International Journal of Science and Research (IJSR)
ISSN: 2319-7064
Impact Factor 2025: 7.089

Reducing False Positives in AML Transaction Monitoring: A Risk-Based Framework Applied to a Composite UK Banking Case

Malki Senevirathne

Independent Researcher, United Kingdom

Abstract: ***Anti-money laundering (AML) transaction monitoring systems in UK financial institutions typically generate false positive rates of 85–95 per cent. Compliance teams spend up to 90 per cent of investigation time on alerts that never result in a Suspicious Activity Report. This paper sets out a risk-based framework for reducing false positives without weakening detection: it combines customer risk segmentation, above-the-line/below-the-line threshold tuning, periodic rule rationalisation, and a hybrid model that pairs automated detection with targeted manual review. Disclosed institutional data was not available, so the framework is applied step by step to an illustrative, composite mid-sized UK retail and commercial bank built entirely from published academic and industry benchmarks, not any real organisation's figures. Applying published threshold-tuning effect sizes to the composite case's modelled baseline of 800 alerts a day suggests that tuning and rationalisation alone could cut alert volume by roughly one-sixth and free up approximately 81 analyst-hours a day for higher-value review, while keeping the great majority of genuine escalations. The paper also sets out the governance evidence and second-line resourcing this approach needs, and identifies empirical validation against real institutional data as the necessary next step to establish whether these modelled gains would hold in practice.***

Keywords: Anti-Money Laundering, Transaction Monitoring, False Positives, Risk-Based Approach, Financial Crime Compliance

1. Introduction

Anti-money laundering (AML) transaction monitoring systems are meant to separate genuine financial crime risk from routine activity. In practice, most of what they flag is neither. Industry benchmarks put false-positive rates in transaction monitoring at 85–95 per cent, with compliance teams spending up to 90 per cent of investigation time on alerts that are ultimately closed without any further action [1]. Academic testing against real banking data reaches a similar conclusion by a different route: Ketenci et al. report a baseline false-positive rate above 90 per cent before any tuning intervention is applied [2]. For a UK retail or commercial bank generating tens of thousands of alerts a month, that gap between what is flagged and what actually matters carries a real cost. Analyst time gets tied up on low-value review work, alerts that do carry genuine risk take longer to turn around, and institutions face mounting pressure to show that monitoring resources are allocated according to risk, not volume, as the risk-based approach formalised in FATF Recommendation 1 and transposed into UK law requires [3].

EY's 2024 UK AML transaction monitoring survey found that most UK institutions are now running some form of threshold tuning or segmentation exercise in response. Its more pointed finding, though, is that alert-quality outcomes have not improved by much even where this work has taken place [4]. That suggests the shortfall is not simply that thresholds are set wrong. Pontes et al.'s interviews with UK AML practitioners identify a more structural issue: a disconnect between what the risk-based approach requires in principle and how it is actually supervised and applied, which undermines its effectiveness regardless of how carefully any one institution tunes its own rules [5]. Russell's decade-long review of FCA enforcement notices reaches a similar conclusion, finding that the industry's primary interpretive guidance overemphasises customer due diligence relative to the wider risk-governance process it is meant to support [6].

This is not a gap in the underlying research techniques. Threshold and segmentation methods capable of closing this gap already exist in the literature. Gupta et al. apply multi-dimensional optimisation to the "above-the-line" threshold-setting practice used across the industry, coordinating thresholds across overlapping scenarios to reduce redundant alerts without losing suspicious-activity coverage [7]; Badics et al. report an 18 per cent reduction in false positives while retaining 98 per cent of true positives using a comparable approach on real bank data [8]. What is comparatively thin is published material that walks through how a risk-based methodology like this is actually built and applied inside an institution, step by step. Most published accounts describe the technique in the abstract or evaluate it after the fact on data no other researcher can see. What is publicly available tends to sit at one of two extremes: peer-reviewed technical papers evaluated on proprietary datasets that cannot be reproduced, or regulatory and vendor commentary that describes the destination without showing the working.

This paper aims to close part of that gap. It sets out a risk-based framework for reducing false positives in AML transaction monitoring, built entirely from published academic and industry sources rather than internal institutional data, and applies that framework step by step to an illustrative, composite retail and commercial banking case. No actual institution's transaction data, thresholds, or alert outcomes are used or disclosed anywhere in this paper. The composite case is built to be representative of the kind of profile a UK-regulated institution would monitor, and every quantitative assumption within it traces back to a cited public source, not a confidential figure. The intent is a worked example that a compliance or risk practitioner could adapt to their own institution, not a validated predictive model.

The remainder of the paper is organised as follows. Section 2 reviews the academic literature on false-positive reduction in transaction monitoring, covering machine learning and risk-scoring approaches, threshold and segmentation methods, explainability and governance constraints, and the practitioner and UK regulatory context. Section 3 defines the specific problem this paper addresses. Section 4 covers the methodology: how the composite case was built, the risk-based framework applied to it, and the limitations of that approach. Section 5 presents the results of applying the framework to the illustrative case and discusses what they suggest. Section 6 concludes, and Section 7 sets out scope for future work using real institutional data.

2. Literature Survey

2.1 The risk-based approach and the persistence of false positives

The risk-based approach that underpins modern anti-money laundering (AML) supervision was formalised by the Financial Action Task Force's Recommendation 1, which requires institutions to direct AML resources according to assessed risk rather than uniform compliance [3]. Savona and Riccardi trace how this shift, adopted into UK and EU law through successive Money Laundering Directives, reframed transaction monitoring from a rule-following exercise into a resource-allocation problem [3]. In practice, the architecture built to operationalise this shift has changed little. Financial institutions still rely predominantly on static, threshold-based rules that flag any transaction crossing a pre-set value or pattern, an approach Chen et al. describe in one of the earliest comprehensive surveys of machine learning applied to AML detection [9]. The empirical record on what this architecture produces is unusually consistent. Ketenci et al., testing detection models against real banking data, confirm a baseline false-positive rate above 90 per cent before any tuning intervention is applied [2]. Later work by Jensen and Iosifidis proposes a unifying vocabulary, distinguishing "client risk profiling" from "suspicious behaviour flagging", precisely because the literature had grown fragmented around different attempts to address the same underlying problem [10].

2.2 Machine learning and risk-scoring approaches

Two adjacent bodies of work respond to the false-positive problem in different ways. The first, and by far the larger, treats it as a modelling problem: layer a machine learning classifier on top of the existing rule-based system and let it triage what the rules produce. Jullum et al., working with data from Norway's largest bank, show that a supervised model trained to predict which flagged transactions warrant investigation outperforms the bank's existing alert system on a purpose-built fairness measure, one of the few such models validated at realistic scale [11]. Eddin et al. extend this to graph-based features, using a triage model that incorporates the relationships between accounts rather than treating each transaction in isolation. They report an 80 per cent reduction in false positives while retaining over 90 per cent of true positives [12]. Halford et al. describe a comparable "bolt-on" scoring model implemented at a UK bank, which moved the institution from reviewing 100 per cent of alerts manually to hibernating 19 per cent and auto-escalating 18 per cent, cutting the time to file a suspicious activity report by 61 per cent [13]. Bakry et al. report a similar XGBoost-based suppression framework, but they are careful to quantify the residual risk: in their test data, 11 per cent of genuinely suspicious customers were incorrectly closed out, a reminder that false-positive suppression and missed true positives sit on the same trade-off curve [14]. A smaller strand of this literature applies the same logic to adjacent screening functions. Alkhalili et al., for instance, target watch-list and sanctions filtering specifically, an area they note had received little prior attention despite extensive research on know-your-customer applications [15].

2.3 Risk-based segmentation and threshold tuning

The second body of work treats the false-positive problem as a tuning problem rather than a modelling problem, and this is the strand that speaks most directly to a risk-based methodology that does not assume full machine learning deployment. Gupta et al. apply linear and non-linear optimisation to the "above-the-line" threshold-setting practice already used across the industry, coordinating thresholds across overlapping scenarios and customer segments to reduce redundant alerts without lowering suspicious-activity coverage [7]. Badics et al. go further, reformulating threshold optimisation as a microeconomic decision problem and reporting an 18 per cent reduction in false positives while retaining 98 per cent of true positives on real bank data. It is the clearest academic articulation found of the above-the-line method as practitioners actually use it [8]. Reite et al. approach the same problem from the client side instead of the threshold side, showing that enriching client risk classification with readily available accounting and credit-score data materially improves the ability to distinguish genuine risk from noise [16]. Alexandre and Balsa, meanwhile, demonstrate a multi-agent system combining machine learning with explicit risk-level rules, validated by AML analysts against six months of live transactions at a real institution [17].

2.4 Explainability and model governance

A third strand concerns whether any of this is deployable in practice, and here the literature turns from detection performance to governance. Weber, Carl and Hinz's systematic review of explainable AI across finance is the most cited paper returned by this search, and its central finding is unambiguous: explainability research is comparatively mature in risk management, portfolio optimisation and equity markets, and comparatively thin in anti-money laundering [18]. Kute et al. reach a similar conclusion from the deep-learning side. They find that just over half of the machine learning methods used in AML research at the time were non-interpretable, despite operating in a domain where regulators expect a documented rationale for every decision [19]. Fritz-Morgenthal, Hein and Papenbrock describe what a governance response to this gap might look like in practice, proposing a risk-based testing and oversight framework for AI/ML models in financial risk functions generally, and anticipating that AML would be one of the areas where model complexity grows fastest [20]. Gerlings' longitudinal study of a European bank implementing explainable AI in transaction monitoring adds an empirical account of what governance

looks like from the inside. She finds that stakeholders close to a model, such as data scientists and investigators, build trust through day-to-day familiarity with it, while stakeholders further from it, such as compliance managers and auditors, rely on explainability tools for oversight and documentation rather than operational trust. That complicates the assumption, common in the technical literature, that explainability straightforwardly produces trust [21]. Turksen et al. document a related tension from the supervisory side, finding that prudential banking supervisors' willingness to trust AI-driven AML tools is shaped as much by cost and legal uncertainty under frameworks like the EU AI Act as by the technical quality of the explanation offered [22].

2.5 Practitioner experience and the UK regulatory context

The practitioner and regulatory literature, much of it UK-specific, provides the clearest evidence that these are not solely technical problems. Oztas et al.'s interviews with eight AML specialists find that practitioners themselves identify explainability, flexibility and the ability to recognise new risk typologies as the central requirements for improving transaction monitoring, a direct, industry-sourced echo of the priorities identified independently in the technical and governance literature above [23, 24]. Gilson offers a useful corrective to any assumption that machine learning is the obvious next step for every institution, arguing that many banks find full ML deployment too costly or resource-intensive and that simpler tuning approaches remain underused and undervalued [25]. Pontes, Rocha and Mac Carthaigh's interviews and focus groups with UK AML practitioners identify a more structural problem: a disconnect between what the risk-based approach requires in principle and how it is supervised in practice. They argue this undermines its effectiveness regardless of how well any individual institution tunes its own systems [5]. Russell's analysis of a decade of Financial Conduct Authority enforcement notices gives this claim empirical weight, finding that JMLSG guidance, the industry's primary interpretive reference for the risk-based approach, overemphasises customer due diligence relative to the broader risk governance process, and fails to distinguish clearly between risk identification, analysis and evaluation [6]. Zavoli and King's empirical study of AML implementation in the UK property market, while outside banking, reaches a comparable conclusion about the gap between regulatory intent and operational reality [26]. Pol offers a more polemical assessment of AML policy effectiveness globally, estimating that the entire AML intervention affects under 0.1 per cent of criminal finances while imposing compliance costs many times greater than the funds it recovers. It is a useful counterweight for any paper focused narrowly on transaction-monitoring efficiency, since it is a reminder that reducing false positives is only a meaningful achievement when understood within this wider effectiveness debate, not as an end in itself [27]. Ogbeide et al. add a behavioural dimension to this picture, finding experimentally that AML risk-assessment experts are, if anything, slightly more overconfident than novices, and that both groups show a systematic preference for false-positive over false-negative errors. That is evidence that some of the bias toward over-alerting is a property of human judgement under uncertainty, not only of poorly tuned rules [28].

2.6 Related reviews, trade-based money laundering, and the research gap

Finally, two recent systematic reviews and a related literature on trade-based money laundering (TBML) help to place this body of work in context. B.K. and Ramasubramanian's review of AML techniques published between 2014 and 2024, and Castelao-López et al.'s PRISMA-guided review covering 2017 to 2024, both document the same broad trajectory: a shift from statistical and rule-based methods toward machine learning and network analysis, with interpretability and cross-institutional data sharing identified as the principal unresolved challenges [29, 30]. Tiwari, Gepp and Kumar's systematic review of TBML research, the first of its kind, finds a comparable pattern in an adjacent domain: a body of work still organised around risk assessment and detection typologies rather than around evaluated interventions, with too few studies to draw firm conclusions about what actually reduces risk in practice [31]. Naheem, writing earlier on AML and TBML risk assessment strategies specifically, argues that most institutional risk assessment remains reactive, responding to enforcement action and emerging case law instead of anticipating risk. That critique motivates the more proactive, cyclical tuning approach this paper applies [32].
Taken together, this literature leaves a specific and well-defined gap. The modelling literature shows, repeatedly and at real-bank scale, that false positives can be substantially reduced. The governance literature shows, with equal consistency, that explainability and institutional trust lag behind detection performance, and that AML specifically remains underexplored within the wider explainable-AI-in-finance literature [18]. The practitioner and regulatory literature shows that UK supervisors and firms alike are still working out what “risk-based” means in operational terms, years after the requirement was formalised. What is comparatively absent is a worked account of how a risk-based tuning methodology, combining segmentation, above-the-line threshold optimisation, rule rationalisation and hybrid automated-plus-manual review, is actually applied end to end, in a way that a compliance practitioner without a data-science team could follow and adapt. That is the gap this paper addresses.

3. Problem Definition

For the purposes of this paper, a false positive is defined as an alert generated by a transaction monitoring system that, following investigation by a compliance analyst, is closed without escalation to a Suspicious Activity Report (SAR) or equivalent internal disclosure. This is a deliberately narrow, outcome-based definition rather than a technical one. It says nothing about why the alert fired, only what happened to it afterwards. Halford et al.'s account of a UK bank moving from manually reviewing every alert to hibernating some and auto-escalating others illustrates the practical stakes of getting this classification right: every alert an institution can correctly identify as a false positive, without missing genuine risk, is time returned to the alerts that matter [13].

The core difficulty is that false positives cannot be reduced in isolation from true positives. Loosening detection thresholds to cut alert volume risks missing the transactions a monitoring system exists to catch. Bakry et al., testing a suppression

framework of this kind, find that even a well-performing model incorrectly closed out 11 per cent of genuinely suspicious customers in their test data, a reminder that false-positive reduction and missed true positives sit on the same trade-off curve, not on separate ones [14]. Over-alerting carries its own risk on the other side of that curve. Ogbeide et al.'s experimental work on AML risk assessment finds that analysts, whether expert or novice, show a systematic preference for false-positive over false-negative errors under uncertainty, and that experience does not reliably correct for this [28]. In an environment already producing far more alerts than a team can meaningfully investigate, that bias compounds rather than corrects itself: the more alerts an analyst reviews, the less scrutiny any individual alert can receive, including the genuine ones.

This paper frames that problem as it would present at a composite, illustrative mid-sized UK retail and commercial bank, built entirely from published figures and never intended to represent any real institution, including the author's employer. The composite bank is a UK-incorporated deposit-taker supervised under the Money Laundering, Terrorist Financing and Transfer of Funds (Information on the Payer) Regulations 2017 and the FCA's wider regulatory regime, running a conventional rule-based transaction monitoring system of the kind described across the literature reviewed in Section 2. Section 4 gives the specific assumptions used to construct this case in full. This section states only the problem the case is built to examine.

Framed this way, the paper addresses two related research questions. RQ1: how does applying a risk-based tuning methodology, of the kind set out in Section 4, change alert volume and the false-positive rate in a representative UK transaction monitoring case, relative to an untuned baseline? RQ2: what governance considerations, including model risk, explainability, audit trail and regulatory defensibility, does applying that methodology raise, and how might a compliance function address them alongside the operational gains?

4. Methodology / Approach

4.1 Research design and rationale

This paper uses a conceptual/applied case study design, built around a composite, illustrative institution rather than disclosed data from any real organisation. That is a constraint, not a preference. Transaction monitoring configurations, alert volumes and SAR outcomes are confidential by regulation and by employer policy alike, and no dataset of that kind could be obtained, let alone published, for a paper of this type. A composite case is the standard way applied compliance research handles that constraint without abandoning the underlying question.

The design choice affects what the paper can and cannot claim. It cannot test whether any specific institution's monitoring improved, since that would require disclosed data no paper of this kind can obtain. What it can test is whether a documented, replicable risk-based methodology, built from the techniques set out in Section 2, behaves as the literature predicts when applied step by step under realistic, published-benchmark conditions. That makes this closer to a worked methodological demonstration than to an empirical case study in the traditional sense, and the paper is explicit about that boundary throughout, including in the limitations set out in Section 4.5.

4.2 Construction of the composite case

The composite case represents a mid-sized UK retail and commercial bank: a deposit-taking institution supervised under the Money Laundering Regulations 2017 and the FCA's wider regime, running a conventional rule-based transaction monitoring system of the kind described throughout Section 2. Its starting assumptions come from published benchmarks rather than any institution's real figures, and each is cited to its source below.

- Baseline false-positive rate: industry benchmarks put transaction monitoring false-positive rates at 85–95 per cent [1], a range corroborated academically by Ketenci et al.'s finding of a baseline false-positive rate above 90 per cent before any tuning intervention [2]. This paper adopts the upper end of that range, 95 per cent, as the composite case's baseline.
- SAR conversion rate: the same industry source puts the share of alerts that result in a filed Suspicious Activity Report at 1–5 per cent [1]. This paper adopts the upper end of that range, 5 per cent, for internal consistency with the 95 per cent false-positive baseline above. Under the definition set out in Section 3, an alert is a false positive precisely when it is closed without SAR escalation, so the two figures are two ways of describing the same split, and are set to sum to 100 per cent.
- Average investigation time per alert: reported industry figures range from 20 to 60 minutes per alert [1]. This paper adopts the midpoint, 35 minutes, as the baseline.
- Daily alert volume: reported figures for large institutions run to 10,000 or more alerts per day [1]. The composite case represents a mid-sized rather than a large institution, so this paper does not reuse that figure directly. Instead it adopts 800 alerts per day as an illustrative, explicitly unsourced modelling assumption, chosen to be representative of a smaller book while keeping the arithmetic in the rest of this section legible. This is an assumption made for modelling purposes, not a benchmark drawn from any published source, and it is flagged as such.

Together, these assumptions give the composite case a baseline of approximately 800 alerts per day, of which roughly 760 (95 per cent) are false positives under the Section 3 definition and roughly 40 (5 per cent) result in a SAR. That consumes approximately 467 analyst-hours per day at the 35-minute average investigation time, and the great majority of it, roughly 443 hours, is spent on alerts that do not result in any disclosure.

4.3 The risk-based framework applied

This section sets out the risk-based framework the case applies, independent of the composite case itself. It draws directly on the threshold-tuning and segmentation literature reviewed in Section 2, and combines five elements.

- Risk-based segmentation: customers are tiered (for example, low, medium and high risk) using established risk factors, including geography, product usage, politically exposed person status, business type and channel, consistent with the risk-based approach formalised in FATF Recommendation 1 and transposed into UK law [3]. This replaces monitoring the whole book under one uniform rule set.
- Threshold tuning: detection thresholds are set per segment rather than uniformly across the whole customer book, following the multi-dimensional optimisation approach Gupta et al. apply to the industry's "above-the-line" threshold-setting practice. This coordinates thresholds across overlapping scenarios and customer segments to reduce redundant alerts without lowering suspicious-activity coverage [7].
- Above-the-line/below-the-line testing: before any threshold change is implemented, it is tested against a holdout sample of historical alerts split into those that would still trigger ("above the line") and those that would no longer trigger ("below the line"), which quantifies the effect on both false positives and true positives before the change goes live. Badics et al. formalise this as a microeconomic decision problem, reporting an 18 per cent reduction in false positives while retaining 98 per cent of true positives on real bank data using this method [8].
- Periodic rule rationalisation: rules are reviewed on a defined cycle and retired or merged where their alert output essentially never converts to a SAR, instead of being left in place indefinitely once implemented.
- Hybrid detection and review: automated detection is paired with targeted manual review, rather than either full automation or full manual review of every alert, following the pattern Halford et al. describe at a UK bank: moving from reviewing 100 per cent of alerts manually to hibernating a share and auto-escalating another share [13]. Bakry et al.'s finding that even a well-performing suppression model incorrectly closed out 11 per cent of genuinely suspicious customers in their test data is treated here as a standing caution against automating suppression beyond what the evidence for a given segment supports [14].

4.4 Application process

The framework in Section 4.3 is applied to the composite case in the sequence a compliance team would realistically execute it.

- Baseline assessment: establish current-state metrics for the composite bank, including alert volume, false-positive rate, SAR conversion rate and investigation time, using the assumptions set out in Section 4.2 as the starting point against which every later step is measured.
- Segmentation: apply the risk-based segmentation from Section 4.3 to the composite bank's customer base, defining risk tiers and assigning the existing alert population across them.
- Tuning cycle: for each segment, calibrate detection thresholds to that segment's own risk profile rather than the single, institution-wide setting assumed in the baseline, using the multi-dimensional optimisation approach set out above.
- Testing: before any proposed threshold change goes live, test it against a holdout sample of historical alerts using the above-the-line/below-the-line method, and quantify the resulting change in alert volume and the estimated impact on true-positive retention.
- Rationalisation: review rule-level performance across the tuned system and retire or merge rules that contribute disproportionately to false positives without a corresponding share of the SARs filed.
- Governance review: before implementation, document each change for model-risk and governance sign-off, covering rationale, testing evidence, expected impact and a rollback plan, consistent with the explainability and audit-trail expectations identified in the governance literature reviewed in Section 2.

4.5 Evaluation approach and limitations

"Success" in the illustrative case is assessed against the same three metrics used to set the baseline in Section 4.2: alert volume, false-positive rate and SAR conversion rate, before and after the framework in Sections 4.3 and 4.4 is applied. The before/after change is modelled by applying effect sizes reported elsewhere in the literature for comparable interventions, including Gupta et al. and Badics et al.'s threshold-tuning results, and Halford et al.'s and Eddin et al.'s results for hybrid detection-and-review models, to the composite case's baseline set out in Section 4.2 [7, 8, 12, 13]. The results this produces, presented in Section 5, are modelled outcomes based on reasonable assumptions and published ranges, stated plainly here so there is no ambiguity: they are not empirically observed results from any real system, and they should not be read as a validated prediction of what any specific institution would achieve.

This approach has four limitations that bear directly on how its results should be read. First, the case is illustrative rather than empirical: its inputs come from published benchmarks rather than one institution's real data, so the figures in Section 5 describe a plausible scenario, not a validated finding. Second, the published effect sizes this paper reuses were measured at institutions with different starting conditions, technology stacks and customer bases than the composite case, so applying them here assumes a degree of transferability the underlying studies do not themselves establish. Third, the paper cannot account for institution-specific factors, such as legacy system constraints, existing rule complexity or staffing levels, that would materially affect how much of a published effect size any real institution could realistically capture. Fourth, the governance and explainability considerations raised in Section 4.4 are discussed qualitatively rather than tested, since no live system is involved in this paper.

Together, these limitations mean the paper's contribution is a transparent, replicable methodology and a plausible illustration of its effect, not a validated predictive claim about false-positive reduction at any specific institution. This distinction comes up again in the Conclusion and is taken up directly in Section 7 (Future Scope), which sets out how the framework could be tested against real institutional data.

5. Results & Discussion

Illustrative before/after results

Applying the modelled effect sizes set out in Section 4.5 to the composite case's baseline from Section 4.2 gives the following illustrative before/after picture. These are modelled outcomes, not measured results. They show what the published effect sizes cited in Section 2 would imply if realised in full against this paper's baseline assumptions, not what any real institution has achieved.

The framework set out in Section 4.3 centres on threshold tuning and rule rationalisation rather than full machine learning deployment, so the primary scenario below applies Badics et al.'s reported effect for above-the-line/below-the-line threshold optimisation: an 18 per cent reduction in false positives while retaining 98 per cent of true positives [8]. Applied to the composite case's baseline of 800 alerts a day, 760 of them false positives and 40 resulting in a SAR, this reduces false positives to approximately 623 a day. True positives fall only marginally, to approximately 39 a day: a loss of roughly one genuine escalation a day, which Section 3 identifies as an unavoidable feature of this trade-off, not a flaw specific to this method.

Table 1: Illustrative before/after metrics for the composite case, modelled from published effect sizes

Metric	Baseline	After tuning (Badics et al. [8])	Upper bound: ML triage (Eddin et al. [12])
Alerts / day	800	662	188
False positives / day	760	623	152
SARs / day	40	39	36
False-positive rate	95%	94.10%	80.90%
SAR conversion rate	5%	5.90%	19.10%
Analyst-hours / day	467	386	110
Analyst-hours saved / day	—	81 (17%)	357 (76%)

For contrast, the table's third column applies Eddin et al.'s more aggressive graph-based triage result, an 80 per cent reduction in false positives while retaining over 90 per cent of true positives [12], as an upper bound on what a fuller machine learning deployment could achieve against the same baseline, not as this paper's chosen outcome. The gap between the two columns is the practical question a compliance function actually faces. Tuning-only interventions of the kind Section 4.3 sets out are achievable inside an institution's existing rule engine, while a result closer to the upper bound requires the kind of new model build, validation and governance overhead described throughout Section 2, for a materially larger return. Halford et al.'s account of a UK bank moving from reviewing 100 per cent of alerts manually to hibernating 19 per cent and auto-escalating 18 per cent, cutting the time to file a SAR by 61 per cent, sits between these two. It is a process change rather than a pure volume reduction, and a reminder that speed to disclosure is a separate metric from raw alert count, one that matters in its own right to a supervisor assessing whether genuine risk is being escalated promptly [13].

Why the framework moves the metrics the way it does

Each element of the framework set out in Section 4.3 contributes to the result through a different mechanism, and the modelled numbers above are best read as their combined effect, not any single lever's. Segmentation reduces volume at source by tiering the customer book so that inherently lower-risk segments are not monitored under the same thresholds as higher-risk ones. That is what makes the subsequent tuning step tractable: tuning one threshold across an unsegmented book means every adjustment is a compromise between segments with genuinely different risk profiles. Threshold tuning itself, tested above-the-line/below-the-line before implementation as Badics et al. and Gupta et al. describe, produces most of the false-positive reduction in the primary scenario, because it targets the specific scenarios and segments generating disproportionate false positives instead of adjusting every rule uniformly [7, 8]. Rule rationalisation compounds this over time. Retiring rules whose alert output essentially never converts to a SAR removes a source of false positives that ongoing tuning cannot fully address, since a badly designed rule can still generate large alert volumes even at its best available threshold. Hybrid detection and review, the fourth element, does not change the underlying false-positive rate at all. It changes what happens to the alerts that remain, redirecting the analyst time freed up by the first three elements toward the higher-risk alerts that most need scrutiny, which is the mechanism behind Halford et al.'s time-to-SAR improvement rather than their alert-volume reduction [13].

Regulatory and governance implications

None of these changes can be made unilaterally by a compliance analytics function. Russell's review of a decade of FCA enforcement notices finds that JMLSG guidance, the industry's primary interpretive reference for the risk-based approach, already overemphasises customer due diligence relative to the wider risk-governance process supervisors expect institutions to demonstrate [6]. Pontes et al.'s interviews with UK practitioners identify the same gap between what the risk-based approach requires in principle and how it is supervised in practice [5]. Applied to this paper's framework, that means every threshold change made under Section 4.4's tuning cycle needs to be evidenced, not just implemented: the rationale for the change, the above-the-line/below-the-line testing evidence behind it, its expected impact on both false positives and true positives, and a documented rollback plan, all retained in an audit trail a supervisor can review after the fact. Rule rationalisation carries a parallel governance requirement, since retiring a rule is itself a model change requiring the same sign-off as introducing one. None of this is unique to the framework this paper sets out. It follows directly from the explainability and model governance expectations identified across the literature reviewed in Section 2, applied here to a tuning-based rather than a machine-learning-based intervention.

Practical implementation challenges

Three practical constraints stand between the modelled results above and what an institution could realistically capture. First, segmentation and tuning are only as reliable as the customer and transaction data behind them. Incomplete or inconsistent KYC data undermines risk tiering before any threshold work begins, and this dependency is invisible in a modelled result built from published effect sizes rather than from data of variable quality. Second, every threshold or rule change identified in Section 4.4 has to pass through change control before it reaches production, and the above-the-line/below-

the-line testing cycle that makes a change defensible also takes calendar time that an institution under regulatory or commercial pressure may be reluctant to spend. Third, second-line and compliance sign-off is a genuine resourcing cost, not a formality. Gilson's finding that many institutions consider even proven tuning approaches under-resourced relative to full machine learning deployment applies with equal force in the other direction: a tuning programme still requires dedicated second-line capacity to review and approve each cycle of change, capacity that competes with the same institution's other compliance priorities [25]. None of these constraints invalidate the framework. They are the reason Section 4.5 presents its results as a plausible illustration rather than a guaranteed outcome, and they are picked up again in Section 7's scope for testing the framework against real institutional data.

6. Conclusion

This paper set out to answer two questions: how a risk-based tuning methodology changes alert volume and the false-positive rate in a representative UK transaction monitoring case, and what governance considerations applying that methodology raises. Modelled against the composite case's baseline, the primary scenario shows that threshold tuning and rule rationalisation alone, without any new machine learning deployment, can be expected to cut alert volume by around a sixth and return roughly 81 analyst-hours a day to higher-value work, while giving up only a small share of genuine escalations. The governance considerations are equally clear from Section 5. None of that gain is available without an evidenced audit trail for every threshold change, sign-off proportionate to what is, in substance, a model change, and second-line capacity dedicated to reviewing it. The operational cost of the approach is real, not incidental to it.

The framework's core value is that it offers a structured, risk-based route to reducing false positives without weakening detection. The literature reviewed in Section 2 and the modelled results in Section 5 point the same way: segmentation, tested threshold tuning, periodic rule rationalisation and hybrid detection-and-review are not competing techniques but complementary ones. Applied together, they move the false-positive rate without materially eroding the institution's ability to catch genuine risk. That is a modest claim by design. This paper does not claim to have solved the false-positive problem, only to have set out, and illustrated, a defensible way to work on it.

Because the framework and every figure behind it are drawn from published, non-confidential sources rather than from any single institution's data, its applicability is not limited to the illustrative case used here. A compliance or risk function at a real UK retail or commercial bank running comparable rule-based monitoring could adapt the same sequence, baseline assessment, segmentation, tuning cycle, testing, rationalisation, governance review, to its own book, provided it is prepared to resource the governance work Section 5 identifies alongside the tuning work itself. What this paper cannot do is confirm what scale of benefit a real institution would actually capture. Section 7 sets out how that gap could be closed.

7. Future Scope

This paper's central limitation, noted in Section 4.5, is that its results are modelled rather than measured. The most direct next step is therefore empirical validation: applying the framework set out in Section 4.3 to one real institution's actual threshold and alert data, under a confidentiality and ethics arrangement suited to that institution's regulatory obligations, and comparing the resulting before-and-after figures against this paper's modelled projections in Section 5. That comparison would test directly how much of the effect sizes reported by Badics et al. and Bakry et al., both of whom worked with real bank data under an institutional arrangement of exactly this kind, actually transfer to a different institution's book, rather than assuming the transferability that this paper's baseline necessarily takes on faith [8, 14].

A second direction is extending the framework rather than replacing it. This paper deliberately confines itself to threshold tuning, segmentation and rule rationalisation, and treats full machine learning deployment as a contrasting upper bound rather than as the framework it applies. The gap between the two columns of Table 1 in Section 5 is, in effect, the size of that opportunity. Future work could study a staged path that layers a scoring or triage model, of the kind Jullum et al., Eddin et al. and Halford et al. describe, on top of an already-tuned rule-based system, rather than treating ML adoption and rule tuning as competing options [11, 12, 13]. Gilson's finding that many institutions see full ML deployment as too costly or resource-intensive relative to simpler tuning suggests this staged approach may be the more realistic adoption path for the kind of mid-sized institution this paper's composite case represents. That is worth testing directly, not assumed [25].

A third direction is a cross-institution comparative study once suitable data-sharing arrangements exist. This paper's own limitations, set out in Section 4.5, include that published effect sizes were measured at institutions with different starting conditions, technology stacks and customer bases, and that this paper cannot say how much of any given effect size a specific institution could realistically capture. A study spanning several UK institutions of different sizes, run for example through a regulator-sponsored data-sharing pilot or sandbox arrangement of the kind used in other UK financial services contexts, could test whether the framework's effect sizes hold consistently across institutions or vary systematically with size and existing system complexity. A single-institution study, however well validated, cannot establish that on its own.

Each of these three directions targets a specific limitation acknowledged earlier in this paper, rather than a general call for further research: single-institution validation tests the modelled-versus-measured gap from Section 4.5, the staged ML extension tests the tuning-versus-full-deployment trade-off set out in Section 5, and the cross-institution study tests the transferability assumption underlying the composite case itself. Taken together, they describe a realistic route from the illustrative methodology this paper sets out to a validated, generalisable one.

References

[1] Rees, A. (2026) 'State of false positives in AML screening: 2026 report', Facctum. Available at: https://www.facctum.com/blog/aml-false-positive-report.
[2] Ketenci, U.G., Kurt, T., Önal, S., Erbil, C., Aktürkoğlu, S. and Şahin, H.İ. (2019) 'A Time-Frequency Based Suspicious Activity Detection for Anti-Money Laundering', IEEE Access.
[3] Savona, E.U. and Riccardi, M. (2019) 'Assessing the risk of money laundering: research challenges and implications for practitioners', European Journal on Criminal Policy and Research.
[4] Ernst & Young LLP (2024) Anti-money laundering (AML) transaction monitoring 2024 survey report. Available at: https://www.ey.com/content/dam/ey-unified-site/ey-com/en-uk/industries/financial-services/documents/ey-uk-aml-tm-survey-12-2024.pdf.
[5] Pontes, R., Rocha, F. and Mac Carthaigh, M. (2021) 'Anti-money laundering in the United Kingdom: new directions for a more effective regime', Journal of Money Laundering Control.
[6] Russell, M. (2025) 'What are the key components of an effective methodology for conducting business-wide risk assessments for money laundering?', Journal of Financial Compliance.
[7] Gupta, A., Chaudhary, P. and Nayak, R. (2021) 'Threshold fine-tuning of money laundering scenarios through multi-dimensional optimization techniques', Journal of Money Laundering Control.
[8] Badics, T., Huszár, A. and Kiss, F. (2022) 'Integral representation method based efficient rule optimizing framework for anti-money laundering', Journal of Money Laundering Control.
[9] Chen, Z., Van Khoa, L.D., Teoh, E.N., Nazir, A., Karuppiah, E.K. and Lam, K.S. (2018) 'Machine learning techniques for anti-money laundering (AML) solutions in suspicious transaction detection: a review', Knowledge and Information Systems.
[10] Jensen, R.I.T. and Iosifidis, A. (2022) 'Fighting Money Laundering With Statistics and Machine Learning', IEEE Access.
[11] Jullum, M., Løland, A., Huseby, R.B., Ånonsen, G. and Lorentzen, J. (2020) 'Detecting money laundering transactions with machine learning', Journal of Money Laundering Control.
[12] Eddin, A.N., Bono, J., Aparício, D., Polido, D., Ascensão, J.T., Bizarro, P. and Ribeiro, P. (2021) 'Anti-Money Laundering Alert Optimization Using Machine Learning with Graphs', ArXiv preprint.
[13] Halford, E., Gibson, I., Newfield, M. and Dhanwala, M. (2025) 'Developing a scoring model for managing money laundering transactions using machine learning', Journal of Money Laundering Control.
[14] Bakry, A.N., Alsharkawy, A.S., Farag, M.S. and Raslan, K.R. (2023) 'Automatic suppression of false positive alerts in anti-money laundering systems using machine learning', The Journal of Supercomputing.
[15] Alkhalili, M., Qutqut, M.H. and Almasalha, F. (2021) 'Investigation of Applying Machine Learning for Watch-List Filtering in Anti-Money Laundering', IEEE Access.
[16] Reite, E.J., Karlsen, J. and Westgaard, E.G. (2024) 'Improving client risk classification with machine learning to increase anti-money laundering detection efficiency', Journal of Money Laundering Control.
[17] Alexandre, C. and Balsa, J. (2023) 'Incorporating machine learning and a risk-based strategy in an anti-money laundering multiagent system', Expert Systems with Applications.
[18] Weber, P., Carl, K.V. and Hinz, O. (2023) 'Applications of Explainable Artificial Intelligence in Finance — a systematic review of Finance, Information Systems, and Computer Science literature', Management Review Quarterly.
[19] Kute, D.V., Pradhan, B., Shukla, N. and Alamri, A. (2021) 'Deep Learning and Explainable Artificial Intelligence Techniques Applied for Detecting Money Laundering – A Critical Review', IEEE Access.
[20] Fritz-Morgenthal, S., Hein, B. and Papenbrock, J. (2022) 'Financial Risk Management and Explainable, Trustworthy, Responsible AI', Frontiers in Artificial Intelligence.
[21] Gerlings, J. (2025) 'The Relevance of Explainable Artificial Intelligence (xAI) in High-Risk Decisions', PhD dissertation.
[22] Turksen, U., Benson, V. and Adamyk, B. (2024) 'Legal implications of automated suspicious transaction monitoring: enhancing integrity of AI', Journal of Banking Regulation.
[23] Oztas, B., Cetinkaya, D., Adedoyin, F., Budka, M., Aksu, G. and Dogan, H. (2024) 'Transaction monitoring in anti-money laundering: A qualitative analysis and points of view from industry', Future Generation Computer Systems.
[24] Oztas, B., Cetinkaya, D., Adedoyin, F., Budka, M., Aksu, G. and Dogan, H. (2023) 'Perspectives from Experts on Developing Transaction Monitoring Methods for Anti-Money Laundering', 2023 IEEE International Conference on e-Business Engineering (ICEBE).
[25] Gilson, C. (2024) 'Are the old ways of transaction monitoring dead?', Journal of Financial Compliance.
[26] Zavoli, I. and King, C. (2021) 'The Challenges of Implementing Anti-Money Laundering Regulation: An Empirical Analysis', The Modern Law Review.
[27] Pol, R.F. (2020) 'Anti-money laundering: The world's least effective policy experiment? Together, we can fix it', Policy Design and Practice.
[28] Ogbeide, H., Thomson, M.E., Gonul, M.S., Pollock, A.C., Bhowmick, S. and Bello, A.U. (2023) 'The anti-money laundering risk assessment: A probabilistic approach', Journal of Business Research.
[29] B.K., M. and Ramasubramanian, V.H. (2025) 'Anti money laundering system in detecting and preventing money laundering activities: a systematic review', Journal of Money Laundering Control.
[30] Castelao-López, J., Corzo Santamaría, T. and Lagoa-Varela, D. (2025) 'Analysis of the main techniques and tools to combat money laundering: a review of the literature', Journal of Money Laundering Control.
[31] Tiwari, M., Gepp, A. and Kumar, K. (2024) 'Trade-based money laundering: a systematic literature review', Journal of Accounting Literature.

[32] Naheem, M.A. (2019) 'Anti-money laundering/trade-based money laundering risk assessment strategies – action or re-action focused?', Journal of Money Laundering Control.

Author Profile

Malki Senevirathne is a risk and compliance professional based in the United Kingdom with 12 years of experience in the financial industry. She is an ICA-qualified professional and holds an MSc from the University of Westminster, an MBA from Cardiff Metropolitan University, and a BBMgt (Bachelor of Business Management) from the University of Kelaniya, Sri Lanka. Her professional interests include anti-money laundering transaction monitoring and financial crime risk management.