ESP-JETA
ESP Journal of Engineering & Technology Advancements
ISSN: 2583-2646/ Volume 6 Issue 4 October-December 2026
Paper Id: JETA-V6I4P107/ Page No: 68-75

Original Article

AI/ML-Powered Real-Time Banking Fraud, AML, and Credit-Risk Intelligence Platform with Generative AI Investigation Copilot and Explainable Governance

Reddi Kishore
Bharathiar University, Coimbatore

Received Date: 16 October 2026 Revised Date: 20 October 2026 Accepted Date: 24 October 2026

Abstract: *Financial institutions still have three distinct technology estates for financial crime detection, anti-money-laundering (AML) surveillance, and credit-risk assessment, each with their own data pipeline, model inventory and investigation queue. This separation dilutes detection as the same customer identity, device fingerprint and counterparty network are signal across all three estates and exacerbates governance as each estate is responsible for meeting the model-risk, explainability and fairness requirements. This paper summarizes the studies conducted in the period 2020-2026 regarding machine learning in real-time banking risk intelligence and why it's important to converge to a single platform. The four strands are: streaming and graph-based fraud detection; network-aware AML transaction monitoring; machine-learned credit scoring and early-warning systems; and the new frontier of using large language models as copilots for investigations. It is observed that the three domains have a common feature, namely the feature substrate, but they differ significantly in the latency budget, the quality of the labels and the regulatory exposure. Also, they are faced with a severe class-imbalance problem and a concept-drift problem. Then, explainability techniques are translated into supervisory responsibilities, and a layered reference architecture is presented where a shared feature store and entity graph are inputs to domain-specific decision engines, a retrieval-grounded copilot supports human investigators, and a governance plane captures explanations, drift metrics and human overrides. Unresolved challenges include privacy-preserving cross-institution learning, adversarial robustness, hallucination control in case narratives generated, and a lack of common benchmarks for assessing unified platforms.*

Keywords: *Machine Learning, Fraud Detection, Anti-Money Laundering, Credit Risk, Large Language Models, Explainable AI, Model Governance, Real-Time Analytics*

I. INTRODUCTION

The bank's decision window has gone from days to milliseconds with instant payment schemes, and risks which were once scored in overnight batch are now scored while a transaction is “in flight.” The systematic review of financial fraud detection finds that there has been a turning point in the field, from expert rule sets to supervised and unsupervised learning, due to the volume and velocity of card-not-present and real-time payment traffic [1]. The same trend is observed in other areas of financial crime, where investigators are increasingly making decisions based on the network of counterparties rather than on individual payments [2] and in credit risk, where statistical scorecards are increasingly being superseded by ensemble and deep models [3].

Even though this is often the path taken, the three functions are usually constructed and managed separately. Typically, a card-fraud engine, an AML transaction monitor and a credit scorecard use different versions of the same core banking data, have different definitions of "customer" and "counterparty," and alert different case-management queues. This is three times as expensive. First, evidence is disjointed: evidence from a mule account can be seen by the AML graph, but not by the fraud engine scoring the payment that is made through the mule account. Secondly, effort is duplicated, as feature pipelines, drift monitors and validation packs are rebuilt three times. Third—and most importantly—the human investigator is forced to reconcile three incomplete stories in the context of a regulatory clock.

There are two developments that make convergence not just desirable, but practical. Large language models have become proficient enough at summarising financial text to generate summaries of the material from a case, write an investigative narrative, and respond to analysts' questions across institutional knowledge bases [4]. Meanwhile, explainable AI has gone from


This is an open access article under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/2.0/)

a research novelty to an operational necessity, and there is now a well-established vocabulary of local and global attribution techniques that can be interrogated by supervisors and validators [5].

This review makes four contributions. It brings together the literature of the last six years (2020-2026) of fraud, AML and credit risk and presents this material in a comparative format to reveal where the domains overlap and where they are truly distinct. It evaluates the evidence base for generative AI investigation copilots and discerns between what has been achieved and what is speculative. It is a mapping of explainability techniques to concrete governance obligations, instead of a generic virtue. Finally, it outlines a layered reference architecture for a unified platform and pinpoints the research gaps that hinder the end-to-end validation of such platforms at present.

II. REVIEW METHODOLOGY AND SCOPE

This is a narrative review, not a formal meta-analysis, because the primary studies report incommensurable metrics on proprietary datasets which cannot be pooled statistically. Query strings were a mix of domain (fraud detection, anti-money laundering, transaction monitoring, credit scoring, credit risk) and method (machine learning, deep learning, graph neural network, large language model, explainable AI) and, if applicable, deployment (real time, streaming, model governance). Items were kept if they reported an empirical evaluation on financial transaction or credit data, systematic synthesis of such work or a governance/explainability framework applicable to regulated financial decisioning. No purely theoretical articles without any financial evaluation, vendor white papers and non-indexed preprints were included.

The literature that was retained is structured around four analytical dimensions that are common across all three risk categories: the data substrate (tabular, sequential or relational); the learning paradigm (supervised, unsupervised, hybrid or generative); the latency regime (batch, near-real-time or in-flight); and the accountability regime (what has to be explained, to whom, and when). These axes are used to organize Sections III-VIII, and they are the basis for the comparative synthesis in Table I. Section IX presents the architecture as a synthesis proposed by this review, but not as a reported deployment, and is offered as a research agenda, not as validated engineering.

III. FROM RULE ENGINES TO LEARNING SYSTEMS

The three risk domains each took a different path and at different speeds to machine learning, and the vestige of those paths remains in their architectures. Fraud detection was first and it went the furthest since there is a relatively short feedback loop: if a chargeback or customer dispute occurs, it gets flagged as a fraud within weeks.

Table 1: Comparative Characteristics of the Three Risk Domains

Dimension	Fraud detection	AML transaction monitoring	Credit-risk intelligence
Decision latency	10–100 ms, in-flight	Minutes to overnight batch	Seconds (application) to monthly (behavioural)
Label source and lag	Chargebacks, disputes; days to weeks	Investigation outcome, SAR filing; months	Default or delinquency event; 6–24 months
Positive-class prevalence	~0.1–1% of transactions	<0.1% of alerts confirmed	1–10% of accounts
Dominant model families	Gradient boosting, sequence models, graph neural networks	Scenario rules plus supervised alert scoring, network analytics	Logistic regression, gradient boosting, constrained neural networks
Primary data structure	Event stream plus entity graph	Entity graph plus aggregated flows	Panel of applicant and behavioural attributes
Drift profile	Fast, adversarial	Slow typology shift, punctuated by new schemes	Macroeconomic, cyclical
Dominant failure cost	Customer friction from false positives	Investigator overload; regulatory penalty for misses	Unfair or unexplained denial; capital misstatement
Principal accountability driver	Payment-scheme liability rules	AML statutes and supervisory examination	Consumer-credit law, capital and provisioning standards

Deterministic velocity rules and threshold rules are still in production as a guard rail, but the discriminative work is now performed by gradient-boosted ensembles and, increasingly, models that take the sequence of transactions as input, not just a row [1]. AML progressed at a slower rate. Because its rules are partially defined by regulation, its labels are not received until after an investigation, and the suppressed alert has legal implications, supervised learning was not used as a replacement for the existing scenario engine, but was used as an alert-ranking layer above it. The slowest was credit risk, not because of technical issues, but because scorecards are adversarial regulated entities: the rejected applicant has the right to an explanation, and thus explainability limited the class of models long before it became a general research area [3].

A common observation in all three is that an increasingly expressive model does not yield as much benefit as a more carefully designed feature space and a more carefully distributed label space. The hybrid design, which integrates unsupervised outlier scores with a supervised classifier, is a case in point: the unsupervised component provides information about behaviour for which no label exists, and using its output as extra features results in more reliable precision than using it as a classifier replacement [6]. The argument for a common feature substrate is similar: a feature that is calculated for fraud is often useful for laundering and for default. Governance evolved alongside. At the beginning of the period, commentators suggested that financial regulators should shift from a reactive approach to supervising the results of AI use, to a proactive approach of mandating controls during the design phase, predicting many of the elements that would later be included in statutory AI regulations [7]. It is this change in stance that makes a modern platform need to view explanation, monitoring and audit as first-class runtime components and not as documentation created after the fact.

IV. REAL-TIME FRAUD DETECTION

The key challenge of fraud detection is that the model must make its prediction before the funds are transferred, making feature engineering as much of a systems problem as it's a statistical problem. Time series aggregation and feature engineering: Nearly every reported improvement in card fraud is due to the ability to summarise a cardholder's recent activity, such as amount relative to a rolling mean, time since last activity at the same merchant category, geographic velocity. Early practice relied on hand-made aggregates, but in more recent practice it has been demonstrated that automated methods can be used to model behaviour sequences directly. Multi-perspective hidden Markov models, computed separately over the cardholder and the terminal and over both genuine and fraudulent histories, generate features which complement the expert aggregates and improve precision-recall performance without manual specification [8]. The operational implication is that these features need to be calculated at scoring time, with point-in-time correctness – which is the job of a streaming feature store.

Unbalanced and undocumented behaviour: When positives are below one per cent, the only meaningful metrics for evaluation are precision based, like average precision, and precision at a fixed alert budget, and accuracy and area under ROC curve are misleadingly flattering. Since labels are also delayed, a purely supervised system is structurally unable to learn about new attack patterns during the time window in which they are most harmful. To overcome this, hybrid architectures have been proposed which use the unsupervised outlier scores, calculated at transaction, card and global level, as engineered features and are fed to the supervised classifier, which has been demonstrated to achieve better detection than either of the individual components [6].

Relational signal: Compromised cards are grouped on merchants, mule accounts are grouped on devices and application fraud is grouped on shared identity attributes. To leverage this structure, graph neural networks propagate representations along transaction edges, and spatio-temporal attention mechanisms add the notion of weighting neighbours by their relational closeness and their timeliness, which is important when a merchant is clean one month and compromised the next [9]. A systematic review of graph NNs for financial fraud reveals that they have consistently outperformed tabular baselines, that most published evaluations rely on static snapshots of the graph, that inductive evaluation with previously unseen nodes is under-reported, and that inference times on large, dynamic graphs are the main challenge for in-flight deployment [10].

Drift: Fraud drift is an adversarial process and not an incidental process, meaning that countermeasures are designed to alter the behaviour of attackers. Practical systems thus combine frequent retraining with explicit population-stability monitoring on the input features, which allows them to detect degradation from the inputs before it is reflected in the delayed labels.

V. AML AND TRANSACTION MONITORING

The issue in AML is not so much missed detection as it is alert volume; conventional scenario engines have false-positive rates that are often reported to be in excess of ninety percent and the issue is the capacity of investigators, not the recall of the

model. The most popular intervention is to keep the scenario rules visible to the regulator, and introduce a learned model to prioritize the resulting alerts based on the likelihood that investigation will substantiate the suspicion. The analysis of Norwegian banking data showed that supervised models trained on the historical results of investigations performed significantly better than rule thresholds at identifying reportable cases, and were also compatible with the current control framework since no reportable case was silently suppressed [11]. It's politically significant as well as technically significant because it means that machine learning can be deployed without the institution needing to say that a model took the place of a prescribed control. There is a separate pipeline for Sanctions/Politically-exposed-person/adverse-media screening, where fuzzy name matching between transliteration variants yields a lot of false positives. It has been demonstrated that supervised classification can materially reduce the amount of data requiring manual review when applied to disposition of historical watch-list hits, an unusually tractable application because the labels (analyst dispositions) are abundant and also occur rapidly [12]. Laundering is not a single transaction, but a pattern of structure, across accounts and over time, that's what makes it laundering. Thus, graph computing is the organising paradigm for financial crime analytics, which can detect communities in counterparty networks, search for motifs that represent known typologies, and propagate risk from a known bad actor to its neighbourhood [2]. The same review mentions the practical challenges: cross-institutional structure is not visible by any single bank and graph construction at national payment volumes is computationally challenging. A lack of shared labelled data is a limitation in AML research, as suspicious-activity reports cannot be published. Practitioner elicitation studies indicate that there is a constant disconnect between the metrics used to optimise detection models and the metrics used to assess institutions: published models optimise detection metrics on synthetic or highly anonymised data, while institutions are assessed on the productivity of investigators, explainability to examiners, and completeness of the audit trail [13]. Therefore, any credible unified platform should be judged on the investigation workflow, rather than just on classifier metrics.

VI. CREDIT-RISK INTELLIGENCE

The most mature of the three is credit risk, and the most constrained, as the outputs are adverse decisions on identifiable individuals and the models are used to input regulatory capital and provisioning. Ensemble methods, and gradient boosting specifically, have been found to consistently lie at the performance frontier on tabular data pertaining to credit applicants over the past 30 years, and logistic regression has been maintained in production for much longer than its accuracy would suggest [3]. The survey also revealed a methodological gap in the literature: many comparative studies are carried out on small public datasets, baselines are tuned less rigorously than proposed methods, and margins of discrimination are reported without calibration, and hence overestimate what can be achieved in production with deep architectures on conventional credit data. A controlled study on multiple real credit portfolios revealed that deep networks were not consistently superior to well-tuned gradient boosting with standard tabular features and that deep networks are useful only in situations where raw sequential or unstructured inputs are available, not necessarily in tabular scoring per se [14]. The architectural implication for a unified platform, with regard to the credit engine and deep models, is that the credit engine can only gain from the deep models if it can take the same sequences of transactions and relational context that the fraud and AML engines already have. Credit risk is legally mandated explanation, not just desirable, and this has resulted in some of the most tangible work in post hoc interpretation in finance. In the case of gradient-boosted credit models, Shapley-value attribution has been demonstrated to provide borrower-level reason codes that are stable for adverse-action notices, and also to be interpretable in terms of risk archetypes, thereby providing the same technique with both customer-facing and portfolio-management applications [15].

New forward-looking impairment criteria moved the analysis from origination (default) to impairment of the credit risk during the life of the exposure. This is where the shared substrate really shines: salary-inflow interruption, an increase in gambling merchant codes or a counterparty already marked by the AML graph are early-warning features that the credit engine can't calculate based on its own data extract.

VII. THE GENERATIVE AI INVESTIGATION COPILOT

Decision comes from investigation, and alerts come from detection. The copilot takes care of the second, more costly and more time-intensive half of the pipeline, in which an analyst brings evidence together from various systems and crafts a defensible narrative. Financial text is sufficiently different from general corpora that the use of domain adaptation measurably helps. Pretraining a transformer encoder on financial disclosures and analyst communications yielded significant improvements over general-purpose models on financial information-extraction tasks, with the greatest improvements in the regime of AML typology classification and adverse-media screening [16] where there was least labelled data. This is extended to generative use in a survey of large language models in finance, which lists applications in sentiment analysis, report summarisation, information extraction and question answering, and identifies as the main challenges to evaluation and factual grounding, not

capability per se [4]. An ungrounded model cannot be used for investigative purposes: the analyst must have statements that can be traced back to a transaction record, a policy clause or a case note. Retrieval-augmented designs overcome this by basing generation on documents retrieved from institutional stores such as alert history, customer file, sanctions determinations, or internal typology guidance, thus allowing each generated assertion to be traced back to the source. Another reason to consolidate detection engines is that the retrieval corpus is the same feature store, entity graph and case archive that is fed into the detection engines. There is evidence for four uses. Firstly evidence assembly – compiling the customer's history of alerts, network neighbourhood and previous dispositions into a structured briefing. Second, narrative drafting: Creating a draft of a SAR or fraud case summary based on facts, which the analyst edits and signs. Third, triage support: providing in natural language an explanation for why a model scored an alert high, using the attribution output and not making up an explanation. Fourth, knowledge access: answering procedural questions according to internal policy, decreasing the escalation load on senior staff. The failure mode is fluid production. Hallucination in natural language generation has been surveyed and two types of hallucination have been identified: intrinsic hallucination, when the output contradicts the input, and extrinsic hallucination, when the output states something that wasn't stated in any input [17]; and it has been noted that standard generation metrics do not detect hallucination reliably. Extrinsic hallucination in a filed report is a compliance failure and is a type of hallucination in a regulated investigation. Three controls follow: each generated claim should be accompanied by a reference to a retrieved record; numerical facts should be calculated using a deterministic tool and inserted, and never generated; and the copilot should be limited to drafting and explaining, but not disposing of an alert or to making a credit decision. The human sign-off is not an analyst courtesy, but a way to keep accountability "locatable.

VIII. EXPLAINABLE AI AND MODEL GOVERNANCE

There are several different requirements of explainability in the banking sector, each aimed at a different audience, and a platform that attempts to satisfy one of them will not do a good job of satisfying any of them. The customer inquires why a decision was made against him and what it would take to alter the decision.

Table 2: Explainability Techniques Mapped to Governance Obligations

Technique	Output	Primary audience	Obligation addressed	Principal limitation
Shapley-value attribution	Signed per-feature contribution to one decision	Customer, investigator	Adverse-action reason codes; alert justification	Ambiguous under correlated features; sensitive to background data
Counterfactual explanation	Minimal change that flips the decision	Customer, relationship manager	Actionable recourse in credit denial	Feasibility and non-discrimination hard to guarantee
Global importance and interaction analysis	Ranked drivers and dependence structure	Model validator	Conceptual soundness review	Aggregate view can hide segment-level failure
Surrogate or constrained transparent model	An inherently readable approximation or model	Supervisor, credit policy	Demonstrable model transparency	Fidelity gap; accuracy cost
Graph evidence subgraph	The neighbourhood and paths driving a network alert	AML investigator	Evidentiary basis for a filed report	No settled evaluation standard for subgraph quality
Drift and stability monitoring	Population and performance shift over time	Validator, supervisor	Ongoing monitoring and recalibration duty	Delayed labels defer detection of true decay
Retrieval citation trail	Source record behind each generated statement	Investigator, auditor	Traceability of generated case narratives	Correct citation does not guarantee correct inference

Investigator inquires what evidence prompted an alert for the case to be worked efficiently. The model validator queries the model for defensible relationships and whether the model is stable across segments. The supervisor asks who is responsible, what was measured and what was done by the institution when performance dropped. This distinction between transparent and post-hoc, and local and global, aligns nicely with the audiences: local post-hoc is for customer and investigator, global is for validator and supervisor [5]. Shapley-value attribution has become the de facto standard for tabular financial models, and the ability to compute exact polynomial-time explanations for tree ensembles, which also provide consistent global importance and interaction effects by aggregating local explanations [18] has made it operationally viable at transaction volumes.

The remaining challenges are well known: ambiguously distributed credit for correlated features, changing the attribution of credit for background-distribution choice, and an explanation of a model is not an explanation of the world. While attribution weights are not actionable, for adverse-action notices counterfactuals of the form “approval would have happened at this income and this utilisation” are more actionable, although the problem of generating minimal, feasible and non-discriminatory counterfactuals remains unsolved. Explainability in the context of financial risk management is proposed to be a part of the model lifecycle, not as validation documentation, and is linked to robustness testing and data governance and continuous monitoring as a single control system [19]. More general reviews of trustworthy AI make the same point, but in the broader context of fairness, privacy and accountability being co-equal requirements that interact and sometimes conflict, since a fairness constraint may decrease accuracy or a privacy control may prevent the collection of attributes required for fairness testing [20]. These compromises are not something the platform can solve, but have to be made clear, decided by a governance body and documented.

IX. A UNIFIED REFERENCE ARCHITECTURE

The architecture in Fig. 1 follows directly from the comparative analysis: what the three domains share belongs in one layer, what differs belongs in separate engines, and what must be demonstrated to third parties belongs in a plane that spans everything.

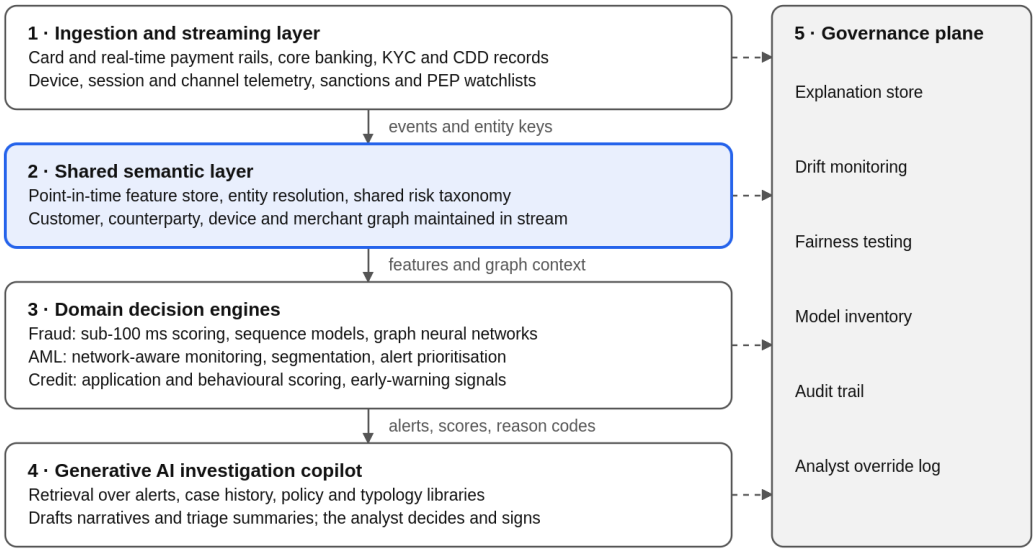


Figure 1: Layered Reference Architecture for a Unified Banking Risk-Intelligence Platform

The ingestion layer maps events related to payments, core banking, onboarding and telemetry to a common schema. The load-bearing part is the shared semantic layer, a point-in-time feature store that ensures that a model at scoring time only contains what was known at that time, entity resolution to ensure that one customer has one identity across channels, and an entity graph that is continuously maintained. This layer is the enabler for a credit early-warning model to take a AML network signal, or a fraud engine to observe that a beneficiary account was opened yesterday [2].

The decision engines are distinct due to the incompatibilities between their latency and label regimes (Table I). The fraud engine scores in flight, the AML engine ranks alerts generated by the scenario layer, on a batch cadence, against investigation

outcomes, and the credit engine provides application and behavioral scoring subject to interpretability constraints. The model and release cycle are not what they have in common, it is the substrate under them.

The copilot is located above the engines and below the decision. It gets from the same substrate, drafts and explains, and hands all material judgements to an accountable human [4]. The governance plane exists across all four layers and is a record of why each decision was made, what was monitored for the input drift, what was monitored for the performance drift, what fairness tests were conducted, and what each analyst override was a supervised signal for the next retraining cycle [19].

The design is defensible by two boundaries. The Copilot never dismisses an alert or approves credit. The governance plane is a runtime dependency of the engines, not a reporting attachment to them: a decision emitted without a stored explanation is an operationally bad decision, not a documentation gap.

X. OPEN CHALLENGES AND RESEARCH DIRECTIONS

Most important financial-crime indicator is between banks, and no bank can see it. The hard problems are non-independently-distributed data across participants, incentive design, and that a shared model still leaks information about its training population [21] – federated learning is the obvious answer, and has been formulated specifically for open-banking settings. The question of whether federated training is compatible with the joint-liability structure of AML is an open legal, as well as technical question. It is not possible to publish labels for confirmed laundering, which is why there is so much research in AML using synthetic transactions. Synthetic financial data, however, suggests that fidelity on the marginal distributions does not imply fidelity on the rare tail distributions which detection models are designed to detect, and that privacy guarantees are often claimed but not proven [22]. So, models with good performance on synthetic benchmarks should not be assumed to be valid until they are tested on production data. Adaptive opponents lie and money-launder. Most published evaluations make the assumption that the data generating process is stationary, and thus are an upper bound on performance. There is a lack of evaluation protocols that involve adaptive attackers and monitoring that identifies probing behaviour. Proxies for protected characteristics are rich with transaction and network characteristics: merchant categories represent lifestyle, geography represents ethnicity, network membership represents community. However, the fairness tests for tabular scorecards do not apply to graph models, which depend on the attributes of an individual's neighbours for their score. There is no accepted benchmark for an investigation copilot. Outcomes of interest are: Investigator throughput, quality of decisions, percentage of unsupported assertions in written narratives, and degree of automation bias in reviewers. Each needs an evaluation protocol for human subjects which the current literature does not provide. Finally, the model risk management philosophy is that models can be tested individually. A platform that shares a common feature store to feed three engines, and a copilot that provides coupling otherwise impossible by single-model validation: every change to the feature definition is reflected simultaneously across fraud scores, laundering alerts and credit decisions. The current state of validation methodology of coupled model estates is an open problem.

XI. CONCLUSION

Fraud detection, anti-money laundering monitoring and credit-risk assessment have, for institutional reasons, been thought of as separate disciplines, not analytical. It turns out that they have the same fundamental problem: the same entities, the same behavioural histories, the same counterparty graph and the same common challenges of extreme class imbalance, delayed labels and non-stationary data. What they do differ in is latency budget and label economics and regulatory exposure, and there's a difference in those, which is where they differ, but it's not a difference in platforms, it's a difference in decision engines. There are three conclusions that are supported by the evidence. The first is that it's probably going to be more beneficial than more model architecture changes in both of the domains, because each domain now has access to signal that it couldn't compute before. Second, generative AI can be used in the investigation process rather than the decision process by using retrieval grounding and human sign-off for the generation of narratives and the explanation of model output to control the tendency of generative AI to make fluent but unsupported assertions. Third, explainability and governance are not a compliance layer that's tacked onto the models after they are built, but runtime components whose output is as important as the risk scores. Evaluation, not modelling capability, is the main hurdle to progress. No common metrics for network-aware financial-crime detection on realistic data, no common measure for whether an investigation copilot enhances the decision quality or just speeds it up, and no common validation methodology for a coupled model estate that spans three regulated domains simultaneously. While those holes remain, unified risk-intelligence platforms will appear a logical possibility, but will not have been tested in practice and the discipline's promised gains will be greater than what institutions can demonstrate to their supervisors.

REFERENCES

[1] A. Ali, S. Abd Razak, S. H. Othman, T. A. E. Eisa, A. Al-Dhaqm, M. Nasser, T. Elhassan, H. Elshafie, and A. Saif, "Financial fraud detection based on machine learning: A systematic literature review," *Applied Sciences*, vol. 12, no. 19, art. 9637, Sep. 2022, doi: 10.3390/app12199637.

[2] E. Kurshan and H. Shen, "Graph computing for financial crime and fraud detection: Trends, challenges and outlook," *International Journal of Semantic Computing*, vol. 14, no. 4, pp. 565–589, Dec. 2020, doi: 10.1142/S1793351X20300022.

[3] X. Dastile, T. Celik, and M. Potsane, "Statistical and machine learning models in credit scoring: A systematic literature survey," *Applied Soft Computing*, vol. 91, art. 106263, Jun. 2020, doi: 10.1016/j.asoc.2020.106263.

[4] Y. Li, S. Wang, H. Ding, and H. Chen, "Large language models in finance: A survey," in *Proc. 4th ACM Int. Conf. on AI in Finance (ICAIF)*, New York, NY, USA, Nov. 2023, pp. 374–382, doi: 10.1145/3604237.3626869.

[5] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82–115, Jun. 2020, doi: 10.1016/j.inffus.2019.12.012.

[6] F. Carcillo, Y.-A. Le Borgne, O. Caelen, Y. Kessaci, F. Oblé, and G. Bontempi, "Combining unsupervised and supervised learning in credit card fraud detection," *Information Sciences*, vol. 557, pp. 317–331, May 2021, doi: 10.1016/j.ins.2019.05.042.

[7] J. Truby, R. Brown, and A. Dahdal, "Banking on AI: Mandating a proactive approach to AI regulation in the financial sector," *Law and Financial Markets Review*, vol. 14, no. 2, pp. 110–120, 2020, doi: 10.1080/17521440.2020.1760454.

[8] Y. Lucas, P.-E. Portier, L. Laporte, L. He-Guelton, O. Caelen, M. Granitzer, and S. Calabretto, "Towards automated feature engineering for credit card fraud detection using multi-perspective HMMs," *Future Generation Computer Systems*, vol. 102, pp. 393–402, Jan. 2020, doi: 10.1016/j.future.2019.08.029.

[9] D. Cheng, X. Wang, Y. Zhang, and L. Zhang, "Graph neural network for fraud detection via spatial-temporal attention," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 8, pp. 3800–3813, Aug. 2022, doi: 10.1109/TKDE.2020.3025588.

[10] S. Motie and B. Raahemi, "Financial fraud detection using graph neural networks: A systematic review," *Expert Systems with Applications*, vol. 240, art. 122156, Apr. 2024, doi: 10.1016/j.eswa.2023.122156.

[11] Balaji, R. (2025) 'Request-Level Anomaly Detection and Risk-Adaptive Authorization for Microservice APIs', ESP Journal of Engineering & Technology Advancements, 5(2), pp. 342-351. doi: 10.56472/25832646/JETA-V5I2P137.

[12] M. Jullum, A. Løland, R. B. Huseby, G. Ånonsen, and J. Lorentzen, "Detecting money laundering transactions with machine learning," *Journal of Money Laundering Control*, vol. 23, no. 1, pp. 173–186, Jan. 2020, doi: 10.1108/JMLC-07-2019-0055.

[13] M. Alkhalili, M. H. Qutqut, and F. Almasalha, "Investigation of applying machine learning for watch-list filtering in anti-money laundering," *IEEE Access*, vol. 9, pp. 18481–18496, 2021, doi: 10.1109/ACCESS.2021.3052313.

[14] B. Oztas, D. Cetinkaya, F. Adedoyin, M. Budka, G. Aksu, and H. Dogan, "Perspectives from experts on developing transaction monitoring methods for anti-money laundering," *IEEE Access*, vol. 12, pp. 86229–86242, 2024, doi: 10.1109/ACCESS.2024.3414848.

[15] B. R. Gunnarsson, S. vanden Broucke, B. Baesens, M. Óskarsdóttir, and W. Lemahieu, "Deep learning for credit scoring: Do or don't?," *European Journal of Operational Research*, vol. 295, no. 1, pp. 292–305, Nov. 2021, doi: 10.1016/j.ejor.2021.03.006.

[16] N. Bussmann, P. Giudici, D. Marinelli, and J. Papenbrock, "Explainable machine learning in credit risk management," *Computational Economics*, vol. 57, no. 1, pp. 203–216, Jan. 2021, doi: 10.1007/s10614-020-10042-0.

[17] A. H. Huang, H. Wang, and Y. Yang, "FinBERT: A large language model for extracting information from financial text," *Contemporary Accounting Research*, vol. 40, no. 2, pp. 806–841, May 2023, doi: 10.1111/1911-3846.12832.

[18] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, "Survey of hallucination in natural language generation," *ACM Computing Surveys*, vol. 55, no. 12, art. 248, Dec. 2023, doi: 10.1145/3571730.

[19] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, "From local explanations to global understanding with explainable AI for trees," *Nature Machine Intelligence*, vol. 2, no. 1, pp. 56–67, Jan. 2020, doi: 10.1038/s42256-019-0138-9.

[20] S. Fritz-Morgenthal, B. Hein, and J. Papenbrock, "Financial risk management and explainable, trustworthy AI," *Frontiers in Artificial Intelligence*, vol. 5, art. 779799, Feb. 2022, doi: 10.3389/frai.2022.779799.

[21] D. Kaur, S. Uslu, K. J. Rittichier, and A. Durresi, "Trustworthy artificial intelligence: A review," *ACM Computing Surveys*, vol. 55, no. 2, art. 39, Mar. 2023, doi: 10.1145/3491209.

[22] G. Long, Y. Tan, J. Jiang, and C. Zhang, "Federated learning for open banking," in *Federated Learning: Privacy and Incentive*, Lecture Notes in Computer Science, vol. 12500, Cham, Switzerland: Springer, 2020, pp. 240–254, doi: 10.1007/978-3-030-63076-8_17.

[23] S. A. Assefa, D. Dervovic, M. Mahfouz, R. E. Tillman, P. Reddy, and M. Veloso, "Generating synthetic data in finance: Opportunities, challenges and pitfalls," in *Proc. 1st ACM Int. Conf. on AI in Finance (ICAIF)*, New York, NY, USA, Oct. 2020, art. 44, doi: 10.1145/3383455.3422554.