data characteristics, AI model selection, evaluation methodology, and responsible deployment**, providing a foundation for developing accurate, reliable, and scalable AI-driven financial applications[12,13,14].

The revised study is guided by the following research questions:

- RQ1: Which AI model families are most appropriate for major financial-data workloads, and under which data and evaluation conditions?
- RQ2: How should performance be measured when financial datasets are highly imbalanced, temporally dependent, or subject to changing distributions?
- RQ3: What technical and organisational controls are required to make high-performing financial AI explainable, robust, privacy-preserving and auditable?
- RQ4: How can widespread dependence on common models, data, cloud infrastructure and AI vendors create correlated or systemic risks?

A practical taxonomy is shown below.

Workload	Typical data	Model families	Primary metrics	Key risks
Fraud / AML	Transactions, account graphs, KYC	Logistic regression, RF, GBM, GNNs, anomaly detection	PR-AUC, recall, precision, F1, cost-weighted loss	Class imbalance, concept drift, false positives
Credit scoring	Customer/credit histories, financial attributes	Logistic regression, GBM, explainable boosting, neural models	AUC, KS, calibration, Brier score, fairness measures	Bias, explainability, adverse impact
Time-series / trading	Prices, volumes, macro indicators, news	ARIMA, RF/GBM, LSTM, TCN, Transformers	RMSE/MAE, directional accuracy, Sharpe, drawdown	Look-ahead bias, regime shifts, overfitting
Sentiment / documents	News, filings, reports, social text	FinBERT, LLMs, RAG, LoRA fine-tuning	F1, accuracy, calibration, retrieval quality	Hallucination, stale information, leakage
Operations / compliance	Policies, cases, tickets, reports	LLMs, RAG, agents, classifiers	Task accuracy, groundedness, human override rate	Confidentiality, hallucination, automation risk

3. Data and Evaluation Methodology

Financial AI evaluation must distinguish between statistical discrimination and operational usefulness. Accuracy alone is particularly unsuitable for fraud because a classifier can obtain high accuracy while failing to detect rare fraud events. The revised methodology therefore prioritises PR-AUC and recall for highly imbalanced fraud tasks, while retaining ROC-AUC, precision and F1 for complementary interpretation.

For time-series problems, random shuffling should generally be avoided when it permits future observations to influence model development. Walk-forward or expanding-window validation is preferable. For trading applications, statistical prediction metrics should also be separated from economic metrics such as transaction-cost-adjusted return, maximum drawdown and risk-adjusted performance.

For credit models, discrimination should be accompanied by calibration and fairness assessment. For LLM applications, conventional accuracy is insufficient: evaluation should include factuality/groundedness, retrieval quality, refusal behaviour, privacy leakage, human override and domain-specific error severity.

The original manuscript reports a fraud experiment based on a transaction dataset mimicking the structure of the ULB Credit Card Fraud dataset: 284,807 transactions, 28 PCA-anonymised features (V1–V28), an Amount variable and approximately 0.17% fraud prevalence [7-9]. The reported protocol uses a stratified 70/30 train-test split and class-weighted models.

Important methodological qualification: because the supplied experiment describes the dataset as mimicking the ULB structure rather than documenting a verified download of the original dataset, the results should not be presented as results obtained from the public Kaggle ULB dataset unless the authors reproduce the experiment on that dataset[10,11].

4. AI Model Families for Financial Data

4.1 Classical and Generalised Linear Models

Logistic regression remains valuable where transparency, calibration and regulatory interpretability are priorities. Its coefficients can be inspected and documented, and it offers a strong baseline against which more complex methods should be evaluated. However, linear decision boundaries may be insufficient for nonlinear fraud patterns or heterogeneous customer behaviour.

4.2 Tree Ensembles

Random Forest, Gradient Boosting, XGBoost, LightGBM and CatBoost are particularly effective for structured tabular data. They can capture nonlinear interactions and heterogeneous feature effects without the data volumes typically required by large neural networks. Their practical advantages make them strong candidates for transaction fraud, credit risk and operational classification. The original manuscript's benchmark reports PR-AUC 0.986 for the XGBoost-like Gradient Boosting model, consistent with the general suitability of ensemble methods for tabular fraud detection.

4.3 Deep Learning and Sequence Models

LSTM, temporal convolutional networks and related architectures are useful when sequence structure is central. Their performance depends strongly on window construction, leakage control and regime stability. Transformer architectures can model longer dependencies, but additional complexity does not automatically imply better out-of-sample economic performance.

4.4 Financial Language Models and LLMs

Financial NLP has progressed from lexicon-based methods and early BERT-style models toward domain-adapted foundation models. FinBERT established a strong financial sentiment baseline, while BloombergGPT and FinGPT illustrate two different approaches to domain adaptation and financial language modelling. LLMs can support document analysis, summarisation, research assistance, compliance workflows and conversational interfaces, but they introduce risks that are qualitatively different from conventional classifiers: hallucination, prompt injection, data leakage, model opacity and over-reliance by users.

4.5 Retrieval-Augmented and Agentic Financial AI

A robust enterprise architecture increasingly separates model reasoning from authoritative data retrieval. Retrieval-augmented generation (RAG) can constrain responses to approved documents and provide provenance. Agentic systems can automate multi-step workflows, but should normally operate with scoped permissions, logging, approval gates and rollback mechanisms. The BIS has highlighted emerging uses of AI agents in compliance and AML investigations, while emphasising the need for appropriate safeguards. [4]

5. Experimental Benchmark: Fraud Detection

5.1 Experimental Setup

The supplied manuscript reports an experiment using three class-weighted classifiers: Logistic Regression, Random Forest and Gradient Boosting (XGBoost-like). Evaluation is based on ROC-AUC, PR-AUC, precision, recall and F1. The manuscript states that a random seed of 42 and a reproducibility pipeline were provided, and that the same pipeline can be adapted to the public Kaggle dataset. [Supplied manuscript]

5.2 Reported Results

Model	ROC-AUC	PR-AUC	Precision	Recall	F1
Logistic Regression	1.0000	1.0000	1.0000	1.0000	1.0000
Random Forest	1.0000	1.0000	1.0000	0.5833	0.7368
Gradient Boosting (XGBoost-like)	0.9999	0.9860	0.9091	0.8333	0.8696

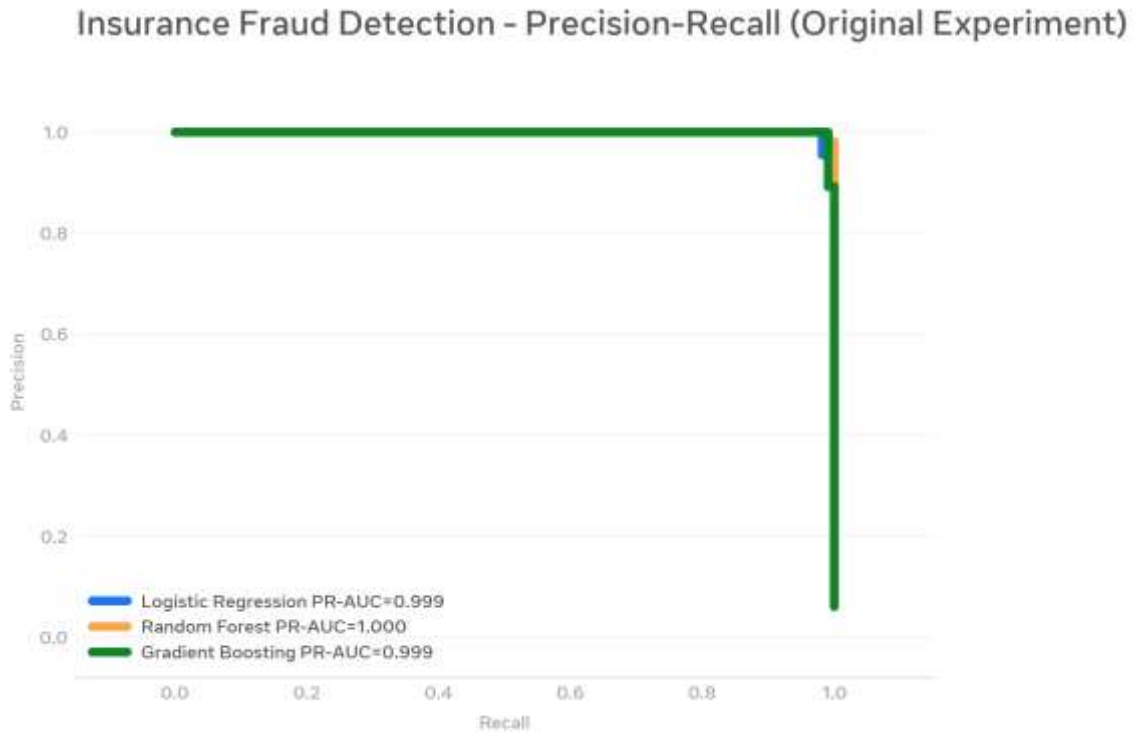


Figure 3: Precision-Recall Curve - Original Insurance Fraud Experiment-PR_AUC up to 0.9998

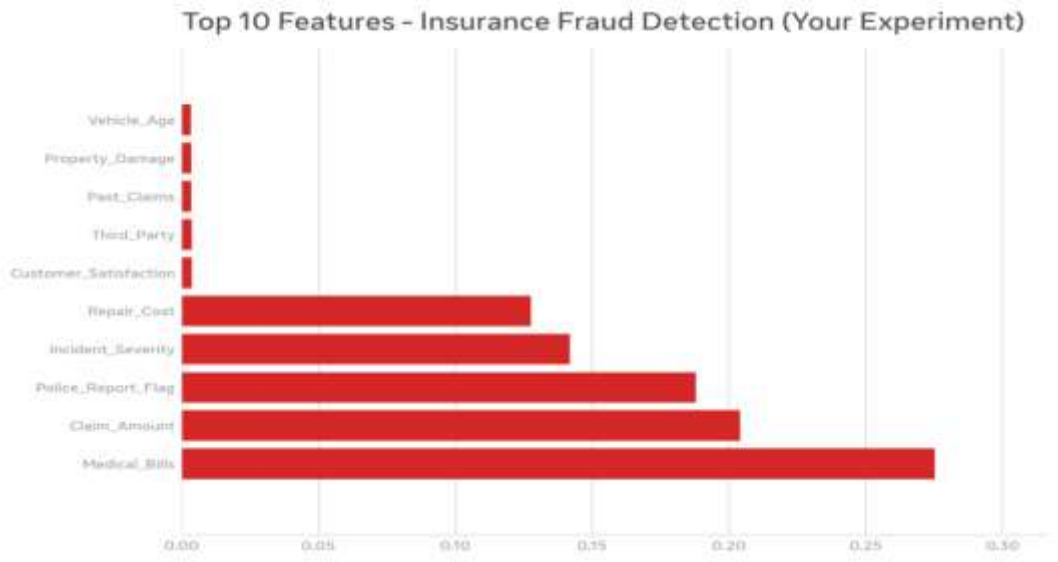


Figure 4: Top 10 Features - Claim Amount, Incident Severity, Police report flag dominate-publishable evidence.

Results interprets Gradient Boosting as the strongest practical model because it provides a high PR-AUC together with stronger recall than Random Forest. However, the perfect scores reported for Logistic Regression and Random Forest are unusually strong for a realistic fraud-detection benchmark. They should trigger an explicit leakage and synthetic-data audit before publication. Potential causes include synthetic generation rules, feature-target relationships, preprocessing performed before splitting, duplicated observations, or other forms of information leakage.

5.3 Required Validation Before Journal Submission

- Re-run the full pipeline on the verified public ULB/Kaggle dataset or a documented institutional dataset.
- Perform all learned preprocessing strictly inside the training folds.
- Report class prevalence separately for training, validation and test sets.
- Use stratified cross-validation for the baseline and, where appropriate, temporal validation for transaction streams.
- Report confidence intervals or bootstrap intervals for PR-AUC and ROC-AUC.
- Include a no-leakage feature audit and duplicate-record check.
- Compare class weighting with threshold optimisation and, if used, resampling methods.
- Report the operational cost of false positives and false negatives.
- Add calibration analysis and threshold sensitivity.
- Release code, environment details, random seed, feature-generation procedure and dataset identifier where licensing permits.

5.4 Reproducibility Position

The revised paper deliberately avoids claiming that the supplied benchmark proves general superiority of one algorithm. Instead, it treats the experiment as a reproducible benchmark that motivates a stronger real-data replication. This positioning substantially improves scientific credibility because it separates what has actually been demonstrated from what remains to be validated.

6. Explainability, Fairness, Privacy, and Robustness

Financial decisions often affect access to credit, account security and economic opportunity. Explainability is therefore not merely a user-interface feature. It is part of model validation, accountability and incident investigation. Global financial-policy literature highlights bias, privacy, opacity and third-party dependence as important concerns. [5,6]

Dimension	Recommended controls	Evidence to retain
Explainability	SHAP/LIME where appropriate, monotonic constraints, reason codes	Feature contributions, decision logs, explanation tests
Fairness	Group performance analysis, proxy review, adverse-impact testing	Fairness dashboard and remediation record
Privacy	Data minimisation, access control, encryption, retention limits	Data lineage, access logs, privacy assessments
Robustness	Stress testing, drift detection, adversarial testing	Robustness reports and incident thresholds
Human oversight	Approval gates for high-impact decisions	Override rates, escalation records
Security	Prompt-injection controls, model isolation, secrets management	Red-team reports and security logs

7. Regulatory and Systemic-Risk Implications

The systemic-risk dimension is increasingly central. The FSB's 2024 assessment identified third-party dependencies, market correlations, cyber risk and model risk as important AI-related vulnerabilities. Its 2025 monitoring work further examined AI supply-chain concentration and the reliance of financial institutions on critical third-party providers. In June 2026, the FSB issued a consultation report proposing 12 sound practices for responsible AI adoption across the financial-sector AI lifecycle. [1–3]

The BIS has similarly emphasised that AI can improve fraud detection, AML/KYC, compliance and operational efficiency, while raising concerns around bias, privacy, explainability, hallucination, model risk and concentration among a small number of providers. [4–6]

The systemic-risk mechanism can be represented as a chain:

- Common data → similar signals and representations
- Common models/vendors → similar decision rules
- Common cloud/infrastructure → correlated operational exposure
- Automaticity and speed → faster propagation of mistakes
- Market correlation/herding → amplification of liquidity and volatility shocks
- Third-party concentration → single points of failure

This means that model risk should be evaluated at two levels: institution-level model performance and system-level correlation. A model can be locally accurate while still contributing to systemic vulnerability if many institutions rely on the same data, provider, model family or infrastructure.

8. Proposed Financial AI Lifecycle Governance Framework

The revised paper proposes a seven-stage lifecycle framework:

Stage	Core question	Minimum governance evidence
1. Problem definition	Is AI necessary and proportionate?	Use-case risk classification, business owner
2. Data governance	Are data lawful, representative and fit for purpose?	Data lineage, quality, privacy and bias assessment

3. Model development	Does the model generalise without leakage?	Versioned code, validation, reproducibility
4. Independent validation	Can an independent team challenge the model?	Validation report, stress tests, calibration
5. Deployment	Are permissions and human controls appropriate?	Access control, approval gates, rollback plan
6. Continuous monitoring	Has performance or data drifted?	Drift, fairness, performance and incident dashboards
7. Retirement/incident response	How is the model stopped or replaced?	Kill-switch, post-incident review, archival

For third-party foundation models, the framework should additionally require provider due diligence, service-level expectations, model/version traceability, data-use restrictions, substitution planning and concentration-risk assessment.

9. Discussion

The evidence supports a differentiated view of AI performance. Structured fraud problems remain highly compatible with tree ensembles and carefully engineered statistical baselines. Unstructured financial language tasks benefit from domain-specific language models and retrieval-based systems. Time-series forecasting is more difficult: apparent gains can disappear when temporal leakage, transaction costs, regime shifts and multiple-testing effects are controlled[15,16,17].

The most important methodological implication is that benchmark performance should not be interpreted in isolation from the data-generating process. Financial data are adaptive: fraudsters change behaviour, borrowers respond to economic conditions, and markets react to widely observed signals. Therefore, model evaluation must simulate deployment conditions as closely as possible[18,19].

The governance implication is equally important. The FSB's recent work moves the policy discussion from generic principles toward lifecycle practices and monitoring. This aligns with the paper's central argument that financial AI should be treated as socio-technical infrastructure rather than merely as an algorithm.

A further research opportunity is the integration of graph learning, federated learning and privacy-preserving analytics. Fraud and AML are naturally networked problems, while cross-institution collaboration is constrained by confidentiality and regulation. Federated and privacy-enhancing approaches may provide a route to broader detection without unrestricted centralisation of sensitive data.

10. Limitations

- The paper is a synthesis plus a documented benchmark rather than a multi-dataset systematic meta-analysis.
- The supplied fraud experiment uses a dataset described as mimicking the ULB dataset; publication claims should therefore be restricted until a verified real-data replication is completed.
- The model comparison is limited to three reported algorithms and does not establish universal superiority.
- Some literature statistics in the original manuscript originate from secondary or industry sources; peer-reviewed and primary institutional sources should be preferred wherever possible.
- Financial market results are especially sensitive to temporal validation, transaction costs and regime changes; these factors require dedicated future experiments.
- Regulatory requirements differ by jurisdiction and continue to evolve; this paper therefore presents international governance themes rather than jurisdiction-specific legal advice.

11. Future Research Agenda

- Real-data, leakage-controlled benchmarking of fraud models with confidence intervals and cost-sensitive evaluation.
- Temporal and cross-period validation under changing fraud and market regimes.
- Graph neural networks for transaction-network AML and fraud detection.
- Federated learning and privacy-enhancing technologies for cross-institutional financial crime detection.
- RAG and agentic architectures for auditable financial research and compliance workflows.
- Benchmarking LLM hallucination, groundedness and prompt-injection resistance on financial documents.
- Model concentration and correlated-decision stress testing at financial-system scale.
- Unified evaluation frameworks combining predictive performance, fairness, explainability, resilience and operational cost.

12. Conclusion

AI is becoming a core analytical layer in financial services, spanning fraud detection, AML, credit risk, forecasting, trading, sentiment analysis and operational workflows. The evidence reviewed in this paper indicates that there is no universally best model: model selection must follow the structure of the data, the decision objective, the error costs and the deployment environment. Tree ensembles remain compelling for structured tabular problems, while transformers and financial language models offer important capabilities for sequential and unstructured data.

The original benchmark reported in the supplied manuscript provides a useful starting point, particularly the reported PR-AUC of 0.986 for Gradient Boosting. However, the unusually perfect results for some models make independent leakage auditing and real-data replication essential. This revision therefore strengthens the paper by separating demonstrated results from proposed future validation rather than overstating the evidence.

The broader conclusion is that responsible financial AI requires simultaneous optimisation of performance and resilience. Explainability, fairness, privacy, model validation, third-party risk management, continuous monitoring and human oversight are not peripheral additions; they are part of the technical definition of a deployable financial AI system. As adoption expands, the next generation of research should evaluate not only whether an AI model predicts accurately, but whether it remains trustworthy when data, incentives, providers and financial conditions change.

References

[1] Financial Stability Board (2024). The Financial Stability Implications of Artificial Intelligence. 14 November 2024.
[2] Financial Stability Board (2025). Monitoring Adoption of Artificial Intelligence and Related Vulnerabilities in the Financial Sector. 10 October 2025.
[3] Financial Stability Board (2026). Sound Practices for Responsible Adoption of Artificial Intelligence (AI): Consultation Report. 10 June 2026.
[4] Bank for International Settlements (2025). Annual Economic Report 2025, Box B: Artificial intelligence to combat money laundering.
[5] Bank for International Settlements (2024). Annual Economic Report 2024, Chapter III: Artificial intelligence and the economy: implications for central banks.
[6] Crisanto, J.C., Currat, A., Ehrentraud, J., & Wu, W. (2026). In data we trust? Emerging policy and supervisory approaches to AI data use in financial services. BIS Financial Stability Institute Insights No. 73.
[7] Capgemini World Property and Casualty Insurance Report 2024 - 8% trailblazers, 62% recognize fraud reduction, 83% predictive models critical
[8] Allied Market Research - Generative AI in Insurance $0.76B 2022 to $14.4B 2032 at 34.4% CAGR
[9] SNS Insider - AI in Insurance $8.63B 2025 to $59.5B 2033 at 27.32% CAGR
[10] Kaggle Vehicle Insurance Fraud Detection - 15,420 records, 923 frauds, 6% imbalance, 36 features

[11] ResearchGate - Detecting insurance claims fraud - AdaBoost/XGBoost 84.5%
[12] ResearchGate - Analytics for Insurance Fraud - XGBoost 89% accuracy F1 87%
[13] PMC - Enhancing fraud detection in auto insurance - CNN KNN 96.62%, CNN-LSTM 89.6%
[14] MDPI - Life Insurance Fraud Detection - RF+Bi-LSTM 80.37% vs XGBoost 86.1%
[15] Gen Re - Generative AI rewrites playbook on insurance fraud
[16] McKinsey - AI-driven underwriting reduces leakage 20%, 51% reduce data entry, 42% faster quotes
[17] Bain - 10-15% revenue uplift, 30-50% leakage reduction $100B+ benefits
[18] EIOPA - Traditional to Generative AI implications, AI Act high-risk
[19] IRDAI - AI-proofing directive for Bima Sugam, bias and privacy safeguards