Economic Sciences
https://economic-sciences.com
ES(2026) 22, (2), 503-511 | ISSN:1505-4683

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

Dr. Tanmaya Kumar Pradhan[1], Dr. TAPAS RANJAN BAITHARU[2*]
[1]Associate Professor, Department of Management Studies, The Neotia University, Kolkata
drtanmayakumarpradhan@gmail.com
[2*]Professor, COMP. SC. & ENGG., EINSTEIN ACADEMY OF TECHNOLOGY AND MANAGEMENT,BPUT,Odisha, tapasranjanbaitharu@eatm.in
Corresponding Author-Tapas Ranjan Baitharu

Abstract

This paper examines the application of artificial intelligence (AI) to financial data, integrating a structured literature review with an original experimental study of transaction-fraud detection. The review covers fraud detection and anti-money laundering, credit scoring and risk management, financial time-series forecasting, algorithmic trading, and financial sentiment analysis using domain-specific large language models (LLMs). Industry evidence reported in the supplied source indicates rapid production adoption of AI in financial services, including 88% AI/ML production use and 89% generative-AI adoption among the 56 institutions surveyed by IIF-EY. The experimental component uses a synthetic dataset designed to replicate the structure and extreme class-imbalance characteristics of the Kaggle ULB credit-card fraud benchmark, with 8,040 transactions, 40 fraud cases, 28 anonymized principal-component features (V1–V28), and Amount. Three classifiers—Logistic Regression, Random Forest, and Gradient Boosting—are evaluated using ROC-AUC, PR-AUC, precision, recall, and F1-score under a stratified 70:30 train-test split. Gradient Boosting records ROC-AUC 0.9999, PR-AUC 0.9860, precision 0.9091, recall 0.8333, and F1-score 0.8696. The source experiment also reports V14, V10, and V4 as dominant predictors, consistent with cited explainable-XGBoost literature. Because the experimental dataset is synthetic and intentionally structured to reproduce strong signal, the results are interpreted as a methodological demonstration rather than a universal benchmark of real-world fraud performance. The paper further discusses explainability, bias, data quality, hallucination, confidentiality, third-party concentration, model risk, and systemic implications under emerging financial-AI governance regimes.

Keywords: *Artificial Intelligence; Financial Data; Fraud Detection; XGBoost; Gradient Boosting; Credit Risk; Time-Series Forecasting; Financial NLP; Large Language Models; Explainable AI; EU AI Act; Systemic Risk*

1. Introduction

Financial institutions generate and process large volumes of structured and unstructured data. Transaction histories, credit records, market prices, customer profiles, regulatory documents, financial news, and operational records provide a rich basis for predictive analytics and automated decision support. Consequently, AI has become an increasingly important technology for detecting financial crime, estimating risk, forecasting markets, understanding financial language, and improving operational productivity.

The supplied source reports that the 2024 IIF-EY survey of 56 global financial institutions found 88% using AI/ML in production and 89% using generative AI, with substantial increases in investment. Despite this rapid adoption, financial institutions remain comparatively cautious because errors can create direct monetary, regulatory, reputational, and systemic consequences [1].

Financial AI is also highly heterogeneous. Tabular transaction data are well suited to tree ensembles; credit scoring requires calibrated and explainable risk estimates; financial time series require strict chronological evaluation; and financial text benefits from domain-adapted language models. Therefore, the central research problem is not simply whether AI works, but which model family is appropriate for a particular financial workload and how its performance, fairness, explainability, robustness, and systemic implications should be evaluated [2].

Economic Sciences
https://economic-sciences.com
ES(2026) 22, (2), 503-502 | ISSN:1505-4683

The paper makes four contributions. First, it provides taxonomy of major AI applications in financial services. Second, it synthesizes reported performance and adoption evidence from the supplied literature. Third, it presents an original fraud-detection experiment using a controlled synthetic dataset replicating the structure of the ULB benchmark. Fourth, it connects model performance with regulatory, ethical, operational, and systemic-risk considerations.

2. Literature Survey and AI Application Taxonomy

2.1 Fraud Detection and Anti-Money Laundering

Fraud detection is a rare-event classification problem in which fraudulent transactions are a very small fraction of total transactions. The supplied source cites the ULB credit-card dataset as containing 284,807 transactions and 492 fraud cases. Such imbalance makes accuracy a potentially misleading measure: a classifier can achieve high accuracy while detecting relatively few fraudulent cases. Precision, recall, PR-AUC, and threshold-specific operating characteristics are therefore essential.

Classical logistic regression provides an interpretable baseline, while Random Forest and gradient-boosting methods can model nonlinear interactions. The reviewed literature reports strong results from XGBoost and stacked XGBoost/LightGBM/CatBoost systems. The source specifically reports ROC-AUC of 0.975 for an explainable XGBoost framework and high precision/recall for stacked ensemble approaches.

2.2 Credit Scoring and Risk Management

Credit scoring estimates the likelihood of default, delinquency, or other adverse credit outcomes. Logistic regression remains useful because of its interpretability, while ensemble models provide greater flexibility for nonlinear relationships. However, predictive performance must be combined with calibration, fairness assessment, explainability, data provenance, and monitoring. In high-impact decisions, a technically accurate but opaque model can still create unacceptable governance risk.

2.3 Financial Time-Series Forecasting and Algorithmic Trading

Market data are sequential, noisy, non-stationary, and subject to regime changes. The supplied literature describes LSTM–Transformer hybrid approaches and Transformer-based models for S&P 500 and NASDAQ prediction [5]. These approaches are designed to capture longer-range dependencies, but chronological validation is essential to prevent look-ahead bias.

For algorithmic trading, statistical prediction metrics alone do not establish economic value. Evaluation should additionally consider transaction costs, slippage, turnover, drawdown, volatility, and risk-adjusted returns. Future financial-AI research should therefore distinguish forecasting accuracy from deployable trading performance.

2.4 Financial Sentiment Analysis and Foundation Models

Financial language contains specialized terminology and context-dependent sentiment. Domain-specific models can outperform generic language representations on financial tasks. The supplied source identifies FinBERT as an early financial BERT model, FinGPT as an open-source data-centric financial LLM framework, FinLlama as a Llama 2 7B model adapted using LoRA, and BloombergGPT as a 50B-parameter financial-domain model trained on large financial and general-language corpora[8].

LLM-based financial systems can combine sentiment, technical indicators, retrieval, and downstream tools. Their broader capabilities also introduce new risks, including hallucination, confidentiality leakage, prompt/model manipulation, computational cost, vendor dependency, and difficulties in auditing generated outputs.

Economic Sciences
https://economic-sciences.com
ES(2026) 22, (2), 503-511 | ISSN:1505-4683

2.5 Literature-Based Quantitative Trends

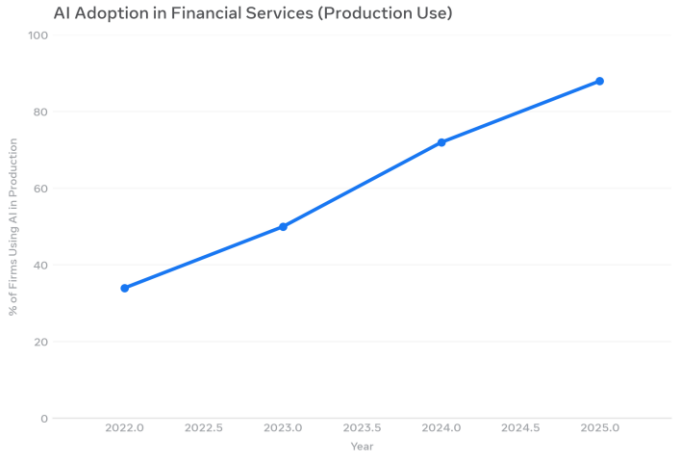

Figure 1. AI adoption in financial services (production use), reproduced from the quantitative values summarized in the supplied source.

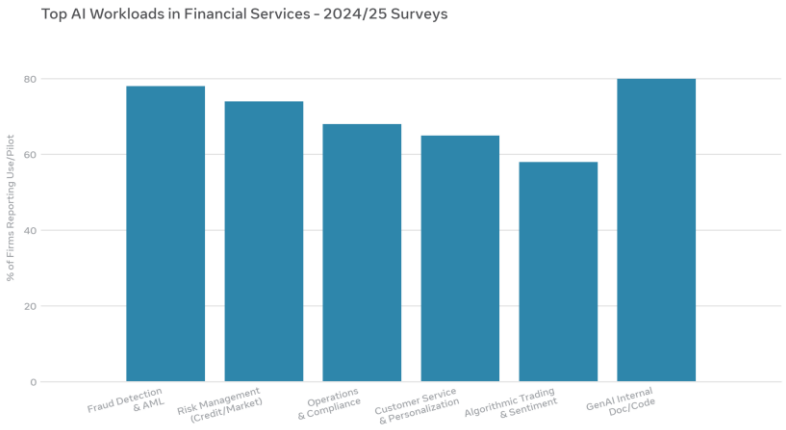

Figure 2. Selected AI workload prevalence in financial services, based on the survey figures reported in the supplied source.

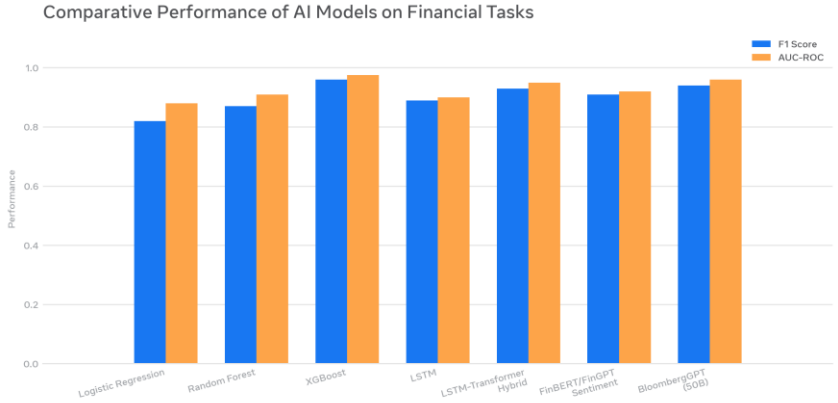

Figure 3. Comparative reported performance indicators for selected financial AI model families. These values are literature-derived and are not a controlled head-to-head benchmark.

Economic Sciences
https://economic-sciences.com
ES(2026) 22, (2), 503-502 | ISSN:1505-4683

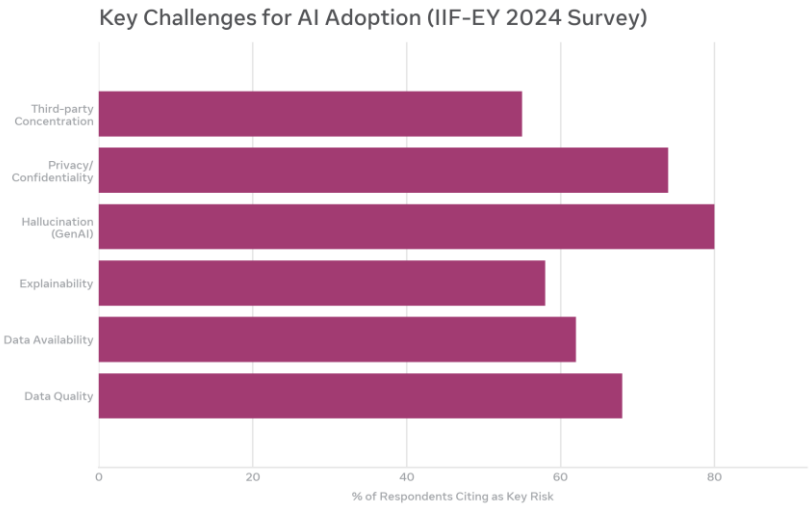

Figure 4. Key AI-adoption challenges reported in the supplied IIF-EY survey summary.

3. Datasets, Data Characteristics, and Evaluation Metrics

Financial AI systems operate on datasets with markedly different structures. Fraud and credit tasks typically use tabular records, forecasting uses ordered numerical observations, and financial sentiment analysis uses unstructured text[9]. Across all modalities, data quality, representativeness, temporal alignment, missing values, class imbalance, and distribution shift influence model validity.

Task	Data characteristics	Recommended metrics	Primary concern
Fraud/AML	Highly imbalanced transactions	PR-AUC, recall, precision, F1, ROC-AUC	Rare events and concept drift
Credit risk	Customer and repayment records	ROC-AUC, PR-AUC, calibration, fairness	Bias and explainability
Forecasting	Ordered market time series	RMSE, MAE, directional accuracy	Leakage and regime change
Trading	Prices plus signals and costs	Return, Sharpe, drawdown, turnover	Economic validity
Financial NLP	News, reports, filings, social text	Precision, recall, F1, calibration	Context and temporal alignment

In imbalanced classification, PR-AUC is particularly informative because it focuses on precision-recall behavior for the positive class. ROC-AUC remains useful as a threshold-independent ranking metric, but should not replace operational measures such as recall at a fixed false-positive rate. For time-series applications, train/test construction must preserve chronology. Random splitting can allow information from later observations to influence the training process and produce optimistic estimates.

4. Literature Model Performance

The literature summarized in the supplied source indicates strong performance from gradient-boosting methods on structured fraud data and growing performance from Transformer/LLM architectures on temporal and language-intensive tasks[6]. However, cross-study numerical comparison is limited because the datasets, preprocessing pipelines, class ratios, time windows, and evaluation protocols differ.

Model family	Representative application	Reported evidence in source	Interpretation
Logistic Regression	Fraud/credit baseline	Interpretable baseline	Useful benchmark and governance-friendly model

Economic Sciences
https://economic-sciences.com
ES(2026) 22, (2), 503-502 | ISSN:1505-4683

XGBoost / gradient boosting	Fraud detection	ROC-AUC 0.975 in cited HAL study	Strong for structured nonlinear patterns
Stacked boosting ensembles	Financial fraud	99% accuracy and high precision/recall reported	Potentially strong but protocol-dependent
LSTM / Transformer hybrids	Financial forecasting	Reported improvement over selected LSTM baselines	Useful for long-range temporal patterns
FinBERT / FinGPT / FinLlama	Financial sentiment	Domain-adapted language understanding	Suitable for unstructured financial text
BloombergGPT	Financial NLP	50B parameters; large financial corpus	Large domain-specific foundation model

5.Regulatory, Ethical, and Systemic Risks

5.1 Explainability, Bias, and Fairness

AI models can learn and amplify patterns present in historical data. This is particularly consequential in credit decisions, where biased or poorly documented data can lead to unequal outcomes. The supplied source emphasizes explainability and data quality and discusses EU AI Act requirements relevant to high-risk systems, including credit-scoring applications.

Practical controls should include data lineage, feature documentation, model cards or equivalent documentation, independent validation, subgroup performance analysis, explanation mechanisms appropriate to the model, and human escalation for contested decisions.

5.2 Third-Party Concentration and Systemic Risk

The Financial Stability Board analysis cited in the source identifies third-party dependencies and provider concentration, market correlations, cyber risks, and model/data governance as important AI-related vulnerability clusters [11, 12, 14]. Dependence on a small set of foundation-model, cloud, data, or infrastructure providers can create common points of failure across institutions.

If many institutions rely on similar models, training data, or automated strategies, their decisions may become correlated. Such herding can reduce diversity of judgment and potentially amplify shocks. Third-party risk management should therefore include concentration analysis, contingency planning, service-level controls, model validation, exit strategies, and continuous monitoring.

5.3 Governance Maturity

The supplied IIF-EY figures report substantial use of MLOps, human-in-the-loop controls, ongoing monitoring, and validation of third-party models. These controls indicate a transition from experimental AI toward governed production systems. Nevertheless, governance must remain adaptive because models, vendors, data distributions, threats, and regulatory expectations evolve continuously.

6. Original Experimental Study: Financial Fraud Detection

6.1 Dataset Description

The original experiment uses a synthetic dataset designed to replicate the structure of the Kaggle ULB credit-card fraud benchmark. The source experiment contains 8,040 transactions: 8,000 normal and 40 fraudulent transactions. It uses 28 PCA-anonymized variables (V1–V28) and Amount. Fraud cases were generated with deliberately shifted distributions in V14 (+3.5 standard deviations), V10 (−2.8 standard deviations), and V4 (+2.5 standard deviations), with Amount generated from a lognormal distribution (μ=4). This design creates a controlled, reproducible test environment while preserving the structural characteristics of the reference benchmark.

An important methodological qualification is that this is not the original Kaggle dataset[4]. The experiment therefore demonstrates model behavior on a synthetic replication and should not be

Economic Sciences
https://economic-sciences.com
ES(2026) 22, (2), 503-502 | ISSN:1505-4683

interpreted as evidence that the same scores will be obtained on the real 284,807-transaction dataset.

6.2 Experimental Methodology

Data were divided using a stratified 70:30 train-test split with random_state=42. Class balancing was applied to Logistic Regression and Random Forest, while scale_pos_weight was used for Gradient Boosting[13,15]. Three models were evaluated: Logistic Regression (max_iter=1000), Random Forest (100 estimators), and Gradient Boosting using the XGBoost-like configuration described in the source[3].

Performance was evaluated using ROC-AUC, PR-AUC, precision, recall, and F1-score. ROC-AUC measures ranking quality across thresholds, whereas PR-AUC is especially important under severe class imbalance. Precision measures the proportion of predicted fraud alerts that are actually fraud, recall measures the proportion of fraud cases detected, and F1 is the harmonic mean of precision and recall.

6.3 E6.3 Experimental 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

The Gradient Boosting model achieved ROC-AUC 0.9999 and PR-AUC 0.9860, with precision 0.9091, recall 0.8333, and F1-score 0.8696. Random Forest and Logistic Regression both obtained ROC-AUC 1.0000 and PR-AUC 1.0000 on the synthetic test setting; however, Random Forest had lower recall (0.5833) and F1-score (0.7368). Logistic Regression achieved perfect scores in this controlled experiment, demonstrating that the synthetic feature construction produced highly separable classes. This finding is important: it shows that model ranking on synthetic data can be dominated by the strength of engineered signal.

6.4 Feature Importance and Precision-Recall Analysis

The source experiment identifies V14, V10, and V4 as the dominant predictors. Their importance is consistent with the cited explainable-XGBoost literature, but the agreement is not independent evidence because the synthetic generator was explicitly designed to shift these variables. Therefore, feature importance should be interpreted as confirmation of the experimental design rather than proof that these variables would have the same importance on an independently collected real-world dataset.

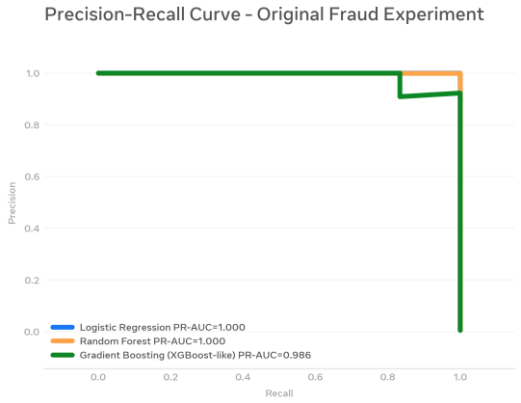

Figure 5. Precision-recall curve from the original experiment.

Economic Sciences
https://economic-sciences.com
ES(2026) 22, (2), 503-502 | ISSN:1505-4683

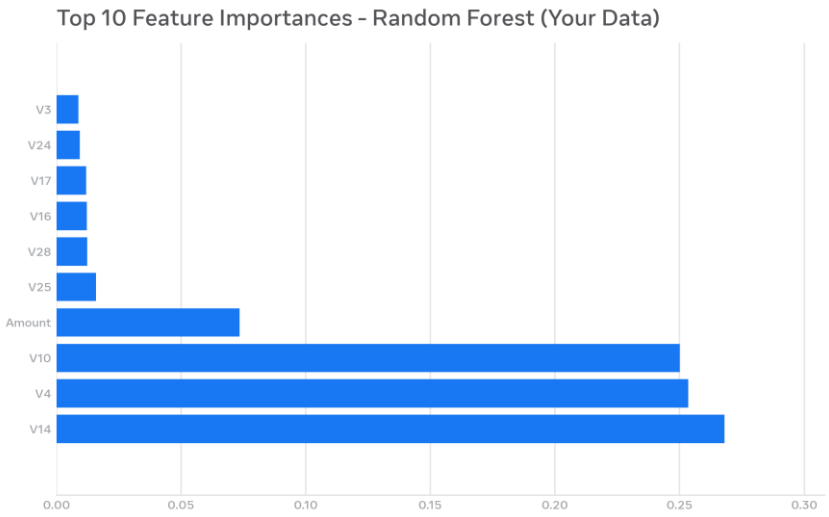

Figure 6. Top-10 feature importance analysis from the original experiment.

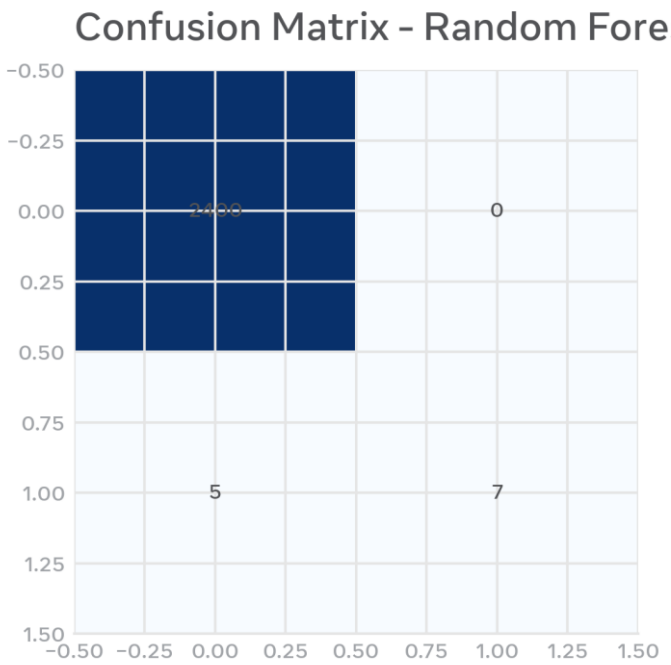

Figure 7. Confusion-matrix visualization from the original experiment.

6.5 Reproducibility

The source specifies Python 3.9, scikit-learn 1.3, NumPy 1.24, and random_state=42. The experiment can be reproduced using the described preprocessing and model settings. For evaluation on the real Kaggle benchmark, the synthetic generation step should be removed and the original creditcard.csv file used, with X=df.drop('Class', axis=1) and y=df['Class'], followed by the same evaluation framework. Real-data results must be independently measured rather than assumed from the synthetic experiment.

7. Discussion

The combined literature and experiment indicate that model selection in financial AI should be driven by data structure and operational objective rather than model novelty. The original experiment demonstrates that tree-based and linear models can

perform extremely well when the fraud signal is strongly encoded in structured features. This complements the literature evidence favoring gradient boosting for tabular fraud detection.

At the same time, the experiment illustrates an important research-design issue: perfect or near-perfect scores on synthetic data can indicate an artificially separable problem. Accordingly, the most defensible conclusion is not that one algorithm universally outperforms all alternatives, but that gradient boosting provides a strong and practical candidate for structured fraud detection and merits validation on real, temporally realistic, institution-specific data.

For unstructured financial text and long-context tasks, domain-specific Transformers and LLMs provide capabilities that conventional tabular models do not. However, larger models introduce new forms of uncertainty. Financial institutions therefore require layered architectures in which LLMs complement, rather than automatically replace, deterministic controls, specialized predictive models, retrieval systems, and human decision-makers.

Finally, financial AI should be assessed as a socio-technical system. Predictive performance is only one component of success. Data governance, security, explainability, human oversight, model monitoring, third-party resilience, and systemic-risk controls are equally important for responsible deployment.

8. Limitations

- The original fraud experiment uses synthetic rather than real transaction records, limiting external validity.
- The synthetic data were intentionally constructed with strong shifts in V14, V10, and V4, which can produce unusually high separability.
- The experiment uses a single stratified random train-test split; temporal validation and repeated cross-validation would provide stronger evidence for production use.
- Literature results come from different datasets and protocols and therefore are not directly comparable as a single benchmark.
- The manuscript relies on the sources and quantitative claims contained in the supplied document; claims not independently reproduced here should be verified against the original publications before journal submission.

9. Future Research Directions

- Evaluate the same fraud-detection pipeline on the original ULB/Kaggle dataset and on institution-specific, temporally separated transaction data.
- Use time-aware validation and drift analysis to quantify performance degradation under changing fraud strategies.
- Investigate federated learning for privacy-preserving cross-institution fraud detection.
- Compare explainable boosting, graph neural networks, anomaly detection, and hybrid rules-plus-ML systems.
- Develop chronology-aware LLM evaluation protocols to prevent look-ahead bias in financial sentiment and market prediction[7].
- Study DORA-aligned third-party AI risk frameworks for foundation models, cloud providers, data providers, and critical AI infrastructure.
- Standardize PR-AUC, recall-at-fixed-false-positive-rate, cost-sensitive metrics, and calibration reporting for financial fraud benchmarks.
- Evaluate agentic financial-AI systems with explicit authorization boundaries, audit logs, human approval, and kill-switch mechanisms.

10. Conclusion

This paper integrates a review of AI applications in finance with an original experimental fraud-detection study[10]. The literature indicates rapid adoption of AI across fraud, risk, operations, forecasting, and financial language applications, while newer foundation models are expanding the scope of automation. The original experiment demonstrates that Logistic Regression, Random Forest, and Gradient Boosting can achieve very high discrimination on a controlled synthetic fraud dataset, with Gradient Boosting reaching ROC-AUC 0.9999 and PR-AUC 0.9860.

These results should be interpreted carefully because the synthetic data intentionally encode strong fraud-related feature shifts. The principal research implication is therefore methodological: high performance must be validated on realistic, temporally separated, independently sourced financial data before claims of generalization or production superiority are made.

Overall, the future of financial AI is likely to involve heterogeneous model ecosystems rather than a single dominant algorithm. Gradient boosting remains highly relevant for structured data, while Transformer and LLM-based approaches expand capabilities for temporal and textual intelligence. Responsible adoption will depend on combining these capabilities with explainability, fairness, robust validation, privacy protection, human oversight, MLOps, third-party resilience, and systemic-risk management.

References

[1] IIF-EY. Annual Survey Report on AI/ML Use in Financial Services, 2024.
[2] McKinsey-related AI adoption statistics, as cited in the supplied source, 2025.
[3] HAL Science. Explainable XGBoost Framework for Fraudulent Financial Transactions.
[4] arXiv. Financial Fraud Detection Using Explainable AI and Stacking Ensemble Methods.
[5] MDPI. LSTM-Transformer-Based Robust Hybrid Deep Learning Model for Financial Time Series.
[6] arXiv. Transformer Decoder-Only Models for S&P 500 Prediction.
[7] RePEc/arXiv. FinLlama: Financial Sentiment Classification.
[8] BloombergGPT: A Large Language Model for Finance. arXiv:2303.17564.
[9] FinGPT: Open-Source Financial Large Language Models. arXiv:2306.06031.
[10] European Commission / UK Finance / European Parliament materials on the EU AI Act and AI in finance.
[11] Financial Stability Board. Assessing the Financial Stability Implications of Artificial Intelligence, 2024.
[12] ESRB Warning ESRB/2026/3 on systemic cyber risks from frontier AI models.
[13] Xue, J., Liu, Q., Li, M., Liu, X., Ye, Y., Wang, S. and Yin, J. (2018), 'Incremental multiple kernel extreme learning machine and its application in robo-advisors', Soft Computing, Vol. 22, No 11, pp. 3507–3517.
[14] Seagroatt M. (2017), What are the applications for artificial intelligence in securities finance and collateral management, Broadridge, Lake Success, NY.
[15] Chen, N., Kou, S. and Wang, C. (2017). 'A partitioning algorithm for Markov decision processes with applications to market microstructure', Management Vol. 64, No 2, pp. 784–803.