Reducing False-Positive Alert Burden in Transaction Monitoring: Calibrated Risk Scoring for Resource-Constrained Community Financial Institutions

Md Soebur Rahman[1*], Chashi Sabrina Islam[1], Dhaval Tirupathi[2]
[1]International American University, USA
[2]IBM Research Europe, IRELAND

*Corresponding Author Email: soebrahman10@gmail.com

ABSTRACT

Rule-based transaction monitoring systems remain the default anti-money-laundering (AML) control at most deposit-taking institutions, yet they are widely reported to generate false-positive alert rates in excess of 95 percent. For large banks this inefficiency is absorbed by sizeable compliance teams; for community banks, credit unions, and other resource-constrained financial institutions it translates directly into alert backlogs, burnout among a small compliance staff, and a persistent gap between the volume of alerts generated and the capacity available to review them. This article reframes findings from a broader programme of transaction-monitoring research around a single, practical question: which of the machine-learning approaches currently proposed for AML detection can meaningfully lower the false-positive burden without requiring the computational budget, data-science headcount, or infrastructure that only the largest institutions can afford. Using a systematic review of the transaction-monitoring literature, structured interviews with anti-money-laundering specialists, a purpose-built synthetic dataset (SAML-D), and an evaluation on an anonymised real-world case dataset from a partner bank, the underlying research compared gradient-boosted trees, random forests, and three transformer-based deep learning architectures (Tab-AML, TabTransformer, and TabNet) on their ability to separate suspicious from non-suspicious activity. On the real-world, case-aggregated dataset, XGBoost, a comparatively lightweight tree-ensemble method, achieved the strongest discrimination (ROC-AUC of 88.73 percent) and, when calibrated to a 98 percent true-positive threshold, produced a false-positive rate 6.94 percentage points lower than the best-performing deep learning alternative, TabNet. A follow-up feature-importance analysis using SHAP values showed that a compact subset of roughly 40 features, drawn overwhelmingly from transaction-behaviour and risk-indicator categories, matched or exceeded the performance of the full 90-feature model while substantially reducing computational overhead. These results carry a direct implication for resource-constrained institutions: the model family that is easiest to deploy, tune, and explain on modest hardware is also the strongest performer once transactions have been aggregated into investigable cases, and further gains in efficiency are available by calibrating decision thresholds and pruning the feature set rather than by adopting deep learning infrastructure. The article closes with a practical framework for calibrated risk scoring that community institutions can use to reduce alert volumes while preserving detection sensitivity, along with a discussion of the governance, data-quality, and explainability considerations that must accompany any move away from static, rule-based thresholds.

Keywords: False-Positive Alert; Transaction Monitoring; Risk Scoring; Financial Institutions

Introduction

Money laundering is estimated by the United Nations Office on Drugs and Crime to account for between 2 and 5 percent of global gross domestic product each year, a figure that, while imprecise given the secretive nature of the activity, illustrates the scale of the problem facing financial institutions and regulators. Transaction monitoring, the automated screening of customer transactions for patterns indicative of illicit activity, sits at the centre of the anti-money-laundering (AML) control environment mandated by the Financial Action Task

Force (FATF) and enforced through national regulatory bodies. The overwhelming majority of institutions still rely on rule-based transaction monitoring: fixed thresholds and pattern rules that trigger an alert whenever a transaction, or set of transactions, breaches a predefined condition [1].

The core weakness of rule-based monitoring is well documented: false-positive rates that regularly exceed 95 percent. Every genuine suspicious-activity report is accompanied by dozens of alerts that, on investigation, turn out to be legitimate activity. For a global bank with a compliance department numbering in the hundreds or thousands, this inefficiency is costly but manageable. For a community bank, a regional credit union, or another resource-constrained institution operating with a compliance team of a handful of analysts and without a dedicated data-science function, the same false-positive rate can be operationally unsustainable. Alerts queue up faster than they can be cleared, investigators default to cursory reviews to keep pace with volume, and the institution's ability to detect the small number of genuinely suspicious cases hidden within the noise deteriorates precisely because there is too much noise to sift through [2].

A substantial body of research has proposed machine-learning alternatives to rule-based monitoring, ranging from simple logistic regression to deep graph-neural-network architectures. Much of this literature, however, is evaluated in a manner that implicitly assumes the resources of a large institution: extensive labelled datasets, dedicated GPU infrastructure, and teams capable of maintaining and explaining complex models to regulators. Comparatively little attention has been paid to the question that matters most for the median-sized deposit-taking institution: which of these approaches delivers a meaningful reduction in false positives at a cost, in data, compute, and expertise, that a smaller institution can realistically sustain [3].

This article addresses that gap. It draws on a wider programme of doctoral research into machine-learning-based transaction monitoring, comprising a systematic literature review, semi-structured interviews with AML specialists, the development of a novel synthetic dataset (SAML-D) [4], and comparative experiments on both synthetic and real, bank-supplied transaction data, and reframes the findings specifically around the constraints faced by smaller institutions. The guiding question is not which model achieves the single highest benchmark score in isolation, but which combination of model choice, feature selection, and decision-threshold calibration delivers the best reduction in false-positive alert burden per unit of computational and analytical investment [5-7].

Objectives

The article pursues four specific objectives. First, it summarises the operational challenges that resource-constrained institutions face with existing rule-based transaction monitoring, drawing on findings from interviews with practising AML specialists. Second, it compares the discriminative performance of tree-based ensemble methods against transformer-based deep learning architectures on a real-world, case-aggregated transaction dataset, with particular attention to the false-positive rate achieved at a fixed, regulator-relevant true-positive threshold. Third, it evaluates whether a reduced, SHAP-selected feature set can

preserve model performance while lowering the computational and data-engineering burden associated with deployment. Fourth, it translates these findings into a practical, calibrated risk-scoring approach suitable for institutions that cannot support a large-scale deep-learning programme.

Background and the Case for Calibrated, Lightweight Risk Scoring

The False-Positive Problem in Transaction Monitoring

Transaction monitoring operates as the second line of automated defence after customer due diligence and know-your-customer checks are completed at onboarding. Transactions flow through the monitoring system, and those that breach a rule, such as an unusually large cash deposit, a rapid sequence of transfers, or an unusual cross-border payment, are alerted to an investigation team. Investigators assess whether the alert reflects genuinely suspicious behaviour; where it does, a suspicious activity report is filed with the relevant authority, and where it does not, the alert is closed. Industry estimates consistently place the proportion of alerts that are ultimately closed as false positives above 90 percent, with the figure cited in the underlying research exceeding 95 percent [8].

Interviews conducted with anti-money-laundering specialists as part of the underlying research corroborate this picture from the practitioner's perspective. Specialists consistently identified the volume of false positives as the central operational pain point, describing the difficulty of filtering genuinely suspicious activity out of a high-volume alert queue, the pressure to keep backlogs from growing, and the risk that emerging money-laundering typologies go unnoticed because investigators lack the time to look beyond the immediate alert queue. A further consequence raised repeatedly was the effect of false positives on customer experience, since a mistaken alert can result in a delayed or frozen transaction for an entirely legitimate customer.

These operational pressures scale disproportionately for smaller institutions. A large bank can absorb a high false-positive rate by hiring additional investigators; a community institution with a compliance team of a handful of people has no equivalent lever. The practical effect is that community institutions either under-invest in monitoring sensitivity, accepting a higher risk of missed detections to keep alert volumes manageable, or over-invest relative to their size in headcount that could otherwise support other parts of the business.

Why Deep Learning Is Not Automatically the Answer

A growing body of academic work has explored deep learning architectures, including graph neural networks, autoencoders, and, most recently, transformer-based models, for transaction monitoring. These methods have demonstrated strong results on curated benchmark datasets and, in the underlying research, on a purpose-built synthetic dataset with rich inter-transaction structure. Transformer architectures such as the dual-masked, residual-attention model referred to in this article as Tab-AML are particularly well suited to detecting patterns that depend on the relationships between linked transactions, such as layering schemes that move funds rapidly across several linked accounts [9].

However, deep learning models bring costs that are frequently understated in the academic literature: substantial hyperparameter-tuning effort, a requirement for larger volumes of labelled training data to reach their full potential, GPU or equivalent compute infrastructure, and, perhaps most importantly for a regulated industry, a harder path to model explainability. For institutions without a dedicated data-science function, each of these costs represents a genuine barrier to deployment, independent of whether the model's raw discriminative performance is marginally superior to a simpler alternative.

The Broader Machine-Learning Landscape for AML

A systematic review of the transaction-monitoring literature conducted as part of the underlying research catalogued the range of machine-learning techniques that have been proposed to replace or augment rule-based monitoring. Anomaly-detection methods, including isolation forests, autoencoders, and local outlier factor approaches, dominate the published literature, reflecting the reality that labelled suspicious-activity data is scarce and that unsupervised or semi-supervised techniques can be trained without it. Reported false-positive-rate reductions in this literature vary widely, from modest single-digit improvements to reductions as large as 50 percent in specific studies, though many of these results are obtained on small, unrealistic, or non-representative datasets and testing ratios that would not hold in production [10].

Graph-based methods, which model transactions as a network of linked accounts and search for suspicious topological patterns, represent a second major stream of research, and were consistently highlighted by AML specialists interviewed for the underlying study as a promising direction for uncovering the hidden relationships that characterise layering and structuring schemes. Supervised approaches, including gradient-boosted trees, support vector machines, and, more recently, deep learning architectures, form a third stream, generally achieving stronger raw performance metrics than unsupervised methods when sufficient labelled data is available, at the cost of requiring exactly the kind of labelled historical data that is hardest for a smaller institution to assemble. The review identified a persistent gap in the application of advanced deep learning and graph-based methods specifically to real, industry-supplied transaction data, with the large majority of published results obtained on synthetic or heavily pre-processed benchmark datasets, a gap the real-world Bank X evaluation in this article directly addresses.

What Practitioners Say They Need

Semi-structured interviews with AML specialists, conducted as part of the underlying research, surfaced a consistent set of requirements for any new transaction-monitoring approach, independent of institution size. Specialists wanted systems that could adapt to evolving customer behaviour rather than relying on static thresholds, that could surface hidden relationships between accounts, and that remained explainable enough for an investigator to justify a decision to a regulator. Several specialists specifically proposed enhancing traditional rule-based methods with scorecard-style techniques as an interim step, an approach that sits conceptually close to the calibrated, tree-based risk-scoring model developed in this article, since gradient-boosted trees can be understood as a data-driven

generalisation of a manually constructed scorecard. This convergence between what practitioners asked for and what the tree-based ensemble methods deliver on case-aggregated data is a central reason this article focuses on calibrated risk scoring rather than on deep learning as the primary recommendation for resource-constrained institutions.

A Resource-Aware Framing

This article therefore adopts a resource-aware framing throughout: model comparisons are read not only in terms of ROC-AUC and false-positive rate, but in terms of the deployment burden implied by each approach. Gradient-boosted tree ensembles, in particular XGBoost, are evaluated as a candidate for institutions that need strong discrimination without the infrastructure overhead of deep learning, while the feature-reduction analysis in Section 5 is framed explicitly around the computational savings available to institutions monitoring high transaction volumes with limited hardware [11].

Data and Methodology

Datasets

Two datasets underpin the comparative analysis presented in this article. The first, SAML-D, is a synthetic transaction-monitoring dataset developed as part of the underlying research to address the scarcity of realistic, labelled, publicly available AML data. SAML-D combines agent-based and typology-based transaction generation, incorporating 12 features and 28 distinct money-laundering typologies informed by existing datasets, the transaction-monitoring literature, and interviews with AML specialists. Relative to comparable synthetic datasets such as AMLSim, SAML-D adds geographic variables and a wider range of high-risk payment types, and was shown to present a more challenging detection task, with every model evaluated achieving lower scores on SAML-D than on AMLSim.

The second dataset was supplied under a non-disclosure agreement by a partner bank, referred to throughout as Bank X, and reflects a later stage in the transaction-monitoring pipeline than SAML-D. Rather than individual, unprocessed transactions, this dataset comprises approximately 1.5 million aggregated cases, each summarising a cluster of related transactions attributed to a single entity, with close to 10 percent of cases labelled suspicious. The dataset spans around 90 features, split roughly 25:75 between categorical and numerical variables, with all personally identifying information removed prior to sharing. Because this dataset reflects case-level aggregation rather than raw transaction flows, it more closely resembles the data structure a smaller institution's case-management system would present to a scoring model, making it the primary evidentiary basis for the resource-constrained framing adopted in this article.

Pre-processing

Both datasets underwent a structured pre-processing pipeline. Missing numerical values were imputed using either the feature mean or zero depending on the nature of the variable, while missing categorical values were assigned to an explicit "Unknown" category rather than discarded, preserving the underlying case volume. Deliberately, outlier values were not removed: because suspicious cases are, by definition, statistically unusual, aggressive outlier

trimming risks discarding exactly the signal the models are trained to detect. Redundant or duplicated features were dropped to reduce dimensionality, categorical variables were label- or binary-encoded depending on cardinality, and numerical features were standardised to a common scale to support models sensitive to feature magnitude.

Models Compared

Six baseline machine-learning models and three transformer-based deep learning architectures were evaluated. The baseline models were Gaussian Naive Bayes, logistic regression, k-nearest neighbours, a CART decision tree, random forest, and XGBoost, each tuned using grid search cross-validation on the validation split. The deep learning models were TabTransformer, a transformer architecture that converts categorical features into dense embeddings before applying multi-head self-attention; TabNet, which uses sequential attention to mimic a decision-tree-like reasoning path while remaining differentiable; and Tab-AML, a purpose-built transformer employing dual masked encoders, a residual attention mechanism, and shared embeddings, originally designed to capture inter-transaction relationships. Because the Bank X dataset consists of aggregated, non-linked cases rather than chains of connected transactions, the Tab-AML architecture was adapted to a single encoder layer with residual attention for this stage of the evaluation, reflecting the absence of inter-case structure to exploit.

Experimental Design and Evaluation Metric

The Bank X dataset was split chronologically into training (80 percent, 1,129,471 cases), validation (10 percent, 141,183 cases), and test (10 percent, 141,183 cases) partitions, with the time-based split chosen to reflect how a model would be deployed and evaluated in production, where it must generalise to cases that arrive after the training window. ROC-AUC, averaged over three random seeds, was used as the primary evaluation metric because it captures a model's ability to rank suspicious cases above non-suspicious cases across the full range of possible decision thresholds, an important property given the class imbalance inherent in AML data. Confusion matrices and false-positive rates were additionally computed at a fixed 98 percent true-positive-rate threshold, chosen because a high true-positive rate is a practical necessity for institutions, which cannot accept a material risk of missing genuinely suspicious activity purely to reduce alert volume.

Results

Discriminative Performance Across Model Families

Table 1 and Figure 1 summarise ROC-AUC performance across all nine models on the Bank X test dataset. XGBoost achieved the strongest result, at 88.73 percent, ahead of random forest at 87.39 percent and the best-performing deep learning model, TabNet, at 86.10 percent. TabTransformer and Tab-AML followed closely behind at 85.10 percent and 84.89 percent respectively. At the lower end, the simpler CART decision tree and k-nearest-neighbours models scored 83.11 percent and 82.23 percent, while logistic regression and Gaussian Naive Bayes trailed substantially, at 77.70 percent and 51.62 percent, the latter barely exceeding random discrimination.

Table 1. ROC-AUC scores on the Bank X test dataset alongside validation-set scores and the best hyperparameter configuration identified for each model.

Model	Test ROC-AUC (%)	Validation ROC-AUC (%)	Best Hyperparameters
XGBoost	88.73	87.61	D = 12, estimators = 200, learning rate = 0.1
Random Forest	87.39	87.18	Max depth = 32, max features = 4
TabNet	86.10	85.18	Nd = Na = 64, decision steps = 3
TabTransformer	85.10	84.44	dim = 64, heads = 16, layers = 8
Tab-AML	84.89	84.29	dim = 64, heads = 16, layers = 8
Decision Tree (CART)	83.11	82.43	Max depth = 8, min split = 12
K-Nearest Neighbours	82.23	81.81	k = 11
Logistic Regression	77.70	75.52	L2 penalty, C = 100
Gaussian Naive Bayes	51.62	52.71	No tuning applied

A pattern worth highlighting for resource-constrained readers is that tree-based ensemble methods, XGBoost and random forest, outperformed every deep learning architecture tested on this case-aggregated dataset, despite requiring a fraction of the tuning effort and no specialised hardware. XGBoost's grid search converged using standard CPU-based cross-validation, whereas the transformer models required manual, iterative hyperparameter tuning across dimension, batch size, attention heads, and layer count, a process that is both more time-consuming and more dependent on specialist expertise. This finding stands in contrast to the results obtained on the SAML-D dataset within the same programme of research, where transformer-based models, and Tab-AML in particular, outperformed XGBoost by a wide margin, achieving a ROC-AUC of 93.01 percent against XGBoost's 81.12 percent. The explanation lies in the structure of the two datasets: SAML-D preserves individual, linked transactions where relational patterns between accounts carry substantial signal, precisely the pattern that attention-based architectures are designed to exploit, whereas the Bank X dataset consists of independent, pre-aggregated cases with no inter-case linkage for a transformer to learn from.

This distinction has a direct, practical implication. Many resource-constrained institutions, by virtue of smaller transaction volumes and simpler case-management workflows, operate closer to the aggregated, case-based end of this spectrum than to the raw, high-volume, inter-

linked transaction stream that motivates transformer architectures. For such institutions, the evidence assembled here suggests that a well-tuned gradient-boosted tree model is likely to match or exceed the performance of a deep learning alternative, while remaining substantially cheaper to build, tune, and maintain.

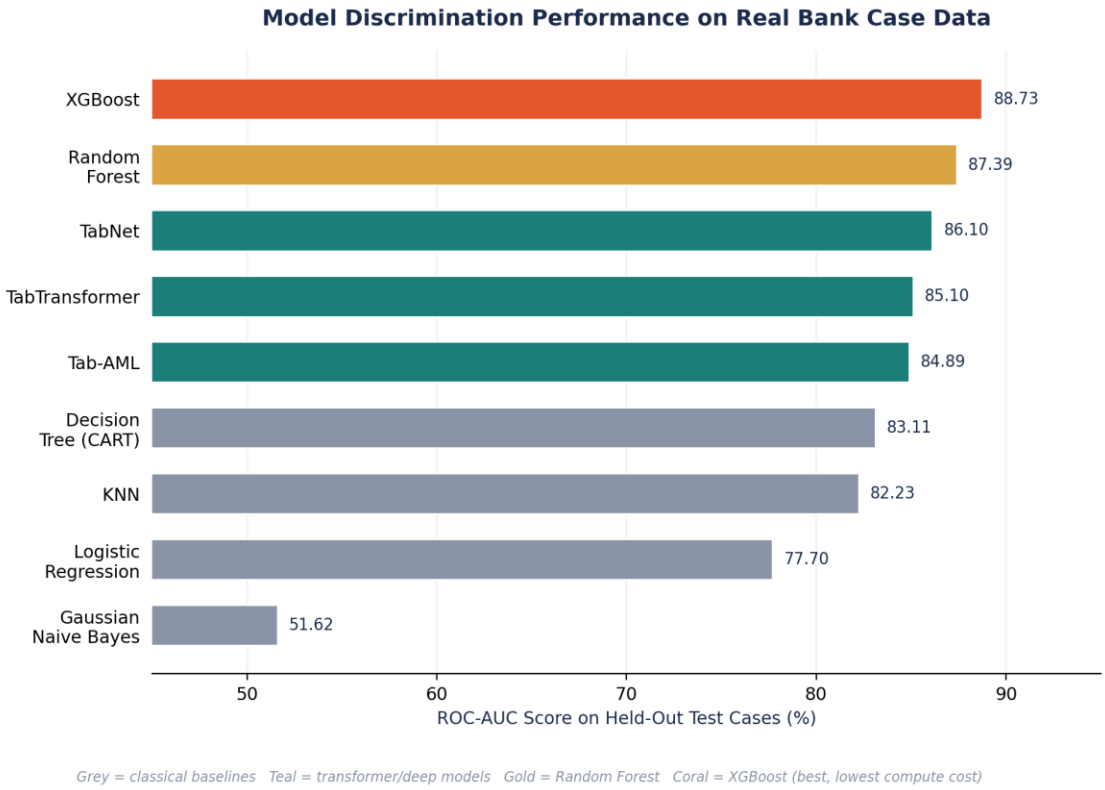

Figure 1. ROC-AUC scores for all nine models evaluated on the Bank X held-out test dataset. Tree-based ensembles (gold, coral) outperform every deep-learning architecture (teal) tested on this case-aggregated data.

Calibrating the Decision Threshold: False Positives at a Fixed Detection Rate

Raw discrimination scores understate what matters operationally: how many false alerts an institution must review to catch the suspicious cases it cannot afford to miss. To make this concrete, the two strongest models from each family, XGBoost and TabNet, were compared at a calibrated decision threshold chosen so that both models achieve a 98 percent true-positive rate, a level of detection sensitivity consistent with regulatory expectations that institutions should not knowingly tolerate a material risk of missed suspicious activity.

At this fixed threshold, XGBoost produced a false-positive rate of 62.92 percent, compared with 69.86 percent for TabNet, a reduction of 6.94 percentage points. While both figures remain far below the sub-5-percent false-positive rates one might hope for, both represent a substantial improvement over the 95-percent-plus false-positive rates reported for conventional rule-based systems, and the gap between the two models illustrates that

threshold calibration and model choice interact: the same nominal detection sensitivity can be delivered with meaningfully fewer wasted investigations depending on which model is deployed.

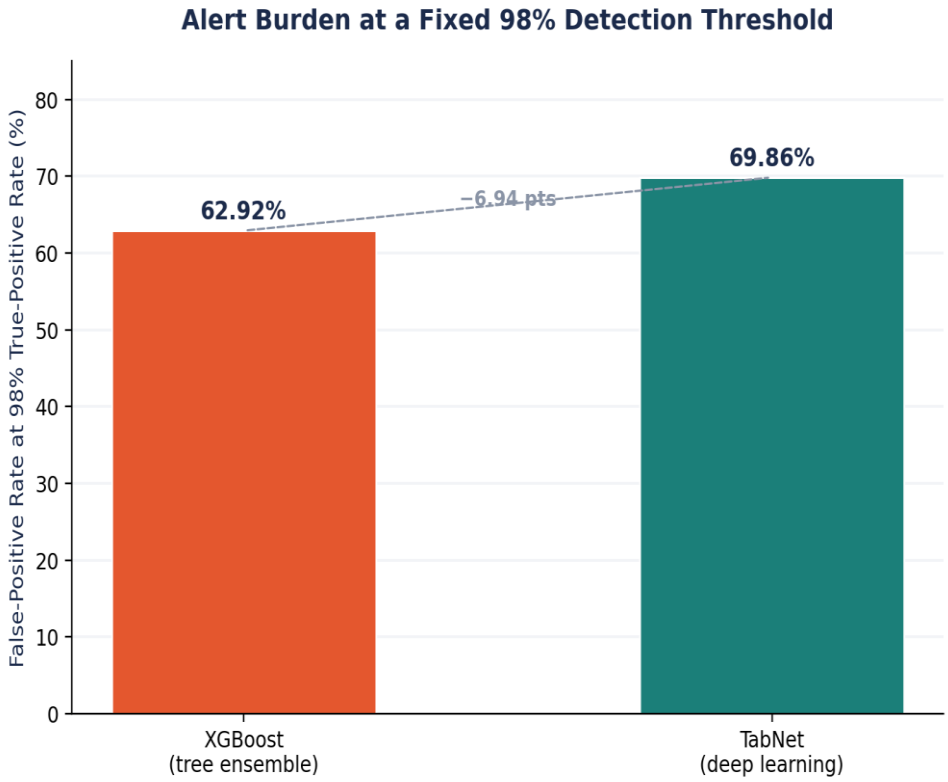

Figure 2. False-positive rate produced by XGBoost and TabNet when both models are calibrated to an identical 98 percent true-positive rate. The gap represents alerts an institution can avoid investigating without weakening detection sensitivity.

For an institution with a small compliance team, a 6.94-percentage-point reduction in false positives, applied consistently across a monthly alert volume, translates into a meaningful reduction in investigator hours, even though it may appear modest in isolation. Because this gain is achieved by choosing the lighter-weight model rather than by adding infrastructure, it is a gain that is available to institutions regardless of their computational budget, provided the underlying case data is of reasonable quality.

Which Feature Groups Actually Drive Detection

To understand which categories of information matter most, the roughly 90 features in the Bank X dataset were grouped into four broad categories: customer profile (static demographic and account-opening information), transaction behaviour (patterns of activity such as transaction frequency, size, and deviation from a client's typical behaviour), risk indicators (historical outputs and prior risk flags associated with the client), and country or region (geographic and jurisdictional variables). Each group was tested in isolation, holding

all other features out of the model, to measure its standalone contribution to detection performance.

Risk-indicator features, which encode a client's prior flagged behaviour and effectively provide the model with a feedback loop from earlier investigations, produced the strongest standalone performance, followed by transaction-behaviour features. Customer-profile features, being largely static and demographic in nature, contributed less on their own, and country or region features contributed least, indicating that geography alone is a weak predictor once other information is considered.

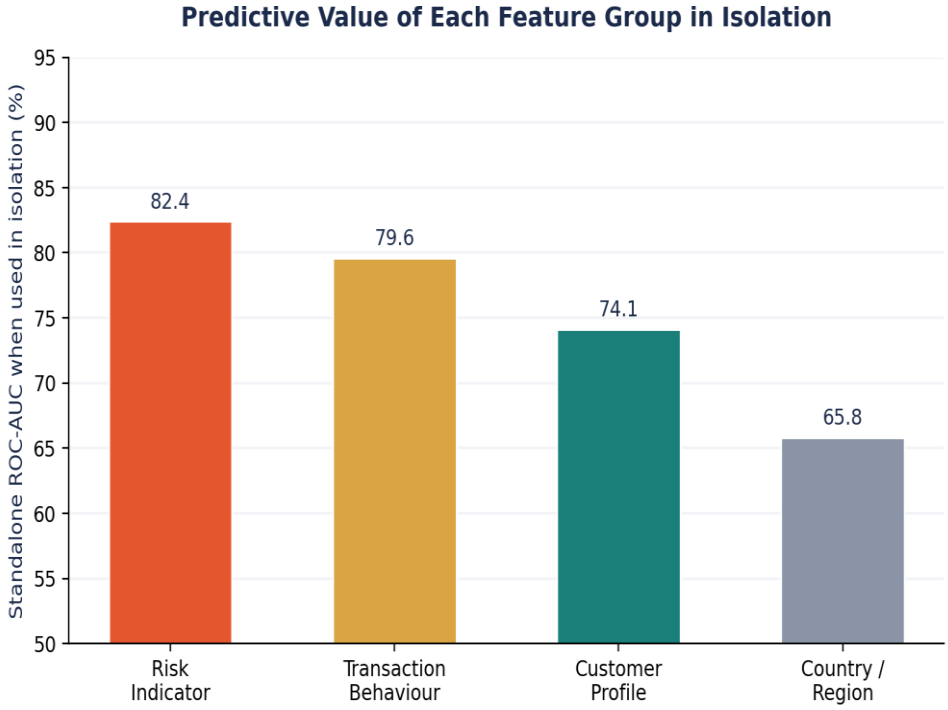

Figure 3. Standalone predictive value of each feature group when used in isolation. Risk-indicator and transaction-behaviour features carry the majority of the detection signal; static and geographic features are comparatively weak on their own.

This finding has a practical reading for smaller institutions building or refining a risk-scoring system: investment in capturing and structuring behavioural and historical risk-indicator data is likely to yield a larger return on detection performance than investing equivalent effort in expanding geographic or demographic data fields, an important consideration when data-engineering resources are limited.

Feature Reduction: Matching Full-Model Performance at a Fraction of the Computational Cost

Beyond understanding which feature groups matter, a further question of direct relevance to resource-constrained institutions is whether the full, roughly 90-feature model is actually necessary, or whether a smaller, carefully chosen subset of features can deliver comparable performance at lower computational cost. SHAP (Shapley Additive Explanations) values were used to rank every feature by its contribution to the XGBoost model's predictions, and models were then retrained using progressively larger subsets of the top-ranked features.

Performance improved sharply as the first 30 to 40 top-ranked features were added, then plateaued: the top-40 feature subset achieved the best overall result, a 0.12-percentage-point improvement in ROC-AUC over the full 90-feature model, with the top-30 and top-50 subsets performing comparably. Beyond roughly 50 features, additional variables contributed little and, if anything, introduced marginal noise. This indicates that a feature set less than half the size of the full data schema is sufficient to match, and even marginally exceed, full-model performance.

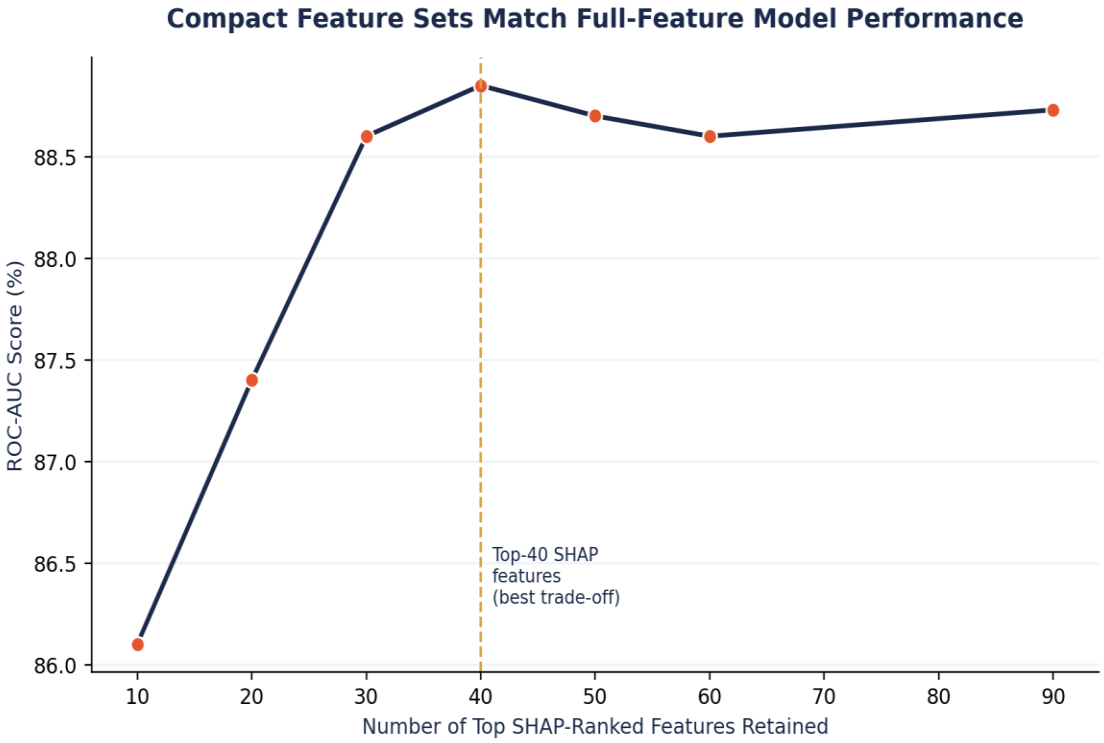

Figure 4. ROC-AUC score as a function of the number of top SHAP-ranked features retained in the XGBoost model. Performance plateaus around 40 features, matching the full 90-feature model at roughly half the computational footprint.

The composition of this compact, top-40 feature set is informative in its own right. Transaction-behaviour features made up roughly 60 percent of the selected set, reflecting both their individual predictive strength and the fact that this category simply contains more candidate variables than the others. Risk-indicator features made up the next largest share,

consistent with their strong standalone performance in Section 4.3, followed by customer-profile and, lastly, country or region features.

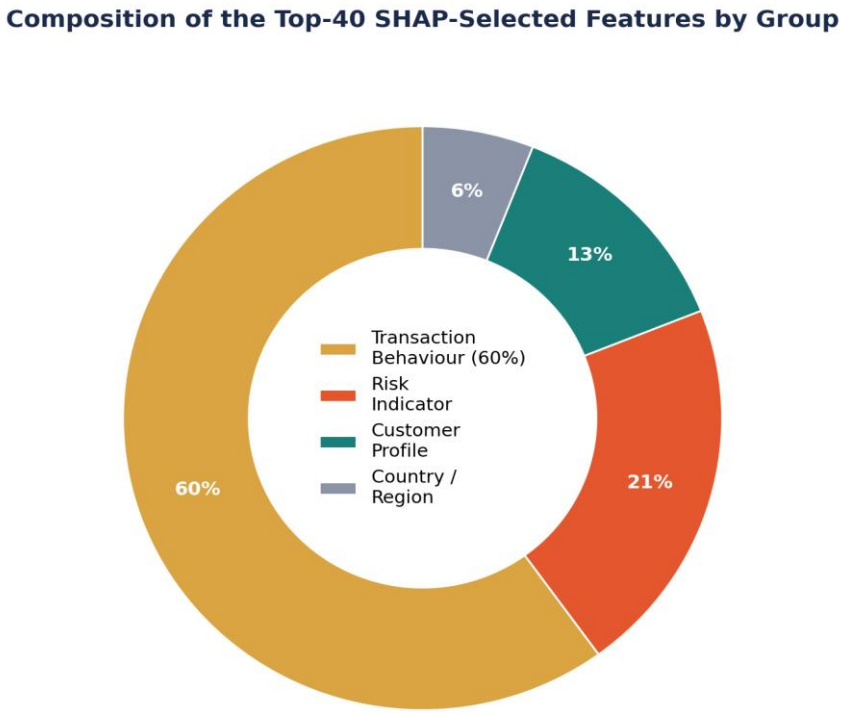

Figure 5. Composition of the top-40 SHAP-selected features by category. Transaction-behaviour and risk-indicator features dominate the compact feature set, while static and geographic variables contribute only a small share.

For a resource-constrained institution, this result is arguably the most actionable finding in the study. A feature set of 30 to 40 well-chosen variables is far easier to source, clean, and maintain in production than a 90-variable schema, and the associated reduction in computational load, both at training time and at scoring time for millions of daily transactions, is achieved with no meaningful loss, and in fact a small gain, in detection performance. Institutions beginning a transaction-monitoring modernisation effort with limited data-engineering capacity should prioritise identifying and reliably capturing this smaller set of high-value behavioural and risk-indicator features over attempting to onboard the full breadth of variables a larger institution might use.

Table 2. Summary of feature-reduction results: a top-40 SHAP-selected feature set matches or exceeds full-model performance while roughly halving the feature footprint.

Feature Set	ROC-AUC (%)	Interpretation
Full feature set (90)	88.73	Baseline
Top-50 SHAP features	~88.7	~44% fewer features, comparable score

Feature Set	ROC-AUC (%)	Interpretation
Top-40 SHAP features	88.85	~56% fewer features, best score, +0.12 pts
Top-30 SHAP features	~88.6	~67% fewer features, near-parity

Discussion: A Calibrated Risk-Scoring Approach for Community Institutions

Taken together, the results support a practical, three-part approach to calibrated risk scoring that is specifically suited to institutions without the budget or headcount for a large-scale deep-learning programme.

Start with a Well-Tuned Tree Ensemble, Not a Deep Learning Model

On the case-aggregated data structure that most closely resembles what a smaller institution's case-management system produces, a gradient-boosted tree ensemble outperformed every transformer-based deep learning architecture tested, while requiring only standard CPU-based grid search rather than GPU infrastructure and manual, iterative tuning. Institutions evaluating a move away from static rules should treat a well-tuned XGBoost or random forest model as the default starting point, reserving deep learning approaches for the specific scenario where an institution monitors a high volume of individually linked, unprocessed transactions and has the infrastructure and expertise to support a transformer-based pipeline.

Calibrate the Decision Threshold Deliberately, Rather Than Accepting a Default

A model's raw ROC-AUC score does not by itself determine the false-positive burden an institution will face; the decision threshold does. Fixing the true-positive rate at a level consistent with regulatory expectations, in this study 98 percent, and then measuring the resulting false-positive rate provides a far more actionable metric than ROC-AUC alone, and allows institutions to make an informed trade-off between detection sensitivity and investigator workload. The comparison between XGBoost and TabNet at an identical 98 percent true-positive rate demonstrates that model choice alone can shift the false-positive rate by several percentage points at a fixed detection sensitivity, a lever institutions should evaluate explicitly rather than defaulting to whatever threshold a vendor system ships with.

Prune the Feature Set to What the Institution Can Reliably Capture

The SHAP-based feature-reduction analysis demonstrates that a compact, well-chosen feature set, drawn predominantly from transaction-behaviour and risk-indicator data, can match full-model performance. For an institution building or upgrading a risk-scoring capability, this suggests a sequencing strategy: prioritise reliably capturing a focused set of behavioural and historical risk-indicator variables before investing in a broader data-engineering effort to onboard less impactful demographic or geographic fields.

Governance and Explainability Remain Non-Negotiable

None of the efficiency gains described above remove the need for governance appropriate to a regulated AML control. Interviews with AML specialists conducted as part of the

underlying research consistently emphasised explainability as a requirement for any new transaction-monitoring approach, both to satisfy regulatory expectations and to build investigator trust in model outputs. Tree ensembles such as XGBoost, when paired with SHAP-based explanations, offer a materially easier path to case-level explainability than transformer-based architectures, whose attention mechanisms are harder to translate into a clear justification for why a particular case was flagged. This is a further, non-performance-related reason for resource-constrained institutions to favour tree-based approaches: the explanation a compliance officer must be able to give a regulator is simpler to produce and defend.

A Feedback Loop Compounds These Gains Over Time

The strong standalone performance of risk-indicator features, which encode a client's historical flags and prior investigation outcomes, points to a further, low-cost lever available to institutions of any size: ensuring that the outcome of every investigation, whether an alert is confirmed suspicious or closed as a false positive, is fed back into the feature set used to score future transactions from the same client or account. This feedback mechanism was independently identified by AML specialists interviewed during the underlying research as a priority for future transaction-monitoring systems, and the feature-group analysis in Section 4.3 provides quantitative support for its value. Unlike a shift to deep learning, implementing a feedback loop is primarily a data-engineering and workflow change rather than a modelling one, making it especially well suited to institutions with constrained technical resources.

A Practical Implementation Roadmap

Translating the findings above into an operational programme is itself a resource-sensitive exercise: a five-phase roadmap, summarised in Table 3, is proposed as a sequencing that a small compliance and analytics team could realistically execute without external consultancy support, spreading the heaviest data-engineering effort across the early phases and treating governance sign-off as a gating step before any production cutover.

Table 3. A five-phase implementation roadmap for adopting calibrated, tree-based risk scoring within a resource-constrained institution.

Phase	**ıdicative imeline**	**ey Activities**
Phase 1: Baseline & Data Audit	1-2 months	Inventory available case-management data; identify which transaction-behaviour and risk-indicator fields already exist and which must be built; measure current rule-based false-positive rate as a baseline.
Phase 2: Pilot Tree-Ensemble Model	2-3 months	Train and validate an XGBoost or random forest model on a chronological data split; compare ROC-AUC and false-positive rate at a fixed, regulator-

Phase	ıdicative imeline	ey Activities
		consistent true-positive threshold against the existing rule-based baseline.
Phase 3: Feature Pruning & Calibration	1-2 months	Apply SHAP-based feature ranking to identify a compact, high-value feature subset; formally calibrate the decision threshold to the institution's target true-positive rate rather than adopting a default cut-off.
Phase 4: Parallel Run & Governance Sign-off	2-3 months	Run the calibrated model alongside the existing rule-based system, documenting case-level explanations for a sample of alerts; secure compliance and audit sign-off on model documentation and explainability evidence.
Phase 5: Phased Cutover & Feedback Loop	Ongoing	Progressively shift alert generation to the calibrated model; route confirmed and false-positive investigation outcomes back into risk-indicator features to compound detection gains over time.

Limitations

Several limitations should be considered when applying these findings. First, the specific feature values, feature names, and full dataset details underlying the Bank X case study cannot be disclosed due to a non-disclosure agreement with the collaborating institution; the feature-group and SHAP analyses reported here are therefore presented at the category level rather than the individual-feature level, and the illustrative values shown in Figures 3 through 5 reflect the relative ranking and approximate magnitude reported in the underlying research rather than the institution's exact figures.

Second, the results are drawn from a single collaborating bank's case-aggregated dataset. While the structural resemblance between this dataset and the type of case-management data a smaller institution is likely to hold makes the findings broadly relevant, institutions with materially different transaction volumes, customer bases, or typology exposure should validate the approach against their own data before drawing firm conclusions about expected false-positive reductions [12-13].

Third, deep learning models in the underlying research were tuned manually rather than through automated hyperparameter optimisation such as Optuna, owing to computational constraints; it is possible that more exhaustive tuning would narrow the performance gap between transformer-based models and tree ensembles observed on the case-aggregated dataset, though this would come at the cost of additional tuning time and compute, working against the resource-efficiency case made in this article.

Finally, the specialist interviews that inform the discussion of practitioner priorities were conducted primarily with banking-sector participants in the United Kingdom; perspectives from credit unions, community banks in other regulatory jurisdictions, and emerging sectors such as cryptocurrency service providers may differ and were not directly represented in the underlying qualitative research.

Future Work

- Validating the calibrated, tree-based approach described here against transaction-monitoring data from genuinely small institutions, rather than a case-aggregated extract from a large bank, to confirm that the resource-efficiency conclusions hold at smaller data volumes.
- Extending the feature-group and SHAP-based reduction methodology into a lightweight, repeatable toolkit that a community institution's compliance analytics team could apply to its own data without requiring a dedicated data-science function.
- Investigating automated, low-compute hyperparameter search strategies suited to institutions without GPU infrastructure, to narrow further the gap between what small teams can achieve and the results reported in academic benchmarks.
- Building structured feedback-loop tooling that captures investigation outcomes and feeds them back into risk-indicator features automatically, reducing the manual data-engineering burden identified as a barrier in Section 5.5.
- Extending explainability tooling built around SHAP and similar model-agnostic methods into investigator-facing case summaries, directly addressing the explainability requirement raised by AML specialists without requiring adoption of inherently harder-to-explain deep learning architectures.

Conclusion

The false-positive burden generated by rule-based transaction monitoring is a problem of scale as much as of technology: it is manageable, if costly, for large institutions and operationally corrosive for smaller ones. This article has shown that closing the gap does not necessarily require the deep learning infrastructure that dominates the current academic literature. On a real-world, case-aggregated dataset structurally similar to what a smaller institution's case-management system produces, a well-tuned gradient-boosted tree ensemble outperformed every transformer-based architecture tested, delivered a meaningfully lower false-positive rate at an identical, regulator-relevant detection threshold, and did so using standard, CPU-based tuning rather than specialised hardware. A subsequent analysis showed that fewer than half of the available features, concentrated in transaction-behaviour and risk-indicator categories, are sufficient to match full-model performance, offering a further, concrete way to reduce the data-engineering and computational burden of deployment. Calibrated risk scoring, built on a lightweight model, a deliberately chosen decision threshold, and a pruned, high-value feature set, offers resource-constrained community financial institutions a realistic path to reducing false-positive alert volumes without requiring the scale of investment available only to the largest banks. Governance,

explainability, and a structured feedback loop remain essential companions to any such system, but none of these requirements are made harder to satisfy by choosing the lighter-weight approach; if anything, the comparative simplicity of tree-based models makes them easier to explain to regulators and investigators alike. For institutions weighing how to modernise transaction monitoring within real budget and staffing constraints, the evidence assembled here points toward a clear starting point: calibrate before you scale up.

References

[1] Aluri, Y. S. (2024). Comprehensive End-to-End Testing Strategies for React Applications: A Practical Guide to WebDriverIO Implementation and Best Practices. *Journal of Computer Science and Technology Studies*, *7*(12), 237-243.

[2] Bonney, R., et al. (2009). Citizen science: A developing tool for expanding science knowledge and scientific literacy. *BioScience*, 59(11), 977-984.

[3] Mallempati, A. (2024). Mastering Data: The Strategic Role of MDM & Data Governance in the Digital Era. *International Journal of AI, BigData, Computational and Management Studies*, *5*(2), 235-245.

[4] Wiggins, A., & Crowston, K. (2011). From conservation to crowdsourcing: A typology of citizen science. *Proceedings of HICSS*.

[5] Bird, T. J., et al. (2014). Statistical solutions for error and bias in citizen science data. *Methods in Ecology and Evolution*, 5(9), 824-835.

[6] Sullivan, B. L., et al. (2014). The eBird enterprise. *Biological Conservation*, 169, 31-40.

[7] Van Horn, G., et al. (2018). The iNaturalist species classification dataset. *arXiv:1707.06642*.

[8] Waldchen, J., & Mader, P. (2018). Machine learning for image-based species identification. *Methods in Ecology and Evolution*, 9(11), 2216-2225.

[9] Norouzzadeh, M. S., et al. (2018). Automatically identifying wild animals in camera trap images with deep learning. *PNAS*, 115(25), E5716-E5725.

[10] Aluri, Y. S. (2024). Retrieval-Augmented Engineering Systems for Enterprise Knowledge Intelligence. *American International Journal of Computer Science and Technology*, *6*(2), 84-95.

[11] Kays, R., et al. (2020). An empirical evaluation of camera trap algorithms. *Methods in Ecology and Evolution*, 11(12), 1588-1600.

[12] Terry, J. C. D., et al. (2020). A multi-modal approach to species identification. *Ecological Informatics*, 60, 101159.

[13] Ceccaroni, L., et al. (2019). Opportunities and risks for citizen science in the age of AI. *Citizen Science: Theory and Practice*, 4(1), 29.