dutch banking association

White Paper

AI
Machine learning in transaction monitoring

strong banks
strong society

Contents

1 Motivation

Banks in the Netherlands have both a societal role and a legal obligation under the Wet ter voorkoming van witwassen en financieren van terrorisme (Wwft) to detect suspicious client transaction behaviour related to money laundering and terrorist financing (ML/TF). One of the key controls for ongoing monitoring is transaction monitoring (TM). TM by banks is performed post-transaction.

Because banks process large volumes of transactions, they rely on automated monitoring systems to detect potential ML/TF risks and generate alerts. These alerts are subsequently reviewed to assess whether the observed transaction behaviour is plausible. When ML/TF risks cannot be reasonably excluded, banks must file a Suspicious Activity Report (SAR) [1] with the Financial Intelligence Unit (FIU). The FIU analyses the SARs, enriches them with additional information and, if deemed to be suspicious, shares them with law enforcement agencies. Banks are also required to take appropriate mitigating measures regarding the client relationship and its associated risks, in line with their individual risk appetite.

The manual handling of TM alerts is time-intensive, requires trained and highly skilled professionals. From both a risk management and effectiveness perspective, it is crucial to allocate resources to the highest ML/TF risks.

Traditionally, TM has been primarily rule-based. A business rule specifies conditions under which an alert should be generated. For example: “Generate an alert if a cash deposit exceeds €10,000.” Although business rules vary in complexity, they typically focus on a single behavioural dimension (e.g., cash usage) and rely on fixed thresholds. Their advantages include interpretability, transparency, clear linkage to risk typologies, and ease of implementation. However, in practice, rule-based systems often produce many false positives – alerts that can be reasonably explained – and may fail to detect more complex suspicious behaviour.

To overcome these limitations, the use of machine learning in TM is gaining traction. Machine learning, a subfield of artificial intelligence, uses algorithms that learn patterns from large datasets to make predictions. Such models can simultaneously analyse multiple behavioural dimensions and identify complex transaction patterns or client characteristics associated with ML/TF risks. As a result, machine learning-based TM can improve the accuracy of risk detection, reduce false positives (enhancing both efficiency and client experience), and detect complex behaviour that rule-based systems struggle to capture.

1 In 2026 Dutch banks still file Unusual Activity Reports to the FIU. This will change to filing Suspicious Activity Reports as of July 2027 under the EU AML Regulation. Methodologies described in this industry white paper are the same for SARs as for UARs.

It is important to stress that in both rule-based and model-based TM, decisions to file a SAR with the FIU are currently made with meaningful human intervention by analysts, not by automated systems.

Although TM has long been part of banks' control framework, the application of machine learning in this domain is relatively new, prompting considerable discussion among internal and external stakeholders. The Dutch Central Bank (DNB) explicitly acknowledges the benefits of responsible AI deployment in TM: *'Another part of the monitoring is done using AI and models. This allows the entity to better detect and investigate potential suspicious patterns and complex transactions'* (GP4.15 of DNB Wwft Q&A and Good Practices). Replacing or augmenting rule-based TM with model-based approaches therefore holds substantial promise for enhancing risk mitigation and improving the resource allocation.

This evolution also affects banks' Ongoing Due Diligence (ODD) frameworks, which aim to mitigate ML/TF risks throughout the business relationship. The Dutch Banking Association (NVB) Industry Baseline ODD emphasises the transition toward automated, trigger-based ODD, supported by Event Driven Reviews (EDRs) and risk-differentiated client reviews. The ODD framework facilitates continuous monitoring of clients – their transactions and behavioural patterns – to detect suspicious activity, maintain accurate client risk classifications, and determine appropriate follow-up actions. This enables banks to apply a risk-based approach by prioritising relevant EDRs instead of performing time-driven periodic reviews by default.

Drawing on banks' experience with developing, implementing, and operating machine learning models for TM, this paper provides insights and recommendations for responsible and effective use of such models in this highly regulated domain. Section 3 explores areas where machine learning appears most beneficial and how these relate to ML/TF risks. Section 4 outlines how to demonstrate that risks are effectively mitigated. As machine learning capabilities evolve rapidly, banks can adopt global best practices, particularly in the domain of responsible AI – which emphasises client impact and safeguards against unfair outcomes. This is examined in section 5. In section 2 an overview is provided of key machine learning concepts and terminology.

2 Machine learning concepts and terminologies

2.1 Model classes in TM

Machine learning models are algorithms that extract information from data to make predictions. Most models used in TM to detect ML/TF risks can be classified into two main categories: supervised and unsupervised learning (defined below). Other modelling approaches – such as graph-based methods or natural language processing – are also employed, but primarily as preprocessing tools whose outputs serve as inputs to the core risk-detection models. This paper focuses on the latter. The overview below provides a concise introduction to key machine learning concepts applied in TM risk-detection systems.

Supervised machine learning refers to models that are trained to predict a predefined target variable or label. The target may be continuous (e.g., price of a house) or categorical (e.g., whether a transaction is fraudulent). During training, the model learns to identify data patterns that are predictive of the target. The underlying data are typically transformed into features – structured inputs suitable for algorithmic processing. Common supervised models relevant for TM include linear and logistic regression, decision trees, random forests, gradient-boosting algorithms, and neural networks.

Unsupervised machine learning, by contrast, is not trained to predict a specific target variable. Instead, it focuses on uncovering structure within the data. In TM, the two most common applications are clustering and anomaly detection. Clustering aims to partition data into subsets (clusters) such that observations within a cluster are similar to each other but distinct from those in other clusters. Applied to transaction behaviour, clustering can reveal groups of clients with similar behavioural profiles. Anomaly detection models pursue the opposite objective: identifying observations that are highly unusual relative to a reference population. Examples of widely used unsupervised models include K-means, DBSCAN, isolation forests, Cluster-Based Local Outlier Factor (CBLOF), and variational autoencoders.

2.2 Ingredients for training a model

Features

Features are the fundamental data elements from which machine learning algorithms extract patterns. They must be carefully designed and implemented prior to model training, with the model's objective in mind. In TM, informative features are those that capture aspects of a client's transaction behaviour relevant for ML/TF risks – e.g., the volume of cash deposits within a defined time window. Features that reflect trusted or low-risk behaviour (e.g., payments to government agencies) can also be valuable when the model's purpose is to distinguish high-risk from low-risk behavioural profiles.

Labels

Labels are the target variables that supervised models learn to predict during training. In TM, labels typically relate to the riskiness or escalation outcome associated with a client's transaction behaviour. Examples used in practice include whether an internal escalation was raised (yes/no), whether a SAR was filed (yes/no), or whether the behaviour ultimately resulted in client offboarding (yes/no). There is no gold standard for labels in TM. Escalating from internal flags to SARs to offboarding does not necessarily imply that the underlying labels become more accurate or more analytically useful.

Models

Machine learning models are statistical algorithms or computational programmes that perform a sequence of predefined mathematical operations. Once trained on historical features and labels, a model can be applied to new, unseen data to generate predictions.

Predictions

Machine learning models generate predictions based on learned patterns in the underlying data. By default, a model does not provide an inherent explanation for why a particular client receives a specific score. Understanding how a model arrives at its output is therefore essential. This transparency is achieved through model explainability techniques (discussed in section 5).

2.3 Governance

The application of machine learning models introduces several new risk dimensions, including the possibility of software or implementation errors leading to inaccurate model outputs, the risk of improper or unintended use of model predictions, and the risk of perpetuating historical biases embedded in the training data. To ensure the safe and responsible deployment of machine learning models, organisations must establish and maintain a comprehensive model governance structure and a robust model risk management framework.

Model governance provides oversight and control across all phases of the model lifecycle and safeguards against privacy, operational, and model-related risks. Key elements of effective governance include continuous performance monitoring, well-defined retraining and update cycles, mechanisms for identifying and mitigating unintended bias, and structured feedback loops that support model improvement. In line with established policy expectations, it is recommended that model governance requirements – such as methodological transparency, documentation of underlying assumptions, implementation details, and risk-mitigation actions – be explicitly incorporated into the technical model documentation [2].

2 See also the NVB Baseline on Technical Model Documentation.

3 Design of TM model effectiveness

3.1 SIRA/risk typology/ML/TF risks

The Wwft requires banks to identify, analyse, and mitigate ML/TF risks to which they are exposed. These risks are consolidated in the Systematic Integrity Risk Assessment (SIRA), which serves as the foundation for designing AML/CFT controls – i.e., the control framework through which these risks are mitigated [3].

AML/CFT controls comprise both automated controls, such as TM, client monitoring, and screening, as well as identification and verification against public data sources such as the Chamber of Commerce. For automated controls in particular, machine learning techniques offer significant opportunities to enhance detection, effectiveness and efficiency.

The SIRA at individual banks can consist of broad risk scenarios that are subsequently decomposed into more granular risk typologies. Both SIRA scenarios and risk typologies relate to ML/TF risks. For clarity and consistency, this paper refers to this collective set simply as ML/TF risks.

3.1.1 Risk manifestations

Risk manifestations describe the concrete ways in which a ML/TF risk becomes visible in client characteristics or transactional behaviour. For example, a structuring risk in the money laundering placement stage may manifest through repeated cash deposits by the same client or related parties within a short timeframe, unusually large cash deposits, or similar patterns indicative of attempts to evade detection.

A clearly defined and comprehensive set of risk manifestations is essential for designing targeted and effective mitigants. Some ML/TF risks – such as basic structuring – have only a small number of identifiable manifestations and can therefore be considered simple. Other ML/TF risks, including more sophisticated schemes like VAT carousel fraud, exhibit a broader and more varied set of manifestations and are thus complex. For these complex risks, it is particularly important that the list of manifestations is both complete (avoiding gaps that allow key behaviours to go undetected) and relevant (ensuring the manifestations genuinely occur in practice). The complexity of a given ML/TF risk determines which mitigation approach is most appropriate, as discussed in section 3.3.

3 For an overview of requirements for an effective SIRA implementation, we refer to Integrity risk analysis (dnb.nl).

3.1.2 Risk exposure assessment

After defining ML/TF risks and their associated risk manifestations, banks should assess their inherent exposure to each risk – that is, both the likelihood of the risk occurring and the potential impact should it materialise. When an inherent risk exceeds the bank's risk appetite, it needs to be mitigated with controls proportionate to the risk.

Using the existing control framework – which includes both rule-based mechanisms and model-based detection – banks should determine the extent to which each inherent risk is mitigated and quantify the remaining residual (or net) risk. If the residual risk still falls outside the risk appetite, additional or enhanced controls should be applied. Where no controls exist for a risk that exceeds the risk appetite, this constitutes a deficiency in the overall AML/CFT control framework. Recommendations on how to perform this assessment are provided in the DNB's guidance on integrity risk analyses (see footnote 5).

Based on this analysis, banks can prioritise improvements to existing mitigants and the development of new ones. Prioritisation should consider the severity of each ML/TF risk – preferably aligned with national priorities – as well as identified control deficiencies and the level of residual risk.

3.2 Integration of machine learning features into ML/TF risks

Features are the foundational building blocks of AI / machine learning models. To integrate such models effectively into a banks existing risk-management framework, it is essential to establish a clear link between model features, and the specific ML/TF risks they are intended to mitigate. This alignment clarifies the risk mitigation purpose of the model and provides the basis for demonstrating its effectiveness in practice (as discussed in section 4).

Constructing high quality features requires close collaboration with subject-matter experts. Discussions should involve:

1. **Analysts**, to understand which data sources and attributes are assessed during investigations and how these are interpreted in context; and
2. **Risk-assessment, design, and compliance teams** (across the first and second lines), to ensure that features align with the bank's risk management framework and cover all relevant ML/TF risks – including those not yet fully mitigated.

Predicting in advance which features will meaningfully contribute to risk detection is inherently difficult. As a result, some features will inevitably be redundant or add little value. Feature selection techniques can be used to remove these non-contributory features. This step is strongly recommended, as it reduces model complexity, minimises the volume of (personal) data processed, lowers computational requirements, and improves model interpretability. It also lowers the chance of 'overfitting', meaning the model is less likely to latch onto quirky coincidences in the historical data that do not hold up in real life – helping the model stay reliable when it encounters new clients and transactions.

3.3 Machine learning approaches in relation to ML/TF risks

The ML/TF risks can be broadly classified into known and new/unknown risks.

3.3.1 Known ML/TF risks

Some known ML/TF risks have only a small number of risk manifestations that adequately capture the underlying behaviour. Structuring, introduced in section 3.1.1, is a typical example. Such low dimensional or simple risks can be mitigated with rule-based monitoring solutions. These involve constructing business rules (if–then logic) that generate alerts once a risk manifestation becomes sufficiently pronounced – typically by breaching a predefined threshold. However, even for simple risks, there is no guarantee that rule-based controls will provide effective and efficient coverage. In some cases, machine learning models may still be required to reach acceptable performance levels. Whether a standalone rule is sufficient should therefore be validated through testing and effectiveness assessments.

For many risks, it is challenging to design rule-based controls that capture only relevant suspicious behaviour. As a result, rule-based systems frequently produce large numbers of false positives. This is particularly true for complex or high dimensional ML/TF risks, which often consist of many interacting and nuanced risk manifestations – none of which provides sufficient control when addressed in isolation.

Machine learning models can be developed to mitigate these more complex risks. In such cases, multiple features are constructed to capture all or part of the relevant risk manifestations linked to the ML/TF risk. When sufficient historical examples of suspicious behaviour exist, supervised machine learning models generally offer the most effective and efficient mitigation technique. Depending on the modelling strategy, institutions may develop either a general model or a set of focus (or agent-based) models, as outlined below.

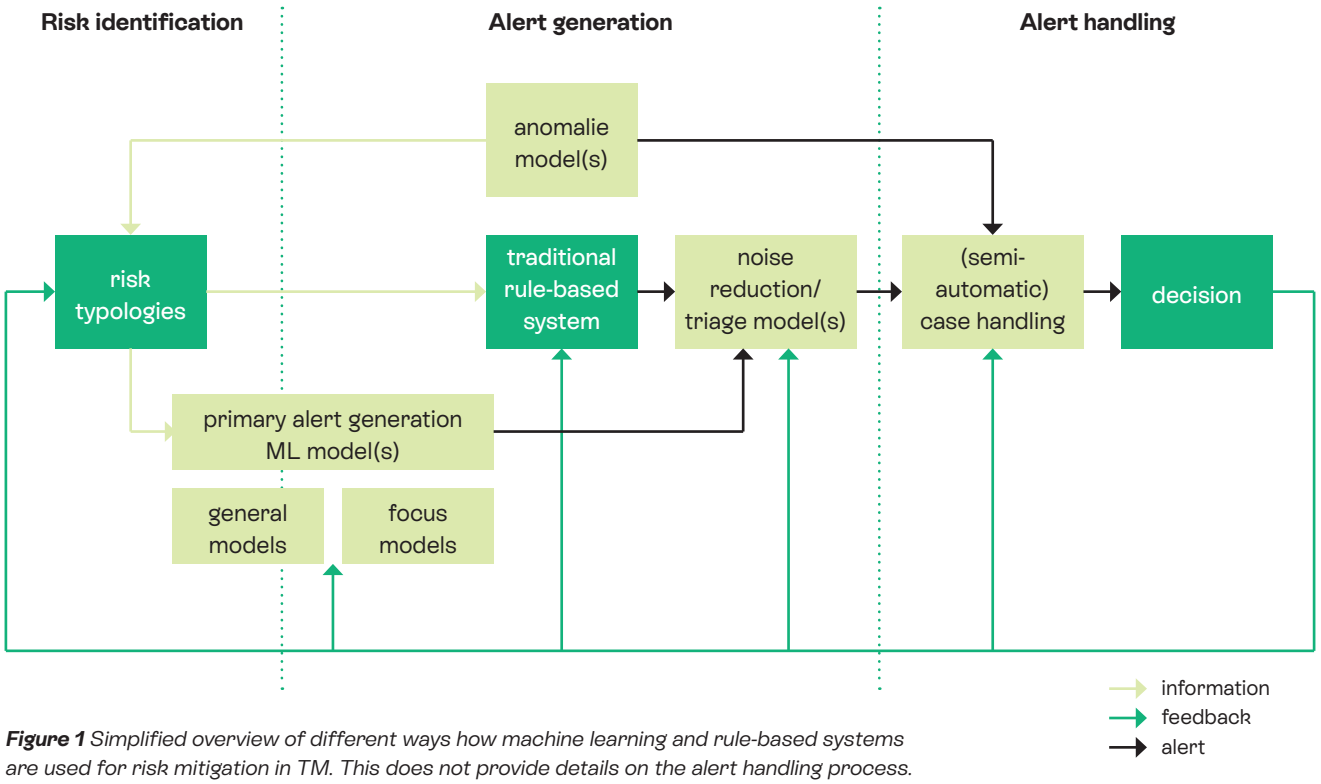

Figure 1 *Simplified overview of different ways how machine learning and rule-based systems are used for risk mitigation in TM. This does not provide details on the alert handling process.*

General model

One approach is to use a single, generalised machine learning model that incorporates multiple feature groups. Each feature group represents the risk manifestations of a particular ML/TF risk. For example, one group may include indicators of structuring in the placement phase (e.g., number of cash deposits per week or per month), while another may capture behaviour related to high risk geographies. In addition to transaction-based features, client-level characteristics – such as risk classifications, PEP/UBO indicators, or adverse media hits – can also be included.

Depending on the availability of labels, the model may be trained using supervised, unsupervised, or semi-supervised methods. In practice, the choice between these approaches is driven by how many reliable training labels exist: supervised models require a substantial number of confirmed historical examples, semi-supervised models can work with only a limited set of labelled cases supplemented by large amounts of unlabelled data, while unsupervised models can be used without any suitable label. Model training should follow established machine learning practices (see section 7).

When using supervised learning, the model identifies combinations of risk manifestations that are strongly associated with positive labels. For example, a client who exhibits both structuring-like cash deposits and subsequent transfers to high risk jurisdictions may be more indicative of suspicious behaviour than clients who exhibit these characteristics separately.

A key challenge of general models lies in assessing and steering their coverage of individual ML/TF risks. Evaluating this requires dedicated effectiveness assessments (see section 4). Although general models allow for flexible detection of complex combinations of risk manifestations, they offer less direct control over individual risks. Despite this, they can serve as a robust base model alongside which more targeted focus models can be deployed.

Focus model

An alternative is to train multiple smaller models – each focused on a single ML/TF risk. Each model uses only the features relevant to that specific risk. These focus models, or agents, aim to quantify the degree to which a particular ML/TF risk manifests. For example, one agent may quantify cash-based structuring risk while another quantifies behavioural changes related to high risk geographies. The agent can be a machine learning model or a rule-based scoring system, depending on the complexity of the risk. An ensemble model can then be trained on the outputs of all agents to capture combinations of ML/TF risk exposures.

A key advantage of this modular approach is the level of control it provides. Different alert thresholds can be defined for different agents, aligned with the severity and expected frequency of each risk. Improvements can be targeted at individual agents: a poorly performing agent can be rebuilt and re-integrated without redesigning the entire detection system. This structure also enhances transparency for analysts, who receive a clearer picture of which risk(s) drove a particular alert.

However, a limitation of the agent-based approach is its difficulty in capturing interactions between risks, as each agent is trained independently. Where such interactions are important, dedicated feature groups can be introduced to capture them. Additionally, if many ML/TF risks require monitoring, the number of focus models can become large, creating challenges in scalability and ongoing maintenance.

3.3.2 New/unknown ML/TF risks

Criminals continuously adapt their methods of ML to avoid detection and disrupt interventions. As a result, new risk typologies emerge that are initially unknown to banks and authorities. When a bank becomes aware of a newly emerging risk and the associated residual risk exposure exceeds its risk appetite, monitoring can begin using targeted rule-based controls, scorecard-style approaches, or unsupervised machine learning models – depending on the complexity of the behaviour and the availability of information.

When new risk typologies are not yet known, the key challenge is how banks can proactively detect them. Rather than relying on chance identification through isolated investigations, banks can use machine learning techniques to search for early signals of emerging crime trends.

Unsupervised machine learning models provide a mechanism for identifying anomalous behaviour. If such a model uses features in which the emerging behaviour manifests, it may generate alerts pointing toward the new risk. However, the challenge is that banks do not know in advance which behavioural dimensions will contain these signals, and existing features may not capture them. Banks can address this by periodically experimenting with new features, reviewing the anomalies produced, gathering insights, and shifting focus to different behavioural aspects over time. Nonetheless, anomalous behaviour frequently turns out not to be suspicious, and unsupervised models generally produce more false positives than supervised ones. The search space can be narrowed – for instance, by restricting analysis to clients with substantial transaction volumes – to focus detection on areas of higher inherent risk.

Once a new risk typology and its manifestations are identified, it must be formally incorporated into the SIRA and associated ML/TF-risk taxonomy as a new risk event. From there, dedicated supervised machine learning models or targeted business rules can be developed to provide effective and sustainable mitigation.

In summary, effectively addressing ML/TF risks requires a diverse set of complementary detection models; no single approach can cover all risks. Continuous development is needed to incrementally strengthen the TM landscape as new risks emerge. Importantly, this does not imply that all components must be in place before enhancing, or partially replacing, traditional rule-based systems with model-based TM controls (see section 4.4).

3.4 Operationalisation approaches

Once a machine learning model is trained, it can be deployed in two main ways within the TM framework: as a noise reduction (triage) model or as an alert generation model.

Noise reduction/triage

When sufficient volumes of rule-based alerts have been reviewed and their outcomes are known, a supervised model can be trained to predict the probability of a rule-based alert being a true hit. Such a noise-reduction or triage model can then score newly generated rule-based alerts. A probability threshold is set: alerts with scores below this threshold are automatically closed and therefore not manually handled.

Because no model can achieve perfect detection performance, the institution must establish a clear TM risk appetite regarding the level of potentially missed suspicious behaviour. Banks should also demonstrate – using the same effectiveness analyses applied in a non-modelled TM environment (e.g., coverage assessments) – that after applying noise reduction, the underlying risk continues to be effectively mitigated.

A supervised model can incorporate a wide range of client- and transaction characteristics, allowing it to capture subtle and complex behavioural patterns. While model performance can be demonstrated on historical data, ongoing monitoring in production is essential to detect model drift, the degradation of performance over time as behaviour evolves. If drift occurs, the model must be retrained and potentially re-validated, or decommissioned. Even without drift, periodic retraining (e.g., annually) is advisable so that the model remains aligned with the latest behavioural developments and financial crime trends.

Such a noise-reduction model closes likely false positives based on a statistical argument, i.e. that the risk of a rule-based alert of becoming a SAR is below a certain probability threshold value. To increase the interpretability of such a model, it is beneficial to include features that directly relate either to the riskiness of behaviour (e.g., weekly cash deposit totals) or to risk reducing indicators (e.g., stable historic behaviour). Mapping features to ML/TF risks, as described in section 3.2, allows analysts to see which ML/TF risks and specific behavioural signals influenced the model's decision.

Alert generation

The same principles apply when models serve as primary alert generators rather than triage mechanisms. Features remain strongly linked to ML/TF risks, and explainability continues to highlight the relevant ML/TF-risk drivers and underlying behaviours.

The distinction is that no pre-existing alerts are being scored. Instead, the model evaluates all clients within its scope. Each client receives a risk score, and those with scores above the alert threshold generate alerts directly. This makes the model an independent detection mechanism rather than a filter on rule-based alerts.

4 Establishment of TM model effectiveness

Operational effectiveness refers to the extent to which a model successfully mitigates the ML/TF risks it is designed to address. Demonstrating this is essential both during model development and on a continuous basis throughout the model's production lifecycle. Monitoring effectiveness is inherently complex: no single metric captures all relevant dimensions of performance. Instead, different metrics highlight different aspects of how well risks are being identified and mitigated. These metrics should be complemented by qualitative interpretation, grounded in the use case and linked explicitly to the underlying ML/TF risks as part of the institution's risk-based approach. The optimal combination of metrics will depend on factors such as model type, model maturity, risk coverage objectives, and the availability of reliable labels. This section presents a set of complementary effectiveness metrics and guidance on when to apply them.

Achieving a TM model with perfect efficiency (i.e., no false positives) and perfect effectiveness (i.e., no false negatives) is generally unattainable. Models operate with incomplete information, meaning banks must navigate an intrinsic trade-off between efficiency and effectiveness. The key objective is therefore to design, operate, and evidence a TM system that supports a risk-based approach, maintaining effective mitigation aligned with the bank's risk appetite.

Firstly, we describe how to measure effectiveness at the model level, referred to as overall effectiveness. Overall effectiveness provides insight into aggregate model performance but may not suffice if a model is designed to cover multiple ML/TF risks. Such a model may perform strongly overall while still leaving specific risks insufficiently mitigated. For this reason, the second subsection focuses on effectiveness at the ML/TF-risk level, which assesses how well individual risks are covered. While this provides a clearer view of risk-specific performance, it does not fully demonstrate the level of residual risk – that is, the behaviour not detected by the model. Methods for assessing the extent of this residual exposure are discussed later in this section.

The overall TM effectiveness can be assessed by examining the number of SARs reported to the FIU and the related transaction volumes over a specific period, split for the different top ML/TF risks and relevant client segments of the bank. This is then related to open source information on ML/TF threats and adjusted for the bank's exposure. Assessing effectiveness over longer periods (e.g., a year) is recommended to enhance accuracy and minimise noise fluctuations.

4.1 Metrics for overall TM model effectiveness

4.1.1 Recall

Recall measures how effectively a model identifies investigation-worthy behaviour – i.e., the share of relevant cases it successfully alerts on. As discussed in section 2.2, investigation-worthy alerts may include cases that previously resulted in SAR filings, as well as cases identified through other internal indicators such as historical review levels. A recall of 100% is generally unattainable and recall alone can be misleading because it treats all missed cases as equally important, even though some behaviours pose higher risk than others. Therefore, recall should always be complemented with a qualitative assessment of the severity and risk level of missed cases.

When sufficient high quality labels exist, institutions may set quantitative recall targets (e.g., 95% recall) to guide model design and ensure consistent, risk-based threshold setting. However, label availability is often limited. In those situations, recall must be supported by qualitative evidence – such as expert assessments or analysis of risk types the model does not capture – to form a complete view of effectiveness.

4.1.2 Precision

Precision reflects the efficiency of the model and is defined as the percentage of investigation-worthy alerts among all alerts generated. Rule-based systems typically have low precision. Precision and recall generally move in opposite directions – increasing one usually reduces the other. For example, lowering the threshold to maximise recall would yield a high volume of false positives and very low precision.

Label quality also affects precision: incomplete labels prevent exact measurement of both recall and, in some cases, precision. During development, partial label sets can approximate performance, though they miss newly detected behaviours and unseen risks. In production, precision is usually the only metric that can be measured directly, while recall must be estimated through methods such as below-the-line sampling or score-curve extrapolation.

4.1.3 Sampling

Sampling can be used to obtain additional labels for assessing a model's performance. Two primary sampling approaches are commonly distinguished, each serving a different purpose and offering different insights.

Below-the-line (BTL) sampling

BTL sampling involves investigating data points with model scores below the alerting cutoff – i.e., clients who would not otherwise generate an alert. Sampling can be performed randomly or using targeted selection criteria. BTL sampling provides insight into behaviour that the model does not surface and therefore helps identify missed investigation-worthy cases. It is particularly valuable for models already in production, as it supports verification that performance has not drifted over time and that the alert threshold remains appropriate.

Above-the-line (ATL) sampling

ATL sampling investigates data points above the alerting threshold. This is especially useful for models targeting new or emerging risks for which no reliable historical labels exist. By examining a sample of these alerted cases, analysts can determine whether the model is indeed capturing the intended behaviour and whether the newly identified patterns warrant further refinement.

Considerations and limitations of sampling

Despite its value, sampling carries two important limitations:

1 **Limited sample sizes**
Investigating sampled cases requires analyst capacity, which is often scarce. As a result, samples tend to be small, leading to a high margin of error. Conclusions drawn from these samples therefore come with considerable uncertainty.

2 **Differences in how sampled alerts are analysed**
Sampled cases are often reviewed using a simplified or accelerated process – for example, without client outreach or using only an initial assessment. This means the resulting labels differ in meaning from standard production labels (e.g., labels from full case investigations). Any metrics derived from these sampled labels must therefore be interpreted with caution, acknowledging these differences.

4.1.4 When to use which metric

As noted in the introduction, demonstrating TM model effectiveness involves nuance. The appropriate assessment method depends on the model's purpose – whether it aims to detect a new ML/TF risk or replace an existing control – and on its stage in the lifecycle (development or production).

The table below provides a non-exhaustive overview of methods for evaluating TM model effectiveness, combining ATL and BTL performance measurements. These methods serve as general guidance; specific models may require additional or alternative approaches.

Effectiveness at the model level is necessary but not sufficient, especially for models designed to cover multiple ML/TF risks. A model may demonstrate strong overall performance while still failing to adequately mitigate one of the underlying risks it targets. Therefore, effectiveness must also be assessed at the ML/TF-risk level, which is described in the following subsection.

Establish overall effectiveness	Target new ML/TF risk	Improving/replacing existing control
During development	• Test samples to estimate performance ATL and BTL.	• Measure recall of previously captured SAR filings (partial recall). • Test samples to estimate performance ATL and BTL.
During production	• Extrapolate performance ATL, potentially complemented by BTL samples. • BTL test samples.	• Monitor number of SARs with regard to previous control (partial recall).

Figure 2 *Non-exhaustive list of methods that can be used to determine TM model effectiveness, depending on the purpose of the model and development phase.*

4.2 Metrics for TM model effectiveness at ML/TF risk level

When a model covers multiple ML/TF risks, effectiveness must also be assessed at the ML/TF-risk level, not just through overall performance. This can be evaluated using three complementary metrics: features per ML/TF risk, alerts per ML/TF risk, and investigation-worthy alerts per ML/TF risk. These metrics should be interpreted together with qualitative, risk-based analysis tailored to the specific model.

4.2.1 Features per ML/TF risk

If model features can be directly or indirectly linked to specific ML/TF risks, the number of features per ML/TF risk can be used to assess how well a model is positioned to address those risks. Features can also be categorised by the type of behaviour they represent – rare versus frequent – based on their value distributions. Features representing rare behaviours may contribute little to overall alert volumes but can be highly relevant for individual cases. Metrics based solely on aggregate alert contribution may therefore underestimate their importance, even though they can capture niche typologies effectively.

Another way to classify features is by examining their correlation with the model score. Strongly correlated features have a greater influence on the score and are key drivers of detection. Weakly correlated features provide additional contextual information. Features negatively correlated with the model score act as risk reducers and should be treated as a separate category. This approach assumes a monotonic relationship between feature values and model scores, a property that can be enforced during model development.

Together, the number and type of features per ML/TF risk provide qualitative insight into whether a model can meaningfully mitigate a given risk. Risks with few or weakly correlated features are unlikely to be well-covered, whereas risks supported by multiple strong features – combined with good overall model performance – are more likely to be addressed effectively.

This analysis is particularly valuable for rare ML/TF risks, which naturally produce few alerts and even fewer investigation-worthy cases. In such situations, alert-based or label-based effectiveness metrics are sparse, making feature-level analysis an important complementary indicator.

4.2.2 Alerts per ML/TF risk

A second way to assess effectiveness at the ML/TF-risk level is to allocate alerts to ML/TF risks based on how strongly each feature contributes to an alert. Using explainability methods such as Shapley, the contribution of each feature to an individual alert score can be calculated and then aggregated across all alerts. By mapping features to their corresponding ML/TF risks, these aggregated feature contributions can be translated into a feature-importance score per ML/TF risk. When normalised by the model's cut-off score, this score can be interpreted as an approximate number of alerts linked to that ML/TF risk.

The resulting alerts-per ML/TF-risk metric provides a quantitative indication of whether the model is effectively surfacing behaviour related to a specific risk. A high volume of alerts associated with a particular ML/TF risk suggests good coverage – provided the underlying feature-to-risk mapping is sound, which must be confirmed through qualitative assessment. This approach is particularly useful for ML/TF risks with limited or no label information.

Conversely, a low number of alerts does not automatically imply poor coverage, as the risk itself may be rare. In such cases, the features per ML/TF-risk analysis offers valuable complementary insight into whether the model is structurally capable of detecting the risk

4.2.3 SARs per ML/TF risk

A third way to assess effectiveness at the ML/TF EC-risk level is to use label information, typically based on SAR filings but also potentially on other relevant indicators (e.g., escalations or client offboarding). For alerts with a SAR label, feature contributions can be calculated using explainability methods such as Shapley and then aggregated, normalised, and mapped to the corresponding ML/TF risks – mirroring the approach used for alerts per ML/TF risk. This produces a SARs per ML/TF -risk metric, which approximates how many SARs the model generates for each risk. If investigators record the ML/TF risk associated with each SAR, this metric can be obtained directly.

When both recall and precision are high and a ML/TF risk has many SARs allocated to it, one can make a strong statement that the model effectively detects the components of that risk addressed by its features. The overall metrics ensure that few investigation-worthy alerts are missed, while the SAR distribution confirms that the model produces SARs in the relevant risk areas. However, the reverse is not necessarily true: a low number of SARs does not imply ineffective detection, as some risks rarely occur. In such cases, effectiveness must be judged using the broader set of qualitative and quantitative metrics. Model-level performance should also always be interpreted in the context of the bank's overall TM effectiveness.

Because label information is often incomplete or unavailable during development, early effectiveness assessments rely on other metrics. Even when historical labels exist, true recall remains uncertain, as past investigation-worthy cases may have gone undetected. To ensure the model remains an effective risk mitigant, its performance must be reviewed periodically through a feedback loop – combining metric monitoring with analyst feedback, financial-crime trend analyses, and below-the-line testing.

4.3 Residual risk

Gaining insight into missed behaviour is essential to determine whether a model captures all relevant suspicious activity. This review must be embedded in a structured feedback loop, forming part of model governance. In this loop, bottom-up insights from model performance are combined with top-down developments – such as updated ML/TF-risk assessments – and external signals (e.g., new regulations or industry guidance) to identify necessary follow-up actions, such as adding new features or adjusting model scope.

During development, the model's recall provides an indication of how many investigation-worthy alerts are missed. Since no model is expected to capture all such cases, banks should apply a risk-based approach, ensuring that no specific ML/TF risk is disproportionately represented in the missed alerts. This requires identifying which ML/TF risks these missed cases relate to and determining whether certain risks are under-detected.

If missed alerts are already linked to ML/TF risks as part of the review process, assessing coverage per risk is straightforward. If not, a (sample of) missed investigation-worthy cases must be manually assigned to the correct ML/TF risk. This assignment cannot rely on model score explanations, because for missed alerts the model score is too low to meaningfully reflect the underlying behaviour. Instead, analysts should use judgement to allocate it to the appropriate risk.

Periodic evaluation of missed-case patterns together with analyst feedback, financial crime trend analyses, and below-the-line testing helps ensure the model remains an effective risk mitigant and that improvements are actioned through the feedback loop.

4.4 Transition

Banks need to determine whether, how, and to what extent they transition from rule-based alert generation to solutions where machine learning is applied – potentially replacing rule-based systems as the primary alerting mechanism. A parallel run is not a default requirement; what matters is demonstrating that the new model provides an effective and efficient risk-based approach to detecting suspicious behaviour. As described earlier, this includes showing improvements in both overall effectiveness and efficiency, as well as across individual ML/TF risks.

As with any detection system, some misses (false negatives) are expected, including cases where the legacy system produces a SARs that the new model does not. Therefore, a “no SAR left behind” standard is unrealistic. Instead, the new machine learning system must demonstrate a net improvement, such as identifying more total SARs, reducing false positives, or showing fewer missed investigation-worthy cases at comparable or better efficiency levels.

When transitioning from rules to models, banks may deactivate individual ML/TF risk scenarios or replace the entire system at once. In either case, deployment requires sufficient evidence and substantiation, following a clear, pre-agreed methodology that enables stakeholders to conclude with confidence that the new model outperforms the rule-based system.

To avoid long or unnecessary parallel runs, all stakeholders must align upfront on the evidence, criteria, and governance needed to demonstrate overall improvement. This ensures the transition is controlled, efficient, and fully defensible.

5 Responsible AI

AI models can significantly enhance TM, but they operate in a highly regulated environment. The key legal frameworks are AML/CFT regulations (such as the EU AML package and the Wwft [4]) and the GDPR.

The EU AML-package requires banks to maintain an effective TM system to detect and report suspicious activity. Improving the efficiency and effectiveness of TM, while complying with AML/CFT rules, is the primary reason to develop AI models.

The GDPR aims to strengthen citizens' right to the protection of their personal data. It ensures that data is processed lawfully, fairly, and transparently, and only for legitimate purposes. This right is not absolute. Recital 4 of the GDPR makes clear that it must be balanced against other public interests, including combating ML and TF – an objective recognised by all EU Member States [5].

Banks therefore have a legal basis to process personal data for AML/CFT purposes, but must still observe GDPR principles such as fairness, transparency, data minimisation, and security. The European legislator acknowledged this balance when drafting the AML Regulation (AMLR), which frequently refers to the Charter of Fundamental Rights and EU data-protection law, such as the Data Protection Regulation (see for example recitals 150 and 151).

AI is not mandated by AML/CFT regulations, but it is increasingly necessary for institutions to meet their obligations. Its use is legitimate under the GDPR when accompanied by safeguards such as purpose limitation, fairness, and privacy-by-design. AI can also help strike the right balance between AML/CFT objectives and protecting individuals' rights.

The AMLR explicitly anticipates the use of AI, explicitly provides a ground for automated decision making and includes safeguards, notably in Article 76. These provisions reinforce that AML/CFT objectives and data-protection rights must be reconciled – an area where challenges remain. One example is transparency: internal model/risk owners, regulators and analysts require insight into model behaviour (see section 5.1), but full disclosure to clients could enable evasion of controls. A recent case involving bunq confirmed that AML/CFT obligations can outweigh an individual's request for detailed model explanations [6]. The appropriate level of transparency remains an ongoing topic of discussion.

4 On the 10th of July 2027, the Wwft will be substituted by the AMLR and the Dutch implementation of the directive contained in the EU anti-money laundering package. The AMLR obliges banks to detect and report suspicious transactions or behaviour to the FIU. "All suspicious transactions, including attempted transactions, should be reported, regardless of the amount of the transaction, and the reference to suspicions should be interpreted as including suspicious transactions, activities, behaviour and patterns of transactions" (recital 136 and Article 69 of the AMLR).
5 Recital 150 of the GDPR.
6 Uitspraken rechtspraak.

GDPR fairness requirements also apply (Art. 5), meaning AI models must avoid discriminatory outcomes, which further constrains model design. This is discussed in section 5.2.

Applicability of the AI Act

The EU AI Act, approved in March 2024, classifies AI systems by risk. High risk systems face strict requirements. Currently, AML/CFT systems are not included in the high risk category.

Legally and from a client protection perspective, the absence of AI Act applicability does not create a protection gap: the GDPR and AMLR already provide strong safeguards tailored to the AML/CFT context. In some cases, when applying AI Act requirements tension with AMLR obligations can arise, for example regarding transparency of the algorithms as discussed above. Nevertheless, many elements of the AI Act (such as risk management, accountability, and human oversight) are already being adopted by the industry due to its maturity in risk management and the need to comply with the GDPR. This is further discussed in sections 5.4 and 5.5.

5.1 Model explainability in TM

The sector recognises the need for transparency in the use of AI systems and shares several common principles.

Transparency toward internal stakeholders and regulators

Banks should be able to explain their monitoring practices to model owners, risk owners, and supervisors. This may require different levels of transparency. The criteria used to detect suspicious behaviour must be explainable, which requires clear documentation of the model's design, methodology, and internal logic. This is often referred to as model transparency.

However, the AMLR explicitly limits transparency toward individuals (see Article 76). Providing detailed information about AI systems to the public – such as through a GDPR access request [7] – could enable criminals to evade detection, as recognised in Recital 146 of the AMLR. Case law is still developing around the boundaries of transparency under Article 23 of the GDPR.

7 In this regard, recital 157 of the AMLR is of relevance: "Access by the data subject to any information related to a suspicious transaction report would seriously undermine the effectiveness of the fight against money laundering and terrorist financing. Exceptions to and restrictions of that right in accordance with Article 23 of Regulation (EU) 2016/679 might therefore be justified. The data subject has the right to request that an authority referred to in Article 51 of Regulation (EU) 2016/679 check the lawfulness of the processing and has the right to seek a judicial remedy referred to in Article 79 of that Regulation. That authority is also able to act on an ex officio basis where provided for under Regulation (EU) 2016/679. Without prejudice to the restrictions to the right to access, the supervisory authority should be able to inform the data subject that all necessary verifications by the supervisory authority have taken place, and of the result as regards the lawfulness of the processing in question."

Interpretability of models

Models must also be interpretable, meaning a human should be able to understand the reasoning behind their decisions. Interpretability can be achieved using explainability techniques tailored to TM machine learning models.

Another important perspective on explainability goes beyond individual models. It involves building a holistic understanding of the entire TM landscape – examining relationships between use cases, dependencies between applications, and the role of machine learning models within the broader ecosystem. In some situations, multiple models or tools must be viewed together to fully explain how ML/TF risks are mitigated. This level of system understanding can be maintained through documentation, architectural overviews, policies, and governance structures.

5.1.1 Model stakeholders and levels of explainability

The purpose and requirements for explainability depend on the role of each stakeholder and their relationship to the model. The table below (non-exhaustive) illustrates how interpretability and transparency can be viewed from the perspective of common stakeholder groups, along with suggested explainability approaches. Different approaches to explainability may be considered, depending on the institution's legal obligations, risk appetite and governance framework.

Role	Purpose of explainability	Explainability approach
Analysts (model user)	Understand why an alert was generated to analyse the potential risk that may be present.	• Global functional documentation. • Detailed local explanations. • Work instructions.
Risk/Model owners	Show how ML/TF risk is mitigated by clarifying when alerts are – and are not – generated, both from an overall effectiveness perspective and on a case-by-case basis to ensure auditability.	• Detailed functional documentation. • Global and local explanations. • Production monitoring framework.
Compliance	Verify that AML/CFT policy is followed and model meets design requirements.	• Detailed functional documentation. • Detailed analysis of residual risk and coverage. • Global explanations. • Production monitoring framework.
Model developers, model validator, regulatory authorities	Technical understanding of the model, code and expected behaviour in production.	• Detailed technical documentation. • Global and local explanations. • Production monitoring framework. • Production code framework.

Figure 3 *Explainability requirements for common stakeholder groups.*

5.1.2 Model interpretability

Explainability techniques help identify why a model makes certain predictions, such as flagging suspicious activity.

Global vs local explainability

- **Global explainability** offers insight into overall model behaviour, usually by summarising feature importance across a broad dataset.
- **Local explainability** clarifies individual predictions by showing which features most influenced a specific model score.

For local explainability, the common approach is to present the key features and their contributions to the score. These top features guide analysts toward the most relevant aspects of a client's behaviour. Mapping features to specific ML/TF risks further strengthens this context.

Clear and concise feature definitions are crucial. Poorly designed features can be misunderstood by analysts and overwhelm them. To avoid this, some banks group related features into feature families. This shifts the focus from technical metrics (e.g., standard deviation of cash deposits) to more intuitive concepts (e.g., abnormal cash patterns over the last three months). Well-designed feature groups improve interpretability, guide investigations more effectively, and can be integrated into TM dashboards to show how often they occur in alerts or lead to SARs.

Model score: 0,97

Feature group	Contribution (%)
Total cash credit last 3 months	80%
Round amounts last month	15%
Total cash debit last month	3%

Figure 4 *Illustrative example of how model explainability can be presented. The model predicts a high score and shows three top feature groups leading to this high score.*

In manual alert handling, humans justify their decisions through narratives. Machine learning models do not use narratives; instead, they provide explanations through features or feature groups and their relative importance. This level of explainability helps stakeholders understand the main drivers behind model-based decisions. The emphasis then shifts to demonstrating overall model effectiveness and ensuring strong governance.

In summary, transparency in TM relies on clear communication, interpretable models, and meaningful explanations. By adopting these principles, banks can strengthen TM while maintaining trust and accountability.

5.2 Fairness in TM

ML/TF risk detection can expose banks to ethical risks, including indirect or perceived discrimination. The AML/CFT domain is particularly vulnerable to bias: for instance, complex ML schemes can involve multiple countries, requiring close monitoring of cross-border transactions. This may lead to uneven alert distributions across client groups (e.g., native vs. non-native clients). If not properly managed, such disparities can result in unfair or insufficiently justified differences in treatment.

As detection methods grow more complex, it becomes harder to assess whether these differences are warranted, especially when underlying logic is less transparent. At the same time, data science techniques offer ways to measure fairness and systematically analyse potential bias.

The legal obligation to mitigate these risks is anchored in Article 5 of the GDPR, which requires that personal data be processed lawfully, fairly, and transparently. The fairness principle already applies to all AI systems used in the AML/CFT, and – together with the GDPR's accountability requirement – obliges banks to demonstrate that their use of AI is fair and compliant.

Though focus is on fairness as dominant risk resulting from AML/CFT model use, other ethical risks can emerge in TM, such as the inability to contest outcomes. Depending on the context and risk profile, banks may consider mechanisms that enable addressing concerns or outcomes where appropriate (e.g. though complaint mechanisms).

5.2.1 Fairness risk assessment

Fairness risks arise when a model produces negative or uneven impacts across stakeholder groups – such as clients, employees, regulators, or society – by creating unequal impact with protected characteristics such as age, gender, or nationality.

The assessment should cover both known risks (e.g., unnecessary outreach to clients) and new, use case-specific risks. Importantly, this evaluation must take place before the model's output can affect anyone, typically during development.

Fairness specific techniques (see section 5.2.2) can help quantify potential ethical risks. If bias is detected, mitigation methods are available (see section 5.2.3). When no fairness specific intervention is needed, ethical risks can be managed through the standard risk management framework.

5.2.2 Bias quantification and fairness metrics

A model is considered biased when it behaves unfairly. Bias can arise from several sources [8]:

- **Biased data** – historical data may already reflect unequal or discriminatory patterns.
- **Biased populations** – real-world differences (e.g., crime rates differing by demographic group) can influence model outcomes.
- **Biased feature design** – monitoring features may unintentionally act as proxies for protected characteristics.
- **Biased model design** – modelling choices can introduce or reduce bias.

If not addressed, these issues can reinforce existing inequalities. This is especially relevant for supervised machine learning models trained on biased historical data, which can create feedback loops that enhance bias. Fairness assessments should therefore consider all potential sources of bias.

To manage ethical risks, data-science techniques – such as fairness metrics – can quantify how different groups are affected. Examples include:

- Alert volume parity, which checks whether alert rates across groups align with their presence in the population.
- False positive rate parity, which shows whether some groups face disproportionately high false-positive rates.

Fairness can be assessed at different levels, including individual models and the broader TM landscape. Such assessments may help banks identify areas where governance actions or model improvements could be most effective.

Because fairness metrics capture different perspectives, the choice of metric must be justified and agreed upon with key stakeholders (e.g., business, legal, privacy, model/risk owners). Multiple metrics may be relevant, and true fairness across all metrics may not be achievable [9]. Translating metric outcomes into real world implications – and using statistical tests when possible – helps to determine whether disparities are acceptable or require action.

Fairness assessments rely on availability of the right data. Many metrics require knowing group membership. Under the GDPR, processing special-category data is generally prohibited (Art. 9), except in narrow circumstances. The AMLR does allow processing such data when necessary for AML/CFT purposes (Art. 76), provided strict safeguards are met (e.g., informing clients, ensuring accuracy, preventing discriminatory decisions, and maintaining high security standards).

8 See e.g. Bias evaluation

Fairness insights are strongest when labels (e.g., whether an alert is investigation-worthy or leads to a SAR) are available. If labels are missing, comparisons to external statistics – such as national datasets – may help. When no suitable data exists but ethical risks have been identified, existing organisational risk-management frameworks may be used instead.

Finally, while fairness checks are often performed during development, some risks require ongoing monitoring (see section 7.2.1). Fairness metrics can be tracked in production like any other performance indicator.

Banks may choose to articulate how fairness is understood in their specific context and how relevant considerations are monitored and governed. While fairness has no single universal definition, establishing and maintaining a transparent, well-governed vision can help to ensure that fairness remains a deliberate and accountable part of their AI-driven FEC activities.

5.2.3 Handling unfairness

If data is available and significant unfairness is detected, data-science techniques can help address it. These may include:

- improving or correcting input data,
- applying bias mitigation methods during model development (see section 7.1.4), and
- adjusting model thresholds, potentially per impacted group.

Beyond clearly unlawful outcomes such as discrimination, ethical impacts often fall into a grey area where acceptability depends on the organisation's ethical stance or strategic choices. This highlights the potential value of a view on fairness and involving relevant stakeholders in discussions regarding ethical aspects and decision making.

5.3 Accountability and human oversight in TM

Accountability and human oversight are key principles in the EU AI Act. Awareness on these principles is increasing. These are often used too in the context of TM, even if the EU AI Act does not include AI systems to abide by the AML provisions in its list of High-Risk AI systems.

Accountability defines the responsibilities of developers, providers, and users to ensure regulatory and ethical compliance. Human oversight ensures that people remain actively involved in monitoring and controlling AI systems, striking the right balance between automation and human intervention. Too much intervention can reduce efficiency; too little can create risks.

Effective oversight requires operational controls that the system cannot override. Oversight roles must have the mandate, competence, and training to intervene when needed. The level of oversight depends on the AI application and its risk. For example, when AI is used to detect suspicious activity, analysts always review alerts.

When models automatically close obvious false positives, first and second line teams perform random checks to confirm continued model accuracy.

If, in the future, models would automatically submit SARs to the FIU, human oversight will become even more critical due to the potential impact on individuals. Strict quality checks would be essential.

The AMLR adds an even stricter requirement: any decision produced by automated processes (Art. 4), including those supported by AI – related to e.g. profiling, entering/maintaining/refusing a business relationship, blocking of transaction or adjusting due diligence measures pursuant to Article 20 AMLR – must be subject to meaningful human intervention to ensure accuracy and appropriateness. This requirement goes beyond general human oversight.

5.4 Responsible AI

Banks have proactively integrated responsible-AI principles into their TM systems, recognising the nature of their work and the need for strong self-regulation. Core principles – explainability, fairness, privacy, accountability, and human oversight – guide the development of safe and trustworthy AI in TM.

The EU's Ethics Guidelines for Trustworthy AI (2019) [10] and the DNB's General Principles for the Use of AI in the Financial Sector (2019) [11] accelerated these efforts, prompting banks to embed Responsible AI into their existing risk-management frameworks.

This proactive stance shows that the industry not only understands responsible AI expectations but also treats them as a strategic priority. By building AI systems that are reliable, transparent, and well-controlled, banks strengthen trust among stakeholders and ensure the ethical use of AI in transaction monitoring.

10 Ethics guidelines for trustworthy AI. Shaping Europe's digital future (europa.eu).
11 Good Practices Fiscale integriteitsrisico's bij cliënten van banken (dnb.nl).

6 Feedback loop

Establishing a robust feedback loop is essential for continuous refinement and optimisation of TM models. This loop should involve cross-functional collaboration between data scientists, TM analysts, second line, domain experts, and senior management. As input for the feedback loop, both internal and external sources should be considered.

The primary purpose of the feedback loop is to gather real-world operational insights into model performance. Each model-generated alert is presented to analysts together with an explanatory narrative that outlines why the behaviour was flagged. As the model's direct users, analysts can provide valuable feedback on the alert outcome but also on the clarity, usability, and relevance of the model's explanations. Quality Assurance processes – such as re-validating samples of alerts by experienced analysts – help ensure that case assessments meet required standards and that the feedback collected is reliable.

Analysts' outcomes can either confirm that the alert reflects genuine ML/TF related activity or show that it concerns non-risky behaviour. Both outcomes offer crucial information. These responses can be incorporated as labels into continuous monitoring, allowing institutions to track key performance metrics and ensuring the model remains within expected effectiveness ranges. Incorporating this feedback into ongoing training – such as through periodic retraining on recent alert data – supports continual model learning and adaptation.

Analysts can also provide operational feedback to improve the end-to-end process. This may include suggestions to enhance model explainability (e.g., clearer feature names), improvements to narratives, or user-interface enhancements that support faster and more effective analysis. Such operational insights help ensure that models remain aligned with analyst workflows and practical needs.

Sources outside the immediate model user group or even outside the organisation can provide useful feedback. This includes, for example, incident reporting, client complaints, supervisory reviews, and reports published by research or advisory institutions and public parties such as the FIU. The latter supports the timely identification and subsequent mitigation of emerging ML/TF risks in existing or new models.

Because ML/TF risk patterns evolve over time, the feedback loop plays a vital role in enabling banks to stay responsive to emerging threats. As new typologies or behaviours are detected, the insights gathered through this mechanism can prompt targeted updates to the model, ensuring it continues to perform effectively in a changing risk landscape.

7 Appendix: Technical considerations

TM models aim to detect transactions or clients with a heightened likelihood of exhibiting ML/TF risks. From a machine learning perspective, this task can be formulated as a binary classification problem in which each observation (e.g., a transaction or client) is assigned to either class 0 (legitimate activity) or class 1 (suspicious activity). As described in section 3.3, a variety of machine learning approaches can be employed depending on the model's purpose and the availability of labels.

In this appendix, we elaborate on several foundational concepts that are essential to sound model development, regardless of the specific modelling technique selected.

7.1 Model development

7.1.1 Train/validation/test set

Model fitting commonly requires partitioning the available data into three unequal subsets: a training set, a validation set, and a test set. The training set is typically the largest to ensure that models with many parameters can be estimated reliably. A fundamental requirement is that all three datasets are mutually exclusive and representative of the population the model will encounter in production. This prevents leakage of information across partitions and supports an unbiased assessment of model performance.

For example, in transaction-level models, multiple observations may belong to the same client. In such cases, all observations for a given client must be assigned to the same subset – training, validation, or test. This prevents the model from learning client-specific idiosyncrasies rather than generalisable behavioural patterns, particularly when client-level features are included.

The training set is used to estimate the model parameters – for instance, the splitting criteria in tree-based algorithms. The validation set serves two key functions: (i) it provides a controlled environment for selecting optimal hyperparameters (the configuration settings that govern how an algorithm learns), and (ii) it offers an initial indication of out-of-sample generalisation before evaluating performance on the test set. The test set – kept entirely unseen until the end – provides an unbiased estimate of model performance based on the final choice of parameters and hyperparameters. For a model to be considered acceptable, performance on the test set should not differ substantially from that on the validation set.

While this three-way split is standard practice, cross-validation is often viewed as a more robust alternative, particularly when data is limited. In cross-validation, the data (excluding the hold-out test set) is divided into several folds, and the model is trained and validated repeatedly on different combinations of these subsets. This

facilitates a more reliable assessment of model performance without sacrificing excessive amounts of data for a single dedicated validation set.

7.1.2 Performance metrics

To evaluate the performance of machine learning models, a key diagnostic tool is the confusion matrix, which provides a breakdown of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) by comparing the model's predictions to the actual labels.

		Prediction	
		0	1
Actual	0	TN	FP
	1	FN	TP

Figure 5 *Confusion matrix of alert outcomes.*

Performance metrics derived from this matrix offer insight into model effectiveness. Common measures include:

- Precision: TP/(TP+FP) – the proportion of correctly predicted positives among all positive predictions.
- Recall: TP/(TP+FN) – the proportion of actual positives correctly identified.
- Specificity: TN/(TN+FP) – the proportion of actual negatives correctly identified.
- Negative Predictive Value: TN/(TN+FN) – the proportion of correctly predicted negatives among all negative predictions.
- Accuracy: (TP+TN)/(TP+TN+FP+FN) – the share of all predictions that are correct.
- F1-score: 2 × (precision × recall)/(precision + recall) – the harmonic mean of precision and recall.

Exact values for these threshold-dependent metrics depend on the classification threshold chosen to label a transaction or client as risky. By contrast, some performance measures do not rely on a threshold, instead providing an aggregate assessment across all possible cut-off values:

- AUROC: Area under the Receiver Operating Characteristic curve, which measures how well a model can distinguish between classes across all classification thresholds, with higher values indicating better distinguishing ability.
- AUPR: Area under the Precision–Recall curve, which summarizes the trade-off between precision and recall across decision thresholds, making it especially informative for evaluating models on imbalanced datasets.

During model development, incorporating threshold independent metrics such as AUROC and AUPR is particularly advantageous as they provide a holistic view of the model's ability to distinguish across all thresholds

In practice, TM suffers from significant class imbalance: legitimate observations vastly outnumber risky ones. Such imbalance can cause models to over-predict the majority class, making it more difficult to identify rare suspicious behaviour. It is therefore crucial to select metrics suited to this context, balancing the need to minimise false positives with the imperative to detect true suspicious activity.

In settings characterised by strong class imbalance – such as TM – AUPR generally provides more informative insights than AUROC. Because AUROC considers the full distribution of negatives, it may yield deceptively high values when the minority class is rarely predicted. AUPR, by concentrating on the minority (risky) class, offers a more accurate reflection of the model's ability to distinguish suspicious activity while limiting unnecessary alerts.

Nevertheless, once the model has been trained and optimised using threshold-free metrics, it remains essential to report precision and recall at the selected operational threshold on the test set. These metrics link directly to business relevance: precision reflects the quality of alerts by showing the share of truly risky cases among all alerts, while recall captures the extent to which genuinely suspicious activities are identified and false negatives minimised.

7.1.3 Stability & robustness

In addition to standard performance metrics, uncertainty-aware methods – such as Bayesian modelling and conformal prediction – can be incorporated during model evaluation to quantify confidence in individual predictions. These approaches help identify cases where the model exhibits low certainty, which may signal areas with limited data coverage or potential blind spots in the detection logic. Although these techniques are not always practical for large-scale production environments due to computational or operational constraints, they can provide valuable diagnostic insights during model development and validation.

The performance of any model is inherently tied to both the fitted model and the specific dataset used for training and testing. As a result, model developers should explicitly acknowledge and assess the uncertainty introduced by the finite nature of the data. One practical approach is bootstrapping, where the model is applied repeatedly to different sub-samples of the dataset after the final parameters and hyperparameters have been selected. This produces a distribution of performance metrics rather than a single point estimate. Reporting both the mean and standard deviation of this distribution offers a more robust understanding of model uncertainty and can also support the definition of control limits for ongoing model monitoring.

The test set additionally serves as an important resource for assessing the temporal stability of model performance. By leveraging the chronological structure within the test data, developers can segment the test set into time-based intervals and evaluate the model's performance across these segments. Ideally, the model should demonstrate consistent behaviour over time, with only minor variations in key performance indicators. This temporal evaluation enables the early detection of issues such as model drift, changes in transaction behaviour, or degradation in predictive power. In doing so, it complements traditional performance assessment by ensuring that the model maintains reliable effectiveness under evolving real-world conditions.

7.1.4 Fairness

In section 5.2.3, we introduced several approaches for addressing model unfairness. Here, we highlight a set of techniques that can be applied during model development to mitigate unwanted bias and promote more equitable model outcomes.

- **Reweighing** adjusts the weights of training samples to ensure that different subgroups are represented more fairly in the learning process. By increasing the influence of under-represented or disadvantaged groups, the model is encouraged to learn patterns that are less biased.
- **Adversarial Debiasing** trains the model alongside an adversarial network that attempts to predict sensitive attributes (e.g., gender, age). The main model is penalised when the adversary succeeds, pushing it to make predictions that are less dependent on sensitive features.
- **Equalised odds Loss** This approach modifies the loss function so that the model is incentivised to achieve similar true-positive and false-positive rates across protected groups. It explicitly enforces fairness by aligning error rates between subpopulations.
- **Subgroup calibration** ensures that predicted risk scores carry the same interpretation across groups. For example, a score of 0.7 should correspond to the same empirical risk for each subgroup, resulting in consistent and interpretable outputs.
- **Reject option classification** In borderline situations – where the model exhibits low confidence – reject option classification explicitly favours decisions that reduce disparities between groups. The method adjusts predictions for uncertain cases to promote fairer treatment of protected groups.
- **Proxy removal strategies** These strategies aim to identify and remove variables (or derived features) that serve as proxies for sensitive attributes.

7.1.5 Threshold selection

Determining the probability threshold for a TM model requires balancing business objectives, regulatory obligations, and operational realities. These considerations are interdependent, and an effective threshold setting process must take all three into account.

Business objectives centre on the bank's strategic priorities and risk appetite. This involves evaluating the consequences of false positives and false negatives – both in terms of missed detections and unnecessary investigations – and understanding

their implications for operational efficiency, client experience, and reputational risk. Close collaboration with business stakeholders ensures alignment with detection priorities and helps quantify the financial and operational impact of different threshold settings.

Regulatory expectations require that the threshold selection process complies with applicable financial-sector standards and supervisory guidelines. Clear documentation of the reasoning behind the chosen threshold, including supporting analyses and justification, is essential for transparency during audits and regulatory reviews.

Operational efficiency considerations involve assessing how different threshold values influence alert volumes and the capacity of investigation teams. The objective is to identify a threshold that supports efficient case handling without compromising the model's ability to detect suspicious behaviour. In some contexts, adaptive or dynamic threshold mechanisms can help maintain efficient operations when transaction patterns or resource availability fluctuate.

A continuous feedback loop (see section 6) strengthens threshold governance by enabling regular reassessment of model performance under real-world conditions. Insights from operational teams, performance monitoring, and drift-detection processes support periodic fine-tuning to maintain alignment with evolving business needs, regulatory developments, and operational constraints.

By integrating these perspectives holistically, institutions can select a probability threshold that meets regulatory expectations, supports business objectives, and ensures that TM operations remain both effective and scalable.

7.2 In production

Model maintenance within an AI-driven TM landscape is an ongoing, iterative process that requires continuous oversight and proactive intervention to ensure sustained effectiveness and relevance. Rather than concluding at deployment, this phase encompasses a broad set of activities aimed at keeping the model aligned with evolving risks, regulatory expectations, and operational realities.

7.2.1 Model monitoring

To ensure that models continue to fulfil their intended business objectives, it is essential to conduct regular, structured assessments of both the model's performance and its underlying components.

7.2.1.1 Monitoring plan

A clearly defined monitoring plan should be developed that is formally embedded within the organisation's governance framework. A comprehensive monitoring plan should include:

- Clear ownership: identification of the teams and designated individuals accountable for monitoring and maintenance activities.

- Performance-assessment procedures: predefined dates and methodologies for evaluating model performance, including the metrics to be tracked (see section 7.2.1.2 and 7.2.1.3), whether manual labelling is required, and the sample size needed to conduct robust performance assessments.
- User feedback cycles: scheduled moments to collect input from model users regarding conceptual soundness, feature completeness, identification of counter-intuitive outcomes, and overall accessibility and usability.

In case of unexpected performance results, the responsible team/person should promptly engage with the model owner to initiate necessary actions.

7.2.1.2 Model drift

Model drift arises when an AI model's performance in production gradually degrades as underlying transaction patterns, client behaviour, or financial-crime typologies evolve. Because such changes may occur subtly over time, continuous monitoring of key performance indicators is essential to ensure the model remains effective against emerging threats. In addition to traditional metrics, unfairness should be treated as a core performance dimension and integrated into routine monitoring and reporting.

A robust monitoring framework evaluates a range of performance metrics – such as precision, recall, or tailored alert-quality indicators – selected in accordance with the model's risk-mitigation purpose. Precision on alerted transactions or clients is straightforward to measure, as it naturally integrates into existing workflows: analysts investigate alerts, enabling direct calculation of the proportion of true positives among all reviewed alerts (TP/(TP+FP)).

As discussed earlier, estimating recall in a production environment presents challenges, particularly because true labels for unalerted cases are typically unknown. Several approaches can be employed to approximate recall.

One method is random sampling, in which a subset of unalerted cases is manually investigated. By combining these sampled cases with alerted cases – and preserving their natural proportion – banks can construct an evaluation set that approximates the full population. Once investigation outcomes are obtained for this set, recall can be estimated.

However, for primary alert-generation models, simple random sampling is often impractical. Given the extreme imbalance in TM, very few sampled cases will represent investigation-worthy activity, necessitating prohibitively large sample sizes. Below-the-line sampling provides an alternative, by focusing on observations lying just below the alert threshold. These cases are more likely to contain overlooked suspicious behaviour due to their proximity to the decision boundary. After manual review of these borderline cases, banks can estimate recall by extrapolating model behaviour from alerted to unalerted cases. Although this approach introduces statistical assumptions and requires further methodological refinement, it provides a promising avenue for estimating recall in production settings where full-population labelling is infeasible.

7.2.1.3 Data drift

Data drift refers to gradual or abrupt changes in the statistical properties of incoming transaction data. Such shifts may arise from evolving client behaviour, seasonal or structural market dynamics, or newly emerging financial crime typologies. Detecting and responding to data drift is essential to ensure that a TM model continues to assess risk accurately and remains aligned with its design intent. A variety of techniques can be used to identify data drift. One of the most applied metrics is the Population Stability Index (PSI), which quantifies how much a variable's distribution has changed over time. PSI can be computed for model inputs (features) as well as model outputs, providing early indications of distributional instability. Higher PSI values suggest that the observed distribution has shifted materially and may signal underlying data drift.

As with model-performance monitoring, defining suitable threshold values for drift detection is a critical step. Thresholds must reflect the characteristics of the underlying data and the model's sensitivity to changes. Importantly, a breached threshold should never be interpreted as conclusive evidence of a problem. Instead, it should serve as an investigative trigger, prompting the model developer teams to conduct targeted, manual analysis. Such follow-up validation ensures that natural behavioural variation – such as periodic trends or legitimate shifts in transaction patterns – is distinguished from true data-quality issues or meaningful drift requiring action.

7.2.2 Model maintenance

Model maintenance typically involves three types of model-tuning interventions, each activated by a specific triggering event.

Periodic model refresh

A scheduled update in which the model is retrained on the most recent available data to ensure continued relevance as behavioural patterns evolve.

Model retraining

Similar to a periodic refresh but triggered by predefined conditions – such as observable degradation in performance or detected drift – rather than calendar intervals.

Model redevelopment

A more substantial effort undertaken when retraining cannot restore performance to acceptable levels, or when new business, regulatory, or technical requirements necessitate fundamental changes to the model's design.

A periodic refresh helps keep the model calibrated to the latest data, but full automation of this process is rarely practical. Models often rely on labels that are not immediately available, and automated retraining alone cannot guarantee that all ML/TF risks remain properly covered. Therefore, a strong model monitoring framework, combined with an active feedback loop, is essential for determining whether a simple refresh is sufficient or a more comprehensive redevelopment is required.

Retraining and redevelopment are essential to maintaining the model's long-term effectiveness, given the dynamic nature of criminal behaviour, data landscapes, regulatory expectations, and technical capabilities. Retraining may be required for several reasons.

- **Data drift** Gradual changes in transaction or client-level data distributions may reduce model effectiveness, necessitating adaptation to updated patterns.
- **Model drift** Even when data remains stable, emerging financial-crime tactics or behavioural shifts may cause the model's predictive ability to degrade. Retraining allows the model to learn from more recent examples and maintain performance.

While retraining is a prudent and necessary maintenance strategy, it may not always be sufficient. Model redevelopment becomes necessary when more fundamental changes are required:

- **Persistent performance degradation** Retraining fails to restore performance to target levels, indicating deeper underlying issues that require structural redesign.
- **Conceptual shifts** Significant change in criminal modus operandi or in the technological ecosystem demands adjustments beyond parameter or data-window updates.
- **Regulatory updates** Revised legislation, supervisory expectations, or internal governance standards require compliance-driven modifications.
- **Technology obsolescence** Legacy modelling techniques or infrastructure limitations prevent the model from scaling or delivering required performance, prompting redevelopment to leverage modern methods and architectures.

Redevelopment may involve adding or modifying features, incorporating new data sources, updating model architectures, or revisiting key methodological assumptions.

In summary, retraining supports incremental adaptation to gradual changes, whereas redevelopment is reserved for more structural shifts in risk, data, regulation, or technology. Both processes are crucial to ensure that the model remains effective, efficient, and aligned with the evolving financial-crime landscape.

Dutch Banking Association
Gustav Mahlerplein 29-35
1082 MS Amsterdam
www.nvb.nl

strong banks
strong society