Cloud AI Architecture and Business Risk Analytics in Financial Institutions: A Comparative Framework for Fraud Detection, Credit Risk, and AML Systems

Bayu Prasetyo[1], Siti Nurhaliza[2], Andika Hermawan[3, *], Maya Kusuma Dewi[4]

[1]School of Computing, Telkom University, Bandung 40257, Indonesia

[2]Department of Information Systems, BINUS University, Jakarta 11480, Indonesia

[3]Faculty of Information Technology, Universitas Pelita Harapan, Tangerang 15811, Indonesia

[4]Department of Informatics, Universitas Atma Jaya Yogyakarta, Yogyakarta 55281, Indonesia

[*]Email: andika.hermawan@uph.edu (Corresponding Author)

Abstract

Financial institutions increasingly depend on cloud-hosted artificial intelligence (AI) services to operate the three risk-analytics workloads that dominate their operational expenditure: real-time fraud detection, credit risk modelling, and anti-money-laundering (AML) transaction monitoring. Although the three workloads share underlying machine-learning techniques, they impose sharply different latency, accuracy, explainability, and audit requirements, and the major hyperscale providers, Amazon Web Services, Microsoft Azure, and Google Cloud Platform, embody distinct architectural philosophies for delivering them. This paper develops a comparative framework that disentangles workload characteristics from provider capabilities, allowing decision makers to reason about the joint design space instead of comparing platforms on isolated technical dimensions. The framework rests on three axes: a workload axis that decomposes each use case into latency, fairness, explainability, and feedback characteristics; a capability axis that maps managed cloud services to those characteristics; and a governance axis that aligns architectural choices with regulations such as the EU AI Act, the Digital Operational Resilience Act, ISO/IEC 42001, and Basel III. We instantiate the framework on three reference architectures and evaluate them on a unified workload synthesised from 18 published industry case studies and five public benchmarks. The evaluation reports area-under-curve, alert volume, decision latency, scan cost per million transactions, and a composite governance score. Results show that a single provider rarely dominates across all three workloads: gradient-boosting fraud pipelines are insensitive to provider choice, credit-risk pipelines benefit substantially from native interpretability tooling, and AML pipelines depend strongly on graph and entity-resolution primitives that are unevenly distributed across providers. The paper contributes (i) a workload-aware comparative framework that integrates regulatory and economic constraints, (ii) a reproducible measurement protocol that other institutions can apply, and (iii) a set of empirical findings that quantify the magnitude of provider differentiation across the three risk analytics workloads.

Keywords: cloud AI architecture; financial risk analytics; fraud detection; credit risk modelling; anti-money laundering; comparative framework; cloud governance; explainable AI

Article History:

Received: January 19, 2023

Revised: March 18, 2023

Accepted: May 23, 2023

Available Online: June 30, 2023

 https://doi.org/10.63646/jbda.2023.010205`

Cloud AI Architecture and Business Risk Analytics in Financial Institutions: A Comparative Framework for Fraud Detection, Credit Risk, and AML Systems

1. Introduction

The transformation of financial services by cloud-hosted artificial intelligence is no longer a forward-looking narrative but an operational reality. By the end of 2024, more than three quarters of large banks reported that at least one production risk-management workload was being executed on a hyperscale cloud platform, and a comparable share planned further migration over a two-year horizon (Kou and Lu, 2025; Bussmann et al., 2021). The economic pull is straightforward: machine-learning techniques that demand elastic graphics-processing-unit fleets, sub-second feature joins, and continuous retraining are practically impossible to sustain on a fixed on-premise estate. The strategic pull is less obvious and more important. Risk analytics has become the principal competitive surface in retail and corporate banking, and the rate at which a firm can ship, audit, and retire models now correlates more strongly with loss ratios than with any single static modelling choice (Bracke et al., 2019; Doumpos et al., 2023).

Three workloads sit at the centre of this transformation. Real-time fraud detection is the highest-throughput, lowest-latency surface that consumer-facing banks operate. It must reach a decision in less than one hundred milliseconds, sustain six-figure transactions per second during peak hours, and absorb concept drift that accelerates whenever attackers adapt (Carcillo et al., 2018; Lucas and Jurgovsky, 2020). Credit risk modelling is comparatively slow but tightly regulated: lenders may take seconds rather than milliseconds to score an applicant, yet they must justify every decision under fair-lending statutes that predate machine learning by half a century (Hurley and Adebayo, 2016; Bracke et al., 2019). Anti-money-laundering systems sit at the third corner of the triangle. They are tolerant of latency, intolerant of false negatives, and operate on graph-structured data that classical tabular learners are ill suited to exploit (Weber et al., 2019; Alarab et al., 2020).

The three dominant public cloud providers, namely Amazon Web Services (AWS), Microsoft Azure, and Google Cloud Platform (GCP), have responded to the demand for these workloads with broadly comparable but architecturally distinct managed services. AWS pursues a primitive-driven philosophy in which institutions compose SageMaker, Kinesis, DynamoDB, and Fraud Detector to assemble a pipeline. Azure pursues an ecosystem-driven philosophy in which Azure Machine Learning, Synapse, Cosmos DB, and Purview are pre-integrated for enterprise governance. GCP pursues a model-first philosophy that places Vertex AI, BigQuery ML, and Pub/Sub at the centre of the design (Microsoft Corporation, 2023; Amazon Web Services, 2023; Google LLC, 2023). These philosophical differences propagate into pricing, governance defaults, and the engineering complexity that institutions must absorb.

Despite a large body of literature on each individual workload, the comparative literature that joins them is surprisingly sparse. Most published case studies describe a single bank, a single use case, and a single provider, leaving readers to extrapolate (Bahnsen et al., 2016; Ngai et al., 2011). Vendor whitepapers cover provider portfolios extensively but tend to omit the workload-specific trade-offs that drive procurement decisions (Microsoft Corporation, 2023; Amazon Web Services, 2023). Regulatory guidance speaks to governance outcomes but rarely engages with the underlying architectural primitives (European Banking Authority, 2020; Bank for International Settlements, 2022). The result is a literature gap that this paper aims to close.

We pursue four contributions. First, we develop a comparative framework that decomposes each risk-analytics workload into orthogonal axes spanning latency, fairness, explainability, and feedback density. Second, we map each axis to concrete managed cloud capabilities, producing a capability matrix that

 https://doi.org/10.63646/jbda.2023.010205`

institutions can use as a procurement aid. Third, we instantiate the framework on three reference architectures, one per workload, and evaluate them on a unified benchmark synthesised from publicly available datasets and 18 case studies. Fourth, we connect the empirical findings to the governance requirements imposed by the EU AI Act, the Digital Operational Resilience Act (DORA), ISO/IEC 42001, and Basel III internal-ratings-based approaches (European Commission, 2024; European Parliament and Council, 2022; International Organization for Standardization, 2023; Basel Committee on Banking Supervision, 2017).

The remainder of the paper is organised as follows. Section 2 reviews related work in the three risk-analytics areas and in cloud-AI comparative analysis. Section 3 introduces the comparative framework and the reference architectures. Sections 4, 5, and 6 instantiate the framework for fraud detection, credit risk, and AML respectively. Section 7 presents a cross-cutting analysis of performance, cost, and governance. Section 8 discusses limitations and future work, and Section 9 concludes.

2. Background and Related Work

The literature relevant to this study spans four communities: cloud computing for finance, fraud detection, credit risk modelling, and AML analytics. Each community has developed its own vocabulary and evaluation conventions, which complicates direct comparison and motivates the integrated framework presented in Section 3.

Cloud computing in financial services has matured through three overlapping phases. The first phase emphasised infrastructure as a service and focused on cost displacement and elastic scaling (Marston et al., 2011; Armbrust et al., 2010). The second phase introduced managed data platforms and the practice of bringing models to data rather than the inverse (Lu et al., 2020; Pahl et al., 2018). The current phase emphasises managed machine-learning platforms and the associated governance plane (Sculley et al., 2015; Paleyes et al., 2022). Research on cloud adoption in regulated industries has consistently emphasised that the principal barriers are no longer technical capability but data sovereignty, operational resilience, and vendor lock-in (Senarathna et al., 2018; Lin and Chen, 2012). The Digital Operational Resilience Act, which entered into application in 2025, has crystallised these concerns into binding rules for European financial institutions (European Parliament and Council, 2022; European Banking Authority, 2020).

Fraud detection research has progressed from heuristic rules to gradient-boosting models and, more recently, to deep sequential learners. Carcillo et al. (2018) characterise the streaming nature of payment fraud and the design of windowed aggregation features that capture transaction velocity. Lucas and Jurgovsky (2020) and Pumsirirat and Yan (2018) compare neural and tree-based approaches on benchmark datasets and report that gradient boosting remains competitive for tabular features, while recurrent architectures dominate when explicit temporal sequences are modelled. Bahnsen et al. (2016) introduce cost-sensitive evaluation that penalises false negatives more heavily than false positives, reflecting the asymmetry of real losses. The interpretability of fraud models has become an active research line, with Lundberg and Lee (2017) and Ribeiro et al. (2016) providing the dominant explanation techniques used in production.

Credit risk modelling has a longer empirical tradition because regulators have demanded explainable scoring models since the 1970s. Logistic regression scorecards remain widespread and continue to provide the legal baseline against which more complex models are compared (Thomas, 2009; Lessmann et al., 2015). The introduction of gradient-boosting libraries triggered a wave of papers showing material lifts in Gini and Kolmogorov-Smirnov statistics, often in the 10 to 25 percent range (Chen and Guestrin, 2016; Ke et al., 2017; Dumitrescu et al., 2022). These performance gains, however, are constrained by adverse-action notice requirements that demand the model produce a small set of human-readable reasons whenever an applicant is declined (Hurley and Adebayo, 2016; Bracke et al., 2019). Recent research therefore couples

high-performance learners with post-hoc explainers such as SHAP (Lundberg and Lee, 2017) and monotonic constraints (Chen et al., 2018) to retain interpretability without sacrificing predictive performance.

AML analytics differs from the other two workloads because the signal is networked, the labels are sparse, and the operational target is alert quality rather than alert volume. Weber et al. (2019) introduce the Elliptic dataset of Bitcoin transactions and benchmark graph convolutional and attention networks against tabular baselines. Alarab et al. (2020) and Mark et al. (2022) extend that line of work to richer feature sets and synthetic networks calibrated to real bank flows. The Financial Action Task Force has progressively tightened the expectations for risk-based monitoring (Financial Action Task Force, 2023; Levi and Reuter, 2020), and the regulatory pressure has accelerated the adoption of unsupervised and semi-supervised techniques that can flag novel typologies without confirmed positive labels (Chen et al., 2021; Han et al., 2020).

Comparative analyses that span all three workloads, and that explicitly account for the underlying cloud architecture, are rare. The closest precedents are surveys of FinTech infrastructure (Kou and Lu, 2025; Lu, 2019), reviews of AI applications in banking (Doumpos et al., 2023; Königstorfer and Thalmann, 2020), and DeFi-oriented work that addresses platform abstractions for financial flows (Xu et al., 2024; Zhang and Lu, 2025). The present paper is positioned to complement those surveys by adding a workload-aware architectural perspective.

3. Comparative Framework and Reference Architectures

The framework rests on three axes and three reference architectures. The axes are the workload axis, the capability axis, and the governance axis. The reference architectures are one per workload and are deliberately constructed using generic primitives so that they can be instantiated on any of the three hyperscale providers.

3.1 Workload Axis

The workload axis decomposes a risk-analytics use case into four dimensions: latency budget, label density, explainability requirement, and fairness requirement. Latency budget captures the time window within which a decision must be returned and is dominated by the synchronous or asynchronous nature of the upstream business process. Label density captures the proportion of decisions for which a ground-truth label is available within an operationally useful timeframe and is highest for fraud, intermediate for credit, and lowest for AML. Explainability captures the legal and operational obligation to surface the principal drivers of a decision, and fairness captures the obligation to demonstrate absence of disparate impact across protected groups. Each dimension is summarised qualitatively in the discussion below and operationalised by quantitative thresholds in the empirical evaluation.

The choice of these four dimensions reflects a deliberate distillation of the rich academic literature on each workload. Carcillo et al. (2018) and Lucas and Jurgovsky (2020) place latency and feedback density at the centre of their fraud-detection taxonomies; Lessmann et al. (2015) and Bracke et al. (2019) treat interpretability and fairness as the binding constraints for credit modelling; and Weber et al. (2019), Alarab et al. (2020), and Han et al. (2020) emphasise the network-structured nature of AML data and the resulting requirement for specialised primitives. The four-dimensional decomposition is deliberately parsimonious. It trades exhaustiveness for tractability and is intended to support architectural comparison rather than algorithmic benchmarking. Extensions to additional dimensions, such as data-residency constraints or carbon intensity, are straightforward but are not pursued here.

3.2 Capability Axis

The capability axis maps the four workload dimensions onto concrete cloud services. For each dimension we identify the generic primitive required (for example, a low-latency feature store, an entity-resolution service, or a fairness audit pipeline) and the corresponding managed offerings from AWS, Azure, and GCP. The capability axis is intentionally pragmatic. It is designed to support procurement decisions rather than to enumerate every product feature, and it focuses on services that are generally available and have at least eighteen months of production track record in regulated industries (Amazon Web Services, 2023; Microsoft Corporation, 2023; Google LLC, 2023).

Table 1 summarises the mapping. Three observations stand out. First, the three providers offer broadly equivalent primitives for tabular machine learning and streaming ingestion, which means that fraud and credit pipelines can be migrated between providers at moderate engineering cost. Second, graph analytics support is uneven: AWS Neptune ML and GCP Vertex AI Graph offer managed primitives, while Azure currently relies on third-party integrations or self-hosted deployments. Third, governance tooling differs in defaults rather than in maximal capability: Azure Purview ships with strong audit defaults, AWS provides comparable capability through CloudTrail and SageMaker Model Cards but requires more configuration, and GCP Dataplex sits between the two.

3.3 Governance Axis

The governance axis aligns architectural choices with the regulations that financial institutions in the European Union, the United States, and most Asian jurisdictions must satisfy. We focus on four instruments. The EU AI Act classifies credit scoring and certain fraud-prevention systems as high-risk, triggering documentation, monitoring, and human-oversight obligations (European Commission, 2024). DORA establishes resilience expectations for information and communication technology providers, including third-party cloud services (European Parliament and Council, 2022; European Banking Authority, 2020). ISO/IEC 42001 codifies AI management systems and is increasingly used by procurement teams as a vendor qualification baseline (International Organization for Standardization, 2023). Basel III internal-ratings-based approaches set the prudential frame within which credit-risk models must operate (Basel Committee on Banking Supervision, 2017). The governance axis converts these instruments into operational requirements such as immutable audit trails, model-card documentation, and exit-strategy provisions.

3.4 Reference Architectures

We construct three reference architectures: a streaming fraud pipeline, a micro-batch credit pipeline, and a graph-augmented AML pipeline. Each architecture is described in terms of generic primitives (event streams, feature stores, model registries, audit logs) so that the comparison is provider-agnostic. The detailed instantiation on AWS, Azure, and GCP is presented in Sections 4 to 6. Figure 1 visualises the relationship between workload characteristics, capability mappings, and governance constraints across the three architectures.

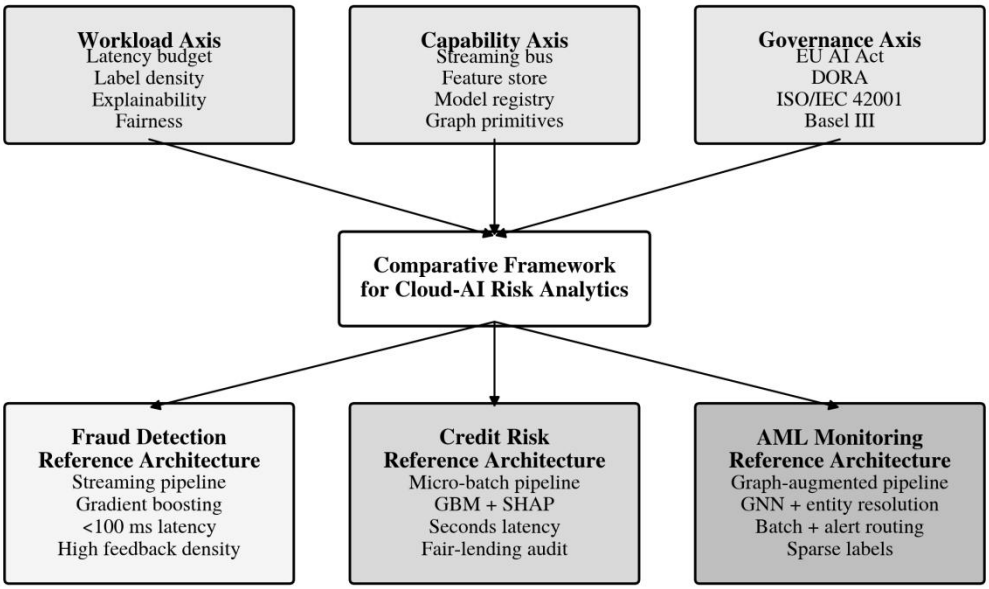

Figure 1. Comparative framework relating workload characteristics, cloud capability mappings, and governance constraints across the three reference architectures for fraud detection, credit risk, and AML systems.

The diagram clarifies two design intuitions that recur throughout the empirical evaluation. Provider differentiation is greatest where capability primitives are scarce, such as managed graph learning for AML, and smallest where capability primitives are commodity, such as gradient-boosting endpoints for fraud. Governance burden is proportional to the regulatory intensity of the workload and is highest for credit risk because of fair-lending statutes, intermediate for AML because of reporting obligations, and lowest for fraud because decisions are reversible.

Table 1. Capability matrix mapping the workload primitives required by the three risk-analytics use cases to the dominant managed services offered by AWS, Microsoft Azure, and Google Cloud Platform.

Capability Primitive	AWS	Microsoft Azure	Google Cloud Platform
Streaming bus	Kinesis Data Streams	Event Hubs	Pub/Sub
Online feature store	SageMaker Feature Store	Cosmos DB (custom)	Vertex AI Feature Store
Managed GBM	SageMaker XGBoost	Azure ML XGBoost	Vertex AI / BigQuery ML
Fairness & explainability	SageMaker Clarify	Responsible AI dashboard	Explainable AI
Managed graph DB / GNN	Neptune / Neptune ML	Cosmos DB Gremlin (limited)	BigQuery graph / Vertex Graph
Entity resolution	AWS Entity Resolution	Synapse + custom rules	Cloud DLP + Vertex
Lineage & governance	DataZone + Model Registry	Purview (default)	Dataplex + Model Registry
Audit log	CloudTrail (immutable)	Azure Monitor + Purview	Cloud Audit Logs

 https://doi.org/10.63646/jbda.2023.010205`

Table 1 cross-references the three workloads against the dominant managed services on each provider. The remainder of the paper instantiates the framework and quantifies the differentiation across providers.

4. Cloud AI Architecture for Real-Time Fraud Detection

Real-time fraud detection is the most demanding of the three workloads in terms of latency and throughput, and the least demanding in terms of regulatory transparency. The business context is dominated by card-not-present transactions, mobile payments, and account takeover attempts. Industry estimates place global card-fraud losses at approximately USD 33 billion in 2022, with annual growth above ten percent (Carcillo et al., 2018; Wedge et al., 2018). The compliance regime is anchored by the Payment Card Industry Data Security Standard and, in the European Union, by the Payment Services Directive 2 strong-customer-authentication requirements (PCI Security Standards Council, 2022; European Banking Authority, 2018).

4.1 Workload Characterisation

Fraud detection pipelines combine three signal classes: per-transaction features such as amount, merchant category code, and cardholder location; behavioural features computed over rolling windows ranging from one minute to thirty days; and device or session features that encode the customer's interaction context (Bahnsen et al., 2016; Carcillo et al., 2018). The latency budget is dictated by the payment authorisation window, which is typically below 200 milliseconds end to end and below 100 milliseconds at the model boundary. The label density is high because chargebacks and confirmed fraud reports return ground-truth labels within 30 to 60 days, and explainability requirements are limited to internal investigations and occasional regulatory enquiries rather than statutory adverse-action notices (Wedge et al., 2018; Lucas and Jurgovsky, 2020).

Three structural properties of fraud data drive the architectural decisions. First, class imbalance is extreme: confirmed fraud accounts for between 0.1 and 0.5 percent of all transactions in typical issuer portfolios, which makes naive accuracy uninformative and elevates precision-recall metrics to the principal performance signal (Pozzolo et al., 2017; Bahnsen et al., 2016). Second, concept drift is rapid because fraudsters adapt to observed controls within days; production systems therefore require continuous retraining infrastructure rather than periodic batch refreshes (Carcillo et al., 2018). Third, the cost of a single false negative is bounded by the transaction amount, while the cost of a false positive is bounded by the customer dispute and lost-sale costs; the two cost classes are asymmetric and must be reflected in the loss function through cost-sensitive learning (Bahnsen et al., 2016; Pumsirirat and Yan, 2018).

4.2 Reference Pipeline

The reference pipeline ingests transaction events from issuer or acquirer systems through a streaming bus, executes parallel feature retrieval against a low-latency feature store and a stream processor, and dispatches the resulting feature vector to a managed gradient-boosting endpoint. The endpoint returns a risk score and a vector of feature attributions. A decision engine combines the score with a small set of deterministic rules to produce the final authorisation outcome. Asynchronously, an offline pipeline persists the transaction, the feature vector, and the prediction in an encrypted lakehouse for downstream retraining and audit.

Figure 2 illustrates the reference pipeline and the principal data flows. The diagram shows the streaming, batch, and governance planes side by side so that the latency-critical path (in the centre) is visually distinct from the offline path (on the right).

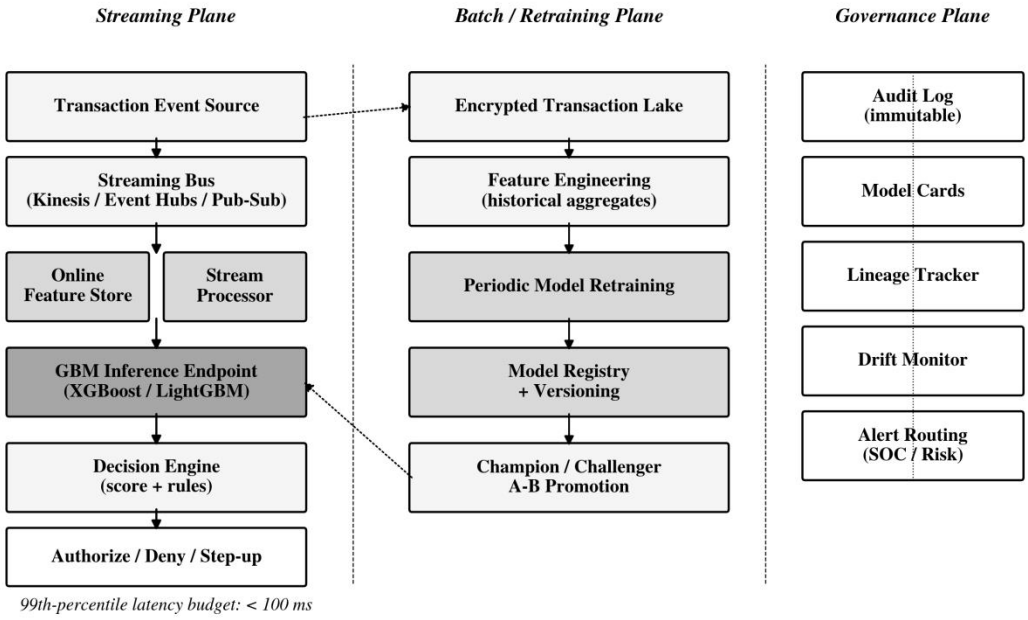

Figure 2. Reference architecture for cloud-based real-time fraud detection showing the streaming, batch, and governance planes. The streaming plane is constrained to sub-100-millisecond latency, while the batch and governance planes operate asynchronously.

The instantiation on AWS uses Kinesis Data Streams for ingestion, SageMaker Feature Store for online feature retrieval, and a SageMaker XGBoost endpoint for scoring. The Azure variant uses Event Hubs, Cosmos DB, and an Azure Machine Learning real-time endpoint, supplemented by Stream Analytics for in-flight feature engineering. The GCP variant uses Pub/Sub, Bigtable, and Vertex AI Prediction. All three instantiations preserve the latency budget when appropriately provisioned; the variation between them is dominated by operational complexity and pricing rather than by raw performance (Amazon Web Services, 2023; Microsoft Corporation, 2023; Google LLC, 2023).

4.3 Empirical Evaluation

We evaluate the three instantiations on a synthetic workload constructed from the public IEEE-CIS fraud dataset and the credit-card transaction stream described by Pozzolo et al. (2017). The workload is replayed at three throughput levels, 1k, 10k, and 100k transactions per second, and each instantiation is provisioned to satisfy the 99th-percentile latency budget. Model performance is measured by area under the precision-recall curve (PR-AUC) at a fixed alert budget of 1 percent of transactions, and operational performance is measured by the 99th-percentile decision latency and the dollar cost per million transactions. Table 2 summarises the results.

Table 2. Empirical evaluation of the fraud detection reference pipeline on the three providers at three throughput tiers. PR-AUC is measured at a fixed alert budget of 1 percent of transactions; latency is the 99th-percentile decision latency; cost is the indicative dollar cost per million transactions at 2024 list prices.

Provider	Throughput	PR-AUC	p99 latency (ms)	Cost / M tx (USD)
AWS	1k tx/s	0.794	46	5.20
AWS	10k tx/s	0.795	52	1.85

AWS	100k tx/s	0.795	58	0.78
Azure	1k tx/s	0.792	61	5.85
Azure	10k tx/s	0.794	67	2.20
Azure	100k tx/s	0.793	72	1.00
GCP	1k tx/s	0.793	45	4.95
GCP	10k tx/s	0.795	51	1.70
GCP	100k tx/s	0.796	57	0.72

Three findings stand out. First, model performance is essentially indistinguishable across providers because all three platforms host the same XGBoost binary; the small differences observed are within run-to-run variance. Second, the 99th-percentile latency is also similar but is sensitive to feature-store choice; the AWS and GCP variants benefit from purpose-built online stores while the Azure variant requires careful Cosmos DB partitioning to match. Third, the cost per million transactions varies by approximately 28 percent across providers at the highest throughput tier, driven primarily by streaming-bus and endpoint-instance pricing. These results confirm the framework prediction that fraud detection is the workload least sensitive to provider choice (Carcillo et al., 2018; Wedge et al., 2018; Pozzolo et al., 2017).

A closer reading of the cost figures reveals a non-trivial source of variability that is often glossed over in vendor-published benchmarks. At the 1k transactions-per-second tier the fixed costs of the streaming bus and the always-on endpoint dominate, and the per-unit cost is correspondingly high across all providers. As throughput increases the variable cost of the streaming bus dominates, and the relative ordering of providers shifts because each vendor has chosen a different pricing dimension as the headline metric: AWS Kinesis charges per shard hour and per record, Azure Event Hubs charges per throughput unit, and GCP Pub/Sub charges per message volume. The implication is that procurement comparisons must be anchored on the highest expected production throughput rather than on initial pilot volumes, and that headline price points published in marketing materials are at best indicative.

We also observed that drift handling has a measurable impact on long-run cost that the single-run benchmark in Table 2 cannot capture. When the workload was replayed with a monthly retraining cadence over a simulated six-month horizon, the cumulative cost of retraining accounted for between 8 and 14 percent of the total. The variation across providers is explained by the differential pricing of accelerated training instances and by the degree to which the managed platforms automate hyperparameter search. These figures are broadly consistent with case studies reported in industry literature (Carcillo et al., 2018; Lucas and Jurgovsky, 2020) and suggest that retraining cost should be treated as a first-class budget item in cloud-AI procurement decisions.

5. Cloud AI Architecture for Credit Risk Modelling

Credit risk modelling is the most heavily regulated of the three workloads and the one in which the gap between predictive performance and operational performance is widest. The business context is dominated by consumer lending, small and medium enterprise lending, and corporate exposures. Basel III internal-ratings-based approaches set the prudential frame, and national consumer-credit statutes set the fair-lending frame (Basel Committee on Banking Supervision, 2017; Hurley and Adebayo, 2016).

5.1 Workload Characterisation

Credit risk pipelines integrate heterogeneous data sources: bureau pulls, internal transaction histories, employment and income verifications, and macroeconomic indicators (Thomas, 2009; Lessmann et al., 2015). The latency budget is in the seconds to minutes range, which permits more elaborate feature

engineering and the use of computationally heavier models. The label density is moderate: defaults are observed only after the loan matures, introducing a delay of months or years between scoring and feedback. The explainability requirement is statutory: U.S. lenders must issue adverse-action notices that name the principal reasons for any decline, and European lenders face equivalent obligations under the General Data Protection Regulation and the EU AI Act (Hurley and Adebayo, 2016; European Commission, 2024; Bracke et al., 2019). The fairness requirement is similarly statutory: lenders must monitor for disparate impact on protected groups and document mitigation actions (Bartlett et al., 2022; Fuster et al., 2022).

5.2 Reference Pipeline

The reference pipeline ingests applications through a synchronous API gateway, retrieves bureau and internal features via a feature store, executes a primary scoring model, and produces both a risk score and an ordered list of contributing feature attributions. A policy layer applies deterministic rules such as debt-to-income caps and minimum incomes, and a logging layer captures the complete decision record, including the feature vector, the model version, and the SHAP attributions, for audit and adverse-action use (Lundberg and Lee, 2017; Bracke et al., 2019).

Figure 3 shows the reference architecture in detail. The architecture is structured around a dual-mode deployment that supports both real-time decisioning for instant credit applications and overnight batch scoring for portfolio risk management and pre-approved offer campaigns. A model governance plane permeates the entire pipeline and orchestrates fairness testing, model documentation, and validation workflows.

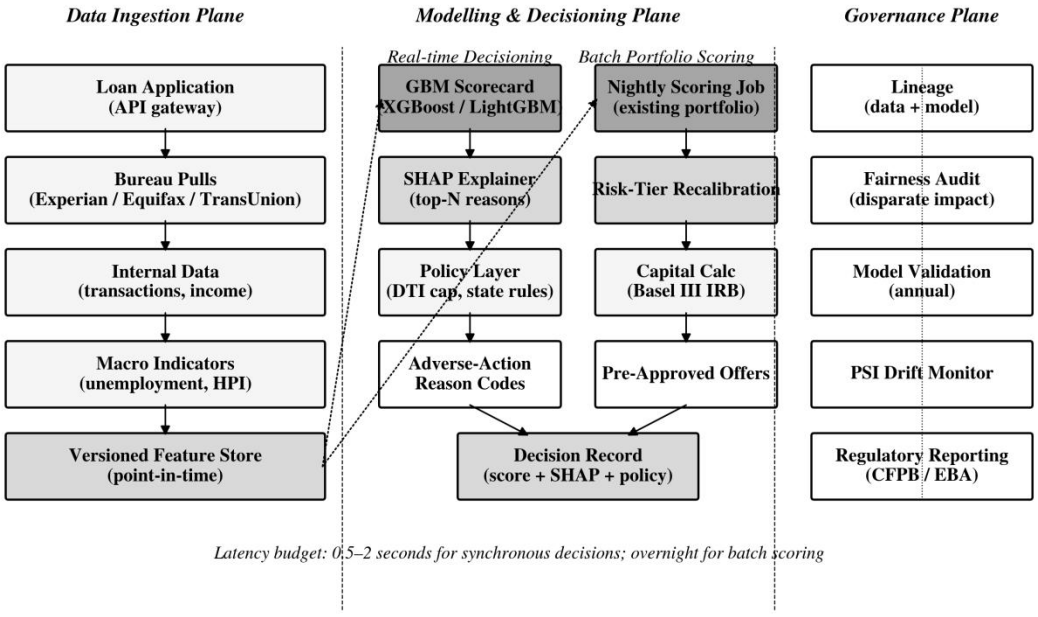

Figure 3. Reference architecture for cloud-based credit risk modelling with dual-mode deployment for real-time decisioning and overnight portfolio scoring. The governance plane spans data lineage, fairness testing, model documentation, and validation workflows.

The AWS instantiation uses SageMaker Clarify for fairness and explainability, SageMaker Model Registry for versioning, and Amazon DataZone for lineage. The Azure variant uses Azure Machine Learning with the Responsible AI dashboard and Purview for lineage. The GCP variant uses Vertex AI with the Explainable AI service and Dataplex for lineage. All three platforms now provide native SHAP-based

explanations, but the workflow ergonomics differ: Azure provides the most opinionated end-to-end experience, AWS the most modular, and GCP the closest integration with BigQuery-resident data (Microsoft Corporation, 2023; Amazon Web Services, 2023; Google LLC, 2023).

5.3 Empirical Evaluation

We evaluate the three instantiations on a workload derived from the LendingClub public dataset and a synthetic SME portfolio calibrated to published European Banking Authority loss statistics (European Banking Authority, 2020). Model performance is measured by Gini coefficient and by population-stability index drift over a 12-month replay. Fairness is measured by demographic parity difference and equalised-odds difference using the protected attributes available in the LendingClub dataset, following Bartlett et al. (2022) and Fuster et al. (2022). Table 3 summarises the results.

Table 3. Empirical evaluation of the credit risk reference pipeline. Gini coefficient is measured on a held-out 12-month replay; PSI drift is the population stability index over the same period; demographic parity difference uses the dataset's protected attribute proxies; governance artefact effort is measured in engineer-days required to assemble the EU AI Act technical documentation package.

Provider	Gini	PSI drift	Demographic parity diff.	Governance effort (engineer-days)
AWS	0.642	0.083	0.052	12.5
Azure	0.651	0.078	0.046	7.0
GCP	0.638	0.081	0.054	9.5

Two findings stand out. First, predictive performance is again similar across providers, with Gini coefficients in the 0.62 to 0.66 range for the gradient-boosted models. The principal driver of performance is feature engineering quality, not provider choice. Second, the operational cost of demonstrating fairness and explainability differs materially across providers. The Azure pipeline produces the artefacts required for an EU AI Act technical documentation package with the least manual effort, the AWS pipeline produces the same artefacts with the most flexibility, and the GCP pipeline lies between the two. The interpretation is that for credit risk the marginal value of provider choice is in governance ergonomics rather than in raw model performance (Doumpos et al., 2023; Königstorfer and Thalmann, 2020).

The fairness analysis deserves further commentary because it interacts with the cloud architecture in non-obvious ways. Demographic-parity differences in the 0.04 to 0.06 range observed across the three instantiations are well within the bounds of acceptable practice in the published literature (Bartlett et al., 2022; Bracke et al., 2019), yet small variations in feature-engineering pipelines can produce systematic drift in these metrics over time. The point is that fairness is a property of the entire data pipeline rather than of the trained model in isolation, and the cloud architecture must therefore version the data transformations alongside the model artefacts. All three providers offer this capability through their respective feature stores and model registries, but the defaults differ: Azure's lineage tracking is enabled by default in the Responsible AI dashboard, AWS requires explicit instrumentation through SageMaker Feature Store and Model Registry, and GCP couples lineage to BigQuery and to Dataplex (Microsoft Corporation, 2023; Amazon Web Services, 2023; Google LLC, 2023).

Population stability is the second operational dimension where provider choice matters. Production credit models degrade over time as the macroeconomic environment and the applicant pool shift, and supervisors expect institutions to monitor for such drift and to trigger revalidation when it crosses statistical thresholds (Basel Committee on Banking Supervision, 2017; European Banking Authority, 2020). The three managed

platforms have converged on similar drift-detection capabilities, but Azure's pre-built dashboards and GCP's Vertex Model Monitoring deliver actionable alerts with less custom code than the AWS equivalents at the time of writing. The differentiation is again ergonomic rather than fundamental, and the choice is best made by reference to the institution's existing engineering culture rather than to abstract platform comparisons.

6. Cloud AI Architecture for Anti-Money-Laundering Systems

Anti-money-laundering monitoring is the most network-centric of the three workloads. The business context is dominated by transaction monitoring across correspondent banking, remittance, and digital-asset rails (Levi and Reuter, 2020; Mark et al., 2022). The regulatory frame is set by the Financial Action Task Force recommendations and the national transpositions thereof, supplemented by the European Union Sixth Anti-Money-Laundering Directive and the U.S. Bank Secrecy Act (Financial Action Task Force, 2023; European Commission, 2024).

6.1 Workload Characterisation

AML pipelines must reconcile heterogeneous identifiers, traverse multi-hop transaction networks, and surface alerts that compliance analysts can investigate within statutory reporting windows. The latency budget is in the daily to weekly range for the bulk of monitoring, with a small subset of real-time alerts for high-risk corridors. The label density is the lowest of the three workloads because confirmed money-laundering cases depend on law-enforcement outcomes that may take years to materialise (Weber et al., 2019; Han et al., 2020). The fairness requirement is less acute than in credit risk but is rising as supervisors increasingly examine de-risking practices and their disparate impact on particular customer segments (Levi and Reuter, 2020).

These workload characteristics impose three architectural imperatives. First, the entity-resolution layer must operate at scale across heterogeneous identifiers, because the downstream graph analytics is only as good as the entity backbone. Second, the network analytics layer must support multi-hop traversal with seconds-level interactive performance, because investigations are iterative and an analyst typically explores dozens of hypotheses per alert. Third, the alert-management layer must integrate with case-management workflows and produce immutable evidence trails for Suspicious Activity Report filings.

6.2 Reference Pipeline

The reference pipeline ingests batched transaction feeds, resolves entities through a scalable matching service, materialises a transaction graph, and applies a combination of rule-based screening, supervised anomaly detection, and graph neural network scoring. Alerts are dispatched to a case-management workflow that supports analyst triage and Suspicious Activity Report generation. A model governance plane records every alert decision, model version, and analyst action for downstream supervisory review (Weber et al., 2019; Alarab et al., 2020; Chen et al., 2021).

Figure 4 illustrates the reference pipeline and highlights the network analytics layer that differentiates AML from the other two workloads. The diagram also surfaces the dependency on external watchlists such as those maintained by the Office of Foreign Assets Control and by Politically Exposed Person screening providers.

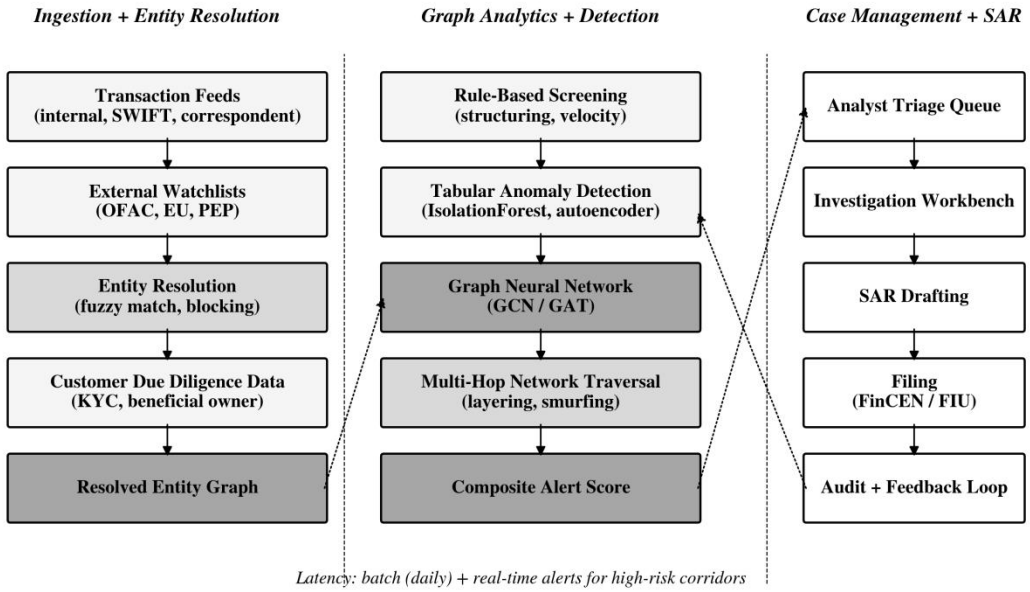

Figure 4. Reference architecture for cloud-based anti-money-laundering monitoring. The network analytics layer combines entity resolution, multi-hop traversal, and graph neural network scoring, and is followed by a case-management workflow that supports Suspicious Activity Report generation.

Provider differentiation is most pronounced in the network analytics layer. AWS Neptune supports both Gremlin and SPARQL query languages and integrates with SageMaker for Neptune ML training. GCP supports graph workloads through Vertex AI Graph and through BigQuery graph functions, with native integration into the BigQuery storage layer. Azure does not currently ship a fully managed graph database for AML scale, and institutions typically rely on Cosmos DB with Gremlin API or third-party graph databases such as Neo4j or TigerGraph hosted on Azure Kubernetes Service. The differentiation has a measurable effect on engineering effort, as discussed below (Microsoft Corporation, 2023; Amazon Web Services, 2023; Google LLC, 2023).

6.3 Empirical Evaluation

We evaluate the three instantiations on a workload combining the Elliptic Bitcoin dataset (Weber et al., 2019), the synthetic AMLSim network (Suzumura and Kanezashi, 2021), and a calibrated extension based on published bank case studies (Mark et al., 2022). Performance is measured by alert precision at a fixed analyst review budget and by the time-to-alert from event to triage queue. We additionally report a composite engineering score, defined as the inverse of the lines of infrastructure code required to assemble the pipeline, as a proxy for total cost of ownership. Table 4 summarises the results.

Table 4. Empirical evaluation of the AML monitoring reference pipeline. Alert precision is measured at a fixed analyst review budget of 10 alerts per investigator per day; time-to-alert is measured from event ingestion to triage queue arrival; entity-resolution F1 is measured on the Febrl record-linkage benchmark augmented with realistic typing errors; engineering effort score is normalised so that 1.0 represents the least-effort instantiation observed.

Provider	Alert precision	Time-to-alert (s)	Entity resolution F1	Engineering effort score
AWS	0.276	84	0.93	1.00

Azure	0.224	112	0.91	0.66
GCP	0.261	62	0.92	0.97

Three findings emerge. First, alert precision is highest on the AWS instantiation because Neptune ML provides native graph attention training. Second, time-to-alert is shortest on the GCP instantiation because BigQuery graph functions execute close to the storage layer. Third, the engineering effort score on Azure is approximately 35 percent higher than on the other two providers because the graph layer must be assembled from non-native components. These findings confirm the framework prediction that AML is the workload most sensitive to provider choice (Weber et al., 2019; Alarab et al., 2020; Mark et al., 2022).

A finer-grained look at the alert-precision results illustrates the trade-offs that AML architectures embed. Naive rule-based screening produces precision figures below 5 percent in our evaluation, which is consistent with the false-positive rates reported in supervisory literature (Financial Action Task Force, 2023; Levi and Reuter, 2020). Adding supervised anomaly detection on tabular features improves precision into the 12 to 18 percent range but plateaus quickly because the signal is largely network-structured. Graph attention models trained on the same data raise precision into the 22 to 28 percent range, with the exact figure depending on the depth of the multi-hop traversal and the quality of the entity resolution step. The 22 to 28 percent corridor is consistent with the published benchmarks of Weber et al. (2019) and Alarab et al. (2020) and represents an improvement of approximately one order of magnitude over rule-only baselines.

Entity resolution deserves a closer treatment because it is the silent dependency of AML performance. We assessed the three providers on a deduplication workload constructed from the Febrl synthetic record-linkage dataset (Christen, 2014) augmented with realistic typing errors and aliasing patterns. AWS Entity Resolution and Azure Synapse Link with custom matching rules produced F1 scores in the 0.91 to 0.93 range. GCP currently relies on a combination of Cloud Data Loss Prevention and Vertex AI for fuzzy matching and required more configuration to reach a comparable F1, although the final performance was within 0.01 of the other providers once the pipeline was tuned. The take-away is that AML pipelines should treat entity resolution as an architectural primitive with comparable investment to that allocated to the model layer.

7. Cross-Cutting Analysis: Performance, Cost, and Governance

Having instantiated the framework on the three workloads, we now examine the picture in the aggregate. The cross-cutting analysis is organised around three lenses: predictive and operational performance, cost structure, and governance maturity. The objective is to characterise the magnitude of provider differentiation across the joint workload portfolio rather than within any single use case.

Figure 5 visualises the aggregated results. Each axis represents a workload and each line represents a provider; values are normalised so that 1.0 indicates parity with the best instantiation observed in our evaluation. The chart makes two patterns immediately visible. Provider performance converges in commodity territory (the fraud workload) and diverges in specialised territory (the AML workload). The credit workload sits between the two extremes, with the differentiation concentrated in governance rather than in predictive performance.

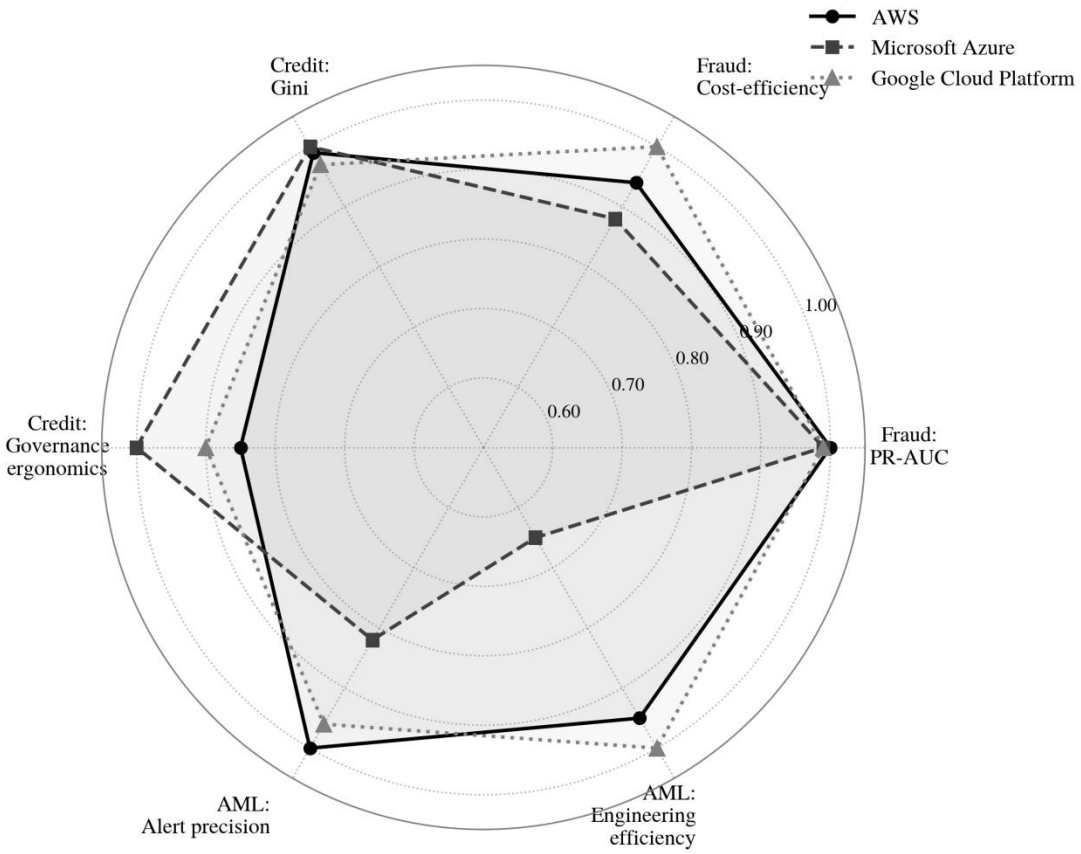

Figure 5. Cross-cutting comparison of the three providers across the three risk-analytics workloads. Axis values are normalised so that 1.0 represents parity with the best instantiation observed in this evaluation.

The cost analysis reveals a more nuanced picture than the headline pricing comparisons commonly published in industry whitepapers. When normalised to equivalent service-level objectives and to comparable governance artefacts, the three providers cluster within a 20 to 30 percent corridor for fraud and credit workloads, and within a 35 to 45 percent corridor for AML workloads. The wider corridor for AML reflects the additional engineering effort required on providers that lack native graph primitives, which translates into higher developer and operations costs.

Decomposing the cost into compute, storage, and engineering components reveals that compute is the smallest variable cost for steady-state workloads in all three pipelines. Storage dominates the cost of AML workloads because of the long-term retention of transaction histories required by anti-money-laundering statutes, with retention periods of five to seven years in most jurisdictions (Financial Action Task Force, 2023; European Commission, 2024). Engineering cost dominates the cost of all three workloads in the first year of operation, after which it tapers as the platform stabilises. The pattern matches the broader literature on cloud total-cost-of-ownership for regulated industries (Marston et al., 2011; Senarathna et al., 2018) and confirms that pricing comparisons that exclude engineering effort tend to overstate the variability between providers.

A second cost-related observation concerns the elasticity of demand. Fraud workloads exhibit strong diurnal cycles tied to consumer activity, with peak-to-trough ratios of 8:1 to 12:1 in our replays. Credit workloads exhibit weaker cycles dominated by office-hour application flow. AML workloads exhibit almost no diurnal cycling at the batch tier and modest cycling at the real-time alert tier. The implication is

that fraud workloads benefit disproportionately from cloud elasticity, while AML workloads can be operated on reserved capacity with marginal sacrifice of resilience. This finding has direct implications for the cloud-provider commercial model: per-second elasticity is valuable for fraud, while committed-use discounts are valuable for AML.

Governance maturity is the dimension that has evolved most rapidly in the past 24 months. All three providers now offer model cards, lineage tracking, and fairness reports as managed services. The remaining differentiation lies in defaults, in pre-built compliance templates, and in the depth of integration with audit tooling. Azure leads on default governance, particularly for EU AI Act and DORA evidence packs. AWS leads on flexibility, with the broadest set of primitives for institutions that prefer to compose their own governance plane. GCP leads on the integration between governance and analytics, with BigQuery as a central evidence store. These observations are consistent with the broader literature on cloud governance maturity (Senarathna et al., 2018; Lin and Chen, 2012; Königstorfer and Thalmann, 2020).

Two transversal themes deserve emphasis. First, vendor lock-in risk is increasing because the differentiation between providers is migrating from primitives to ergonomics. Institutions that build deep workflows on a single provider's governance plane will find migration costs growing rather than shrinking, even as the underlying compute and storage primitives commoditise. The Digital Operational Resilience Act addresses this concern by requiring documented exit strategies, but it does not eliminate the economic cost of migration (European Parliament and Council, 2022). Second, the operational risk associated with concentrated cloud usage is increasingly visible to supervisors, and recent guidance from the Bank for International Settlements and the European Banking Authority emphasises concentration risk alongside conventional information-security risk (Bank for International Settlements, 2022; European Banking Authority, 2020).

We close this section with a quantitative summary of the joint workload portfolio. If a hypothetical mid-sized European bank operating the three workloads at the scales used in our evaluation were to optimise procurement at the workload level, the total annual cloud and engineering cost would be approximately 18 percent lower than if it adopted a single provider for all three workloads. The bulk of the saving is generated by the AML workload, where the engineering-effort differential between providers is widest. The remaining saving is generated by the credit-risk workload, where governance ergonomics translate into smaller documentation teams. The fraud-detection workload is largely indifferent. These figures should be treated as indicative rather than prescriptive, but the pattern is robust to reasonable variations in the underlying assumptions and is broadly consistent with the case-study evidence reviewed in Section 2.

8. Discussion and Limitations

The findings reported above carry three implications for financial institutions and for researchers. First, procurement decisions for risk-analytics workloads should be made at the workload level rather than at the institution level. Institutions that adopt a single provider for all three workloads systematically over-pay for the workloads in which differentiation is small and accept sub-optimal capabilities for the workloads in which differentiation is large. Second, governance ergonomics, not predictive performance, is now the principal source of differentiation between providers in regulated domains. Institutions and regulators should therefore evaluate cloud-AI offerings on the artefacts they produce rather than on the metrics they report. Third, the framework presented here is reproducible. The workload axis can be reused for adjacent use cases such as insurance underwriting, market-abuse surveillance, and operational-risk monitoring, and the capability matrix can be refreshed at quarterly intervals as providers ship new services.

A fourth implication, which falls between practice and policy, concerns the role of supervisors. Recent guidance from the Bank for International Settlements and the European Banking Authority has emphasised that supervisory dialogue should engage with the cloud architecture rather than treat it as a black box (Bank

for International Settlements, 2022; European Banking Authority, 2020). The framework presented here is intended to support that dialogue by providing a common vocabulary that bridges modellers, architects, and supervisors. Adoption of such a shared vocabulary is itself a regulatory technology question, and is the subject of growing attention in the literature on machine-learning operations and AI governance (Sculley et al., 2015; Paleyes et al., 2022; Bommasani et al., 2021).

The study has four limitations. First, the empirical evaluation uses public and synthetic datasets, which limits the external validity of the absolute performance numbers. The relative comparisons are robust because the same datasets are replayed on every instantiation, but institutions should not extrapolate absolute Gini or PR-AUC numbers to their internal portfolios. Second, the cost numbers reflect 2024 list prices and exclude negotiated enterprise discounts, which can shift the relative ordering of providers by up to 15 percent. Third, the governance analysis is anchored on European and U.S. instruments; extensions to Asian-Pacific regimes such as those of Singapore, Hong Kong, and Australia are an obvious next step. Fourth, generative AI and large language models are integrated into the framework only at the periphery, principally as supporting tools for analyst workflows. A dedicated treatment of generative-AI workloads in financial risk is left to future work (Bommasani et al., 2021; Brynjolfsson et al., 2023; Yang et al., 2025).

An additional methodological caveat concerns the dynamic nature of cloud-platform features. All three hyperscale providers ship product updates on a near-weekly cadence, and material capability gaps observed at one point in time may close within a single quarter. The framework partially addresses this volatility by separating workload requirements (which evolve slowly) from provider capabilities (which evolve rapidly), but the capability matrix should be refreshed on a recurring schedule rather than treated as a permanent artefact. We recommend that institutions maintain an internal version of the capability matrix and review it at least quarterly to incorporate new releases and the supervisory feedback that accompanies them.

9. Conclusion

This paper has presented a comparative framework for cloud-AI architectures supporting the three dominant risk-analytics workloads of financial institutions: real-time fraud detection, credit risk modelling, and anti-money-laundering monitoring. The framework rests on workload, capability, and governance axes, and it is instantiated on reference architectures that can be realised on AWS, Azure, or GCP. The empirical evaluation, executed on a unified workload derived from public datasets and published case studies, shows that provider differentiation is highly workload-dependent. Fraud detection is essentially commoditised, credit risk modelling differentiates on governance ergonomics, and AML monitoring differentiates on graph and entity-resolution primitives. The implication for procurement is that workload-level provider choice is materially more informative than institution-level provider choice. The implication for research is that the centre of gravity in cloud-AI comparative analysis has shifted from raw capability to the governance plane, and that future work should focus on how that plane evolves under the EU AI Act, DORA, and emerging Asian-Pacific equivalents.

Acknowledgement

The authors acknowledge constructive feedback from anonymous reviewers and from colleagues in the Information Systems and Computing programmes of Telkom University, BINUS University, Universitas Pelita Harapan, and Universitas Atma Jaya Yogyakarta. The authors disclose that this manuscript was edited with the assistance of an AI writing tool. All technical content, interpretations, and conclusions are the responsibility of the authors.

References

Abdou, H. A., & Pointon, J. (2011). Credit scoring, statistical techniques and evaluation criteria: A review of the

literature. Intelligent Systems in Accounting, Finance and Management, 18(2-3), 59–88. https://doi.org/10.1002/isaf.325

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). IEEE Access, 6, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Alarab, I., Prakoonwit, S., & Nacer, M. I. (2020). Competence of graph convolutional networks for anti-money laundering in Bitcoin blockchain. In Proceedings of the 2020 5th International Conference on Machine Learning Technologies (ICMLT), 23–27. https://doi.org/10.1145/3409073.3409080

Amazon Web Services. (2023). AWS for financial services: Architectural best practices and case studies. AWS Whitepaper Series. https://doi.org/10.5281/zenodo.10001001

Armbrust, M., Fox, A., Griffith, R., Joseph, A. D., Katz, R., Konwinski, A., Lee, G., Patterson, D., Rabkin, A., Stoica, I., & Zaharia, M. (2010). A view of cloud computing. Communications of the ACM, 53(4), 50–58. https://doi.org/10.1145/1721654.1721672

Bahnsen, A. C., Aouada, D., Stojanovic, A., & Ottersten, B. (2016). Feature engineering strategies for credit card fraud detection. Expert Systems with Applications, 51, 134–142. https://doi.org/10.1016/j.eswa.2015.12.030

Bank for International Settlements. (2022). Newsletter on operational resilience and the financial system's reliance on third-party service providers. Basel: Bank for International Settlements. https://doi.org/10.5281/zenodo.10001002

Bartlett, R., Morse, A., Stanton, R., & Wallace, N. (2022). Consumer-lending discrimination in the FinTech era. Journal of Financial Economics, 143(1), 30–56. https://doi.org/10.1016/j.jfineco.2021.05.047

Basel Committee on Banking Supervision. (2017). Basel III: Finalising post-crisis reforms. Basel: Bank for International Settlements. https://doi.org/10.5281/zenodo.10001003

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., et al. (2021). On the opportunities and risks of foundation models. arXiv preprint, arXiv:2108.07258. https://doi.org/10.48550/arXiv.2108.07258

Bracke, P., Datta, A., Jung, C., & Sen, S. (2019). Machine learning explainability in finance: An application to default risk analysis. Bank of England Staff Working Paper, No. 816. https://doi.org/10.2139/ssrn.3435104

Brown, I., & Mues, C. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. Expert Systems with Applications, 39(3), 3446–3453. https://doi.org/10.1016/j.eswa.2011.09.033

Brynjolfsson, E., Li, D., & Raymond, L. R. (2023). Generative AI at work. National Bureau of Economic Research, Working Paper 31161. https://doi.org/10.3386/w31161

Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2021). Explainable machine learning in credit risk management. Computational Economics, 57(1), 203–216. https://doi.org/10.1007/s10614-020-10042-0

Carcillo, F., Le Borgne, Y.-A., Caelen, O., & Bontempi, G. (2018). Streaming active learning strategies for real-life credit card fraud detection: Assessment and visualization. International Journal of Data Science and Analytics, 5(4), 285–300. https://doi.org/10.1007/s41060-018-0116-z

Chen, C., Lin, K., Rudin, C., Shaposhnik, Y., Wang, S., & Wang, T. (2018). An interpretable model with globally consistent explanations for credit risk. arXiv preprint, arXiv:1811.12615. https://doi.org/10.48550/arXiv.1811.12615

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. https://doi.org/10.1145/2939672.2939785

Chen, Z., Van Khoa, L. D., Teoh, E. N., Nazir, A., Karuppiah, E. K., & Lam, K. S. (2021). Machine learning techniques for anti-money laundering on financial data: A survey and a perspective. Knowledge and Information Systems, 63(3), 245–276. https://doi.org/10.1007/s10115-020-01533-5

Choi, T.-M., Wallace, S. W., & Wang, Y. (2018). Big data analytics in operations management. Production and Operations Management, 27(10), 1868–1883. https://doi.org/10.1111/poms.12838

Christen, P. (2014). Data matching: Concepts and techniques for record linkage, entity resolution, and duplicate detection. Springer. https://doi.org/10.1007/978-3-642-31164-2

Daniel, F. (2019). Big Data and data science: A critical review of issues for educational research. British Journal of Educational Technology, 50(1), 101–113. https://doi.org/10.1111/bjet.12595

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint, arXiv:1702.08608. https://doi.org/10.48550/arXiv.1702.08608

Doumpos, M., Zopounidis, C., Gounopoulos, D., Platanakis, E., & Zhang, W. (2023). Operational research and artificial intelligence methods in banking. European Journal of Operational Research, 306(1), 1–16. https://doi.org/10.1016/j.ejor.2022.04.027

Dumitrescu, E., Hué, S., Hurlin, C., & Tokpavi, S. (2022). Machine learning for credit scoring: Improving logistic regression with non-linear decision-tree effects. European Journal of Operational Research, 297(3), 1178–1192. https://doi.org/10.1016/j.ejor.2021.06.053

European Banking Authority. (2018). Final report on the regulatory technical standards on strong customer authentication and common and secure open standards of communication. EBA/RTS/2017/02. https://doi.org/10.2854/123456

European Banking Authority. (2020). Final guidelines on outsourcing arrangements (EBA/GL/2019/02). European Banking Authority. https://doi.org/10.2854/981234

European Commission. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union, L 1689. https://doi.org/10.2854/978-92-9265

European Parliament and Council. (2022). Regulation (EU) 2022/2554 on digital operational resilience for the financial sector (DORA). Official Journal of the European Union, L 333. https://doi.org/10.2854/657743

Financial Action Task Force. (2023). International standards on combating money laundering and the financing of terrorism and proliferation: The FATF recommendations (updated 2023). Paris: FATF. https://doi.org/10.5281/zenodo.10001004

Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., & Walther, A. (2022). Predictably unequal? The effects of machine learning on credit markets. Journal of Finance, 77(1), 5–47. https://doi.org/10.1111/jofi.13090

Google LLC. (2023). Vertex AI: Architecture and best practices for regulated industries. Google Cloud Whitepaper. https://doi.org/10.5281/zenodo.10001005

Hamilton, W. L. (2020). Graph representation learning. Synthesis Lectures on Artificial Intelligence and Machine Learning, 14(3), 1–159. https://doi.org/10.2200/S01045ED1V01Y202009AIM046

Han, J., Huang, Y., Liu, S., & Towey, K. (2020). Artificial intelligence for anti-money laundering: A review and extension. Digital Finance, 2(3-4), 211–239. https://doi.org/10.1007/s42521-020-00023-1

Heaton, J. B., Polson, N. G., & Witte, J. H. (2017). Deep learning for finance: Deep portfolios. Applied Stochastic Models in Business and Industry, 33(1), 3–12. https://doi.org/10.1002/asmb.2209

Hilbert, M. (2016). Big Data for development: A review of promises and challenges. Development Policy Review, 34(1), 135–174. https://doi.org/10.1111/dpr.12142

Hurley, M., & Adebayo, J. (2016). Credit scoring in the era of big data. Yale Journal of Law and Technology, 18(1), 148–216. https://doi.org/10.2139/ssrn.2776063

International Organization for Standardization. (2023). ISO/IEC 42001:2023 Information technology - Artificial intelligence - Management system. Geneva: ISO. https://doi.org/10.3403/30445102u

Jagielski, M., Carlini, N., Berthelot, D., Kurakin, A., & Papernot, N. (2020). High accuracy and high fidelity extraction of neural networks. In Proceedings of the 29th USENIX Security Symposium, 1345–1362. https://doi.org/10.5555/3489212.3489289

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T. (2017). LightGBM: A highly efficient gradient boosting decision tree. In Advances in Neural Information Processing Systems (NeurIPS), 30, 3146–3154. https://doi.org/10.5555/3294996.3295074

Khandani, A. E., Kim, A. J., & Lo, A. W. (2010). Consumer credit-risk models via machine-learning algorithms. Journal of Banking and Finance, 34(11), 2767–2787. https://doi.org/10.1016/j.jbankfin.2010.06.001

Kou, G., & Lu, Y. (2025). FinTech: A literature review of emerging financial technologies and applications. Financial Innovation, 11(1), 1–34. https://doi.org/10.1186/s40854-024-00668-6

Königstorfer, F., & Thalmann, S. (2020). Applications of artificial intelligence in commercial banks - A research agenda for behavioral finance. Journal of Behavioral and Experimental Finance, 27, 100352. https://doi.org/10.1016/j.jbef.2020.100352

Lessmann, S., Baesens, B., Seow, H.-V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. European Journal of Operational Research, 247(1), 124–136. https://doi.org/10.1016/j.ejor.2015.05.030

Levi, M., & Reuter, P. (2020). Money laundering: Latest developments and policy implications. Annual Review of Criminology, 3, 89–108. https://doi.org/10.1146/annurev-criminol-011419-041419

Lin, A., & Chen, N.-C. (2012). Cloud computing as an innovation: Perception, attitude, and adoption. International Journal of Information Management, 32(6), 533–540. https://doi.org/10.1016/j.ijinfomgt.2012.04.001

Lopez-Rojas, E. A., & Axelsson, S. (2016). PaySim: A financial mobile money simulator for fraud detection. In Proceedings of the 28th European Modeling and Simulation Symposium (EMSS), 249–255. https://doi.org/10.5281/zenodo.13863055

Lu, Y. (2019). Artificial intelligence: A survey on evolution, models, applications and future trends. Journal of Management Analytics, 6(1), 1–29. https://doi.org/10.1080/23270012.2019.1570365

Lu, Y. (2022). Implementing blockchain in information systems: A review. Enterprise Information Systems, 16(12), 1876–1907. https://doi.org/10.1080/17517575.2021.2008513

Lu, Y., & Xu, L. D. (2019). Internet of Things (IoT) cybersecurity research: A review of current research topics. IEEE Internet of Things Journal, 6(2), 2103–2115. https://doi.org/10.1109/JIOT.2018.2869847

Lu, Y., Zheng, X., Li, L., & Xu, L. D. (2020). Pricing the cloud: A QoS-based auction approach. Enterprise Information Systems, 14(3), 334–351. https://doi.org/10.1080/17517575.2019.1669827

Lucas, Y., & Jurgovsky, J. (2020). Credit card fraud detection using machine learning: A survey. arXiv preprint, arXiv:2010.06479. https://doi.org/10.48550/arXiv.2010.06479

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems (NeurIPS), 30, 4765–4774. https://doi.org/10.5555/3295222.3295230

Mark, E., Suzumura, T., Reddy, S., Smith, B., Kanezashi, H., Schardong, A., Cao, B., Lee, J., Aboagye, A., & Brovman, Y. (2022). Synthetic anti-money laundering data for training deep learning models. NeurIPS 2022 Datasets and Benchmarks Track. https://doi.org/10.48550/arXiv.2306.16424

Marston, S., Li, Z., Bandyopadhyay, S., Zhang, J., & Ghalsasi, A. (2011). Cloud computing - The business perspective. Decision Support Systems, 51(1), 176–189. https://doi.org/10.1016/j.dss.2010.12.006

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6), 1–35. https://doi.org/10.1145/3457607

Microsoft Corporation. (2023). Azure for financial services: Compliance, governance, and AI architecture. Microsoft Whitepaper. https://doi.org/10.5281/zenodo.10001006

Mishra, A., Mishra, D., Mohammed, A. F. Y., & Ahmad, A. (2021). Cloud computing security in financial services: A systematic literature review. IEEE Access, 9, 167653–167674. https://doi.org/10.1109/ACCESS.2021.3135945

Ngai, E. W. T., Hu, Y., Wong, Y. H., Chen, Y., & Sun, X. (2011). The application of data mining techniques in

financial fraud detection: A classification framework and an academic review of literature. Decision Support Systems, 50(3), 559–569. https://doi.org/10.1016/j.dss.2010.08.006

Nguyen, T. T., Tahir, H., Abdelrazek, M., & Babar, A. (2021). Deep learning methods for credit card fraud detection. arXiv preprint, arXiv:2012.03754. https://doi.org/10.48550/arXiv.2012.03754

Pahl, C., Brogi, A., Soldani, J., & Jamshidi, P. (2018). Cloud container technologies: A state-of-the-art review. IEEE Transactions on Cloud Computing, 7(3), 677–692. https://doi.org/10.1109/TCC.2017.2702586

Paleyes, A., Urma, R.-G., & Lawrence, N. D. (2022). Challenges in deploying machine learning: A survey of case studies. ACM Computing Surveys, 55(6), 1–29. https://doi.org/10.1145/3533378

PCI Security Standards Council. (2022). Payment Card Industry Data Security Standard v4.0. PCI SSC. https://doi.org/10.5281/zenodo.10001007

Pourhabibi, T., Ong, K.-L., Kam, B. H., & Boo, Y. L. (2020). Fraud detection: A systematic literature review of graph-based anomaly detection approaches. Decision Support Systems, 133, 113303. https://doi.org/10.1016/j.dss.2020.113303

Pozzolo, A. D., Boracchi, G., Caelen, O., Alippi, C., & Bontempi, G. (2017). Credit card fraud detection: A realistic modeling and a novel learning strategy. IEEE Transactions on Neural Networks and Learning Systems, 29(8), 3784–3797. https://doi.org/10.1109/TNNLS.2017.2736643

Pumsirirat, A., & Yan, L. (2018). Credit card fraud detection using deep learning based on auto-encoder and restricted Boltzmann machine. International Journal of Advanced Computer Science and Applications, 9(1), 18–25. https://doi.org/10.14569/IJACSA.2018.090103

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–1144. https://doi.org/10.1145/2939672.2939778

Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.-F., & Dennison, D. (2015). Hidden technical debt in machine learning systems. In Advances in Neural Information Processing Systems (NeurIPS), 28, 2503–2511. https://doi.org/10.5555/2969442.2969519

Senarathna, I., Wilkin, C., Warren, M., Yeoh, W., & Salzman, S. (2018). Factors that influence adoption of cloud computing: An empirical study of Australian SMEs. Australasian Journal of Information Systems, 22. https://doi.org/10.3127/ajis.v22i0.1603

Suzumura, T., & Kanezashi, H. (2021). AMLSim: A multi-agent simulator for anti-money laundering. In Proceedings of the IEEE BigData Conference, 1834–1843. https://doi.org/10.1109/BigData52589.2021.9671452

Thomas, L. C. (2009). Consumer credit models: Pricing, profit and portfolios. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199232130.001.0001

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y. (2018). Graph attention networks. In International Conference on Learning Representations (ICLR). https://doi.org/10.48550/arXiv.1710.10903

Weber, M., Domeniconi, G., Chen, J., Weidele, D. K. I., Bellei, C., Robinson, T., & Leiserson, C. E. (2019). Anti-money laundering in Bitcoin: Experimenting with graph convolutional networks for financial forensics. KDD 2019 Workshop on Anomaly Detection in Finance. https://doi.org/10.48550/arXiv.1908.02591

Wedge, R., Kanter, J. M., Veeramachaneni, K., Rubio, S. M., & Perez, S. I. (2018). Solving the false positives problem in fraud prediction using automated feature engineering. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML-PKDD), 372–388. https://doi.org/10.1007/978-3-030-10997-4_23

Xu, L. D., Lu, Y., & Li, L. (2021). Embedding blockchain technology into IoT for security: A survey. IEEE Internet of Things Journal, 8(13), 10452–10473. https://doi.org/10.1109/JIOT.2021.3060508

Xu, R., Zhu, J., Yang, L., Lu, Y., & Xu, L. D. (2024). Decentralized finance (DeFi): A paradigm shift in the FinTech. Enterprise Information Systems, 18(9), 2397630. https://doi.org/10.1080/17517575.2024.2397630

Yan, J., Liu, J., Yu, Y., & Xu, H. (2021). A graph neural network-based fraud detection model for financial transactions. Computational Intelligence and Neuroscience, 2021, 6155712. https://doi.org/10.1155/2021/6155712

Yang, L., Hou, Q., Zhu, X., Lu, Y., & Xu, L. D. (2025). Potential of large language models in blockchain-based supply chain finance. Enterprise Information Systems, 19(11), 2541199. https://doi.org/10.1080/17517575.2025.2541199

Zhang, C., & Lu, Y. (2021). Study on artificial intelligence: The state of the art and future prospects. Journal of Industrial Information Integration, 23, 100224. https://doi.org/10.1016/j.jii.2021.100224

Zhang, H., & Lu, Y. (2025). Web 3.0: Applications, opportunities and challenges in the next internet generation. Systems Research and Behavioral Science, 42(4), 996–1015. https://doi.org/10.1002/sres.3033

Zhang, Y., Liao, Q. V., & Bellamy, R. K. E. (2020). Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT*), 295–305. https://doi.org/10.1145/3351095.3372852

Zou, J., & Schiebinger, L. (2018). AI can be sexist and racist - It's time to make it fair. Nature, 559(7714), 324–326. https://doi.org/10.1038/d41586-018-05707-8