Sonora: An Autonomous Analyst for Anti-Money Laundering Based on Explainable Artificial Intelligence

Vlademir Lenin Donato Batista[1] **, Paulo André Lima de Castro**[1]

[1]Instituto Tecnológico de Aeronáutica (ITA)
São José dos Campos – SP – Brasil

vlademirbatista@outlook.com, pauloac@ita.br

Abstract. *This paper presents the development and evaluation of Sonora, an autonomous analyst based on explainable artificial intelligence (XAI) designed to support Anti-Money Laundering (AML) monitoring in financial institutions. Using real-world, imbalanced data from a Brazilian financial institution, we developed a predictive model to identify suspicious cases. A multi-stage evaluation process began with the selection of Gradient Boosting as a robust baseline model over other standard classifiers. This model was then benchmarked against an alternative model, XGBoost, selected from a pool of state-of-the-art boosting algorithms. After hyperparameter optimization for both finalists, a threshold-based analysis confirmed that Gradient Boosting delivered the best performance for a recall-focused strategy, achieving a recall of approximately 96%. We deliberately focus on recall rather than accuracy to minimize regulatory risk and ensure that critical cases are surfaced for human review. To foster institutional trust, Sonora integrates instance-level SHAP explanations, highlighting the most influential features in each classification, and it acts as a decision-support tool, not a replacement for human analysts.*

1. Introduction

Money laundering involves concealing the origin of illicit financial resources through layers of complex transactions [GAFI 2012]. The increasing sophistication of these schemes, combined with increased transaction volumes and regulatory demands, undermines the effectiveness of traditional systems [Han et al. 2020]. Manual AML processes are time-consuming, error-prone, and heavily reliant on analyst interpretation [Jullum et al. 2020]. The regulatory framework in Brazil, for instance, was established by Law No. 9,613/1998 [Brasil 1998] and is complemented by directives from entities such as the Central Bank [Banco Central do Brasil 2020].

This paper addresses the need for improved, scalable, and interpretable solutions in Anti-Money Laundering (AML) monitoring, a challenge highlighted in recent reviews on the use of Artificial Intelligence (AI) in this field [Chen et al. 2018]. We propose Sonora, an autonomous analyst framework. In addition to improving performance and transparency, Sonora delivers a replicable pipeline for developing regulated AI in financial compliance settings. The system employs a Gradient Boosting model, developed through a methodical process encompassing data preprocessing, hyperparameter tuning, and careful consideration of data splitting strategies and decision thresholds. This methodology aims to enhancing the detection of suspicious financial activities (identified as the `Report` class), prioritizing high recall as a key objective in the AML context.

A central feature of Sonora is the integration of SHAP (SHapley Additive exPlanations) [Lundberg and Lee 2017, Lundberg et al. 2020], which provides instance-level, human-understandable explanations for the model's classifications. The combination of predictive modeling, with an emphasis on identifying high-risk cases, and transparent reasoning seeks to improve the operational efficiency of AML analysts, facilitate more informed decision-making, and support adherence to regulatory compliance.

The remainder of this article is organized as follows. First, we review related work that explores the application of artificial intelligence and explainability in anti-money laundering. Next, we present the proposed solution, detailing the system design and modeling approach. We then describe the experiments and discuss the results obtained. A comparative analysis with existing approaches is provided to contextualize the contributions. Finally, we conclude with key findings and suggest directions for future research.

1.1. Objectives

The main objective of this work is to develop and evaluate an explainable machine learning model for Anti-Money Laundering (AML) monitoring, embodied in a proposed system architecture named Sonora, and to demonstrate its potential for effective integration into financial institutions. Gradient Boosting [Friedman 2001] is selected as the primary algorithm for the predictive component after initial comparative analysis. The specific goals are to:

- Evaluate the predictive performance of various machine learning classifiers and select a primary model for detailed optimization and interpretation.
- Investigate and apply techniques, such as data splitting optimization and decision threshold adjustment, to achieve high recall for the class representing suspicious activities (`Report` class), a critical requirement in AML.
- Utilize interpretable methods, specifically SHAP, to generate clear, instance-level explanations for the model's predictions, thereby enhancing transparency and supporting analyst decision-making.
- Assess the optimized model's performance on independent test sets using standard evaluation metrics, focusing on its practical utility and the trade-offs involved in prioritizing high recall.

2. Related Work

The application of Artificial Intelligence (AI) to Anti-Money Laundering (AML) has gained increasing relevance as financial institutions seek to overcome the limitations of traditional, rule-based systems. Chen et al. (2018) and Han et al. (2020) highlight the growing complexity of financial transactions and the need for scalable and interpretable detection solutions. Explainable models—particularly those leveraging SHAP (Lundberg and Lee, 2017; Lundberg et al., 2020)—have become essential in building trust and aligning AI predictions with regulatory requirements. Recent research has explored various AI strategies across multiple stages of the AML pipeline. Balaji (2024) focuses on minimizing false positives, while Raj et al. (2024) investigate anomaly detection using neural networks. Konstantinidis and Gegov (2024) propose combining deep neural networks with SHAP to improve transparency.

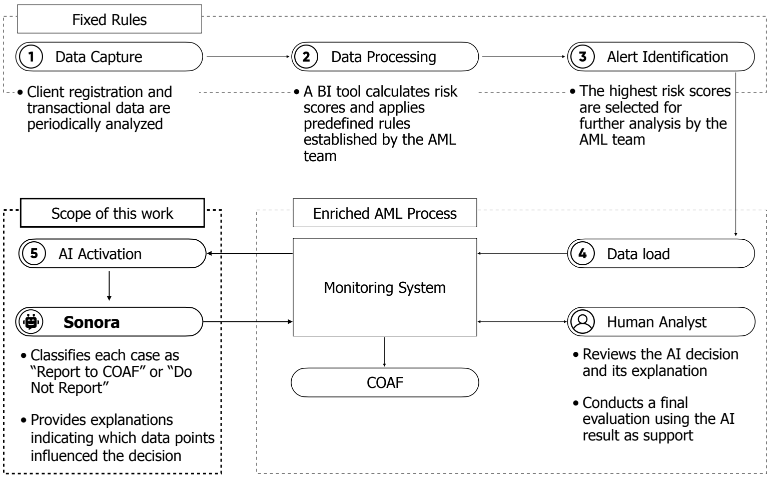
In addition, graph-based approaches have emerged as promising tools for modeling complex financial relationships. Assumpção et al. (2022) apply Graph Neural Networks (GNNs) to capture client connections and risk propagation. Bellei et al. (2024) leverage subgraph representation learning on blockchain data, while Weber et al. (2019) explore Graph Convolutional Networks for AML using Bitcoin transaction records. These studies expand AML research into decentralized contexts, reinforcing the need for adaptable and explainable AI frameworks.

3. Proposed Solution: Sonora

Sonora is an Autonomous Analyst powered by explainable artificial intelligence to support anti-money laundering (AML) monitoring. It assists compliance teams by classifying alerts from rule-based systems and providing interpretable justifications for each decision.

Its architecture centers on a Gradient Boosting classifier trained on labeled alerts from financial operations, producing a `Report` or `Dismiss` output. To ensure auditability and analyst trust, Sonora integrates SHAP (SHapley Additive exPlanations) [Lundberg and Lee 2017], quantifying the contribution of each feature to the prediction.

The workflow (Figure 1) begins with structured alerts filtered by institutional rules, followed by a preprocessing stage that applies the necessary data treatments to prepare the dataset for model inference, as detailed later in this work. Next, the Gradient Boosting classifier generates the prediction, the SHAP-based layer produces the explanation, and both are delivered for analyst review and regulatory documentation.



Fixed Rules
1 Data Capture
• Client registration and transactional data are periodically analyzed
2 Data Processing
• A BI tool calculates risk scores and applies predefined rules established by the AML team
3 Alert Identification
• The highest risk scores are selected for further analysis by the AML team
Scope of this work
5 AI Activation
Sonora
• Classifies each case as "Report to COAF" or "Do Not Report"
• Provides explanations indicating which data points influenced the decision
Enriched AML Process
Monitoring System
COAF
4 Data load
Human Analyst
• Reviews the AI decision and its explanation
• Conducts a final evaluation using the AI result as support

Figure 1. Architectural workflow of the Sonora autonomous analyst, illustrating the modular stages from structured alert intake to SHAP-based explanation delivery, designed for integration within compliance operations.

4. Experiments and Results

This section presents the experiments conducted to evaluate Sonora. It details the dataset, preprocessing steps, model training, and evaluation.

4.1. Dataset and Preprocessing

The dataset comprises 10,856 structured alerts collected from a Brazilian financial institution operating as both a bank and an insurance provider. Each alert corresponds to a transaction reviewed by human analysts and labeled as either `Report` (suspicious) or `Dismiss` (not suspicious). Reflecting the natural class imbalance observed in operational AML monitoring, only 27.9% of the cases were labeled as `Report` (3,028 alerts), while 72.1% were labeled as `Dismiss` (7,828 alerts) — highlighting the predominance of non-suspicious alerts in real-world scenarios.

Table 1. Data dictionary of the original dataset prior to preprocessing. This includes all raw features before the removal of non-informative fields (e.g., free-text notes) and the application of data transformations.

Feature	**Description**
Income Reported	Client's declared monthly income
Transaction Amount	Monetary value of the alerted transaction(s)
Alerted Product	Alerted financial product type
Operational Score	Total Score
KYC/Reputation Score	Reputation score derived from KYC profile
Participant Score	Score based on participant-related attributes
Product Score	Score based on characteristics of the alerted product
Partner/Associate Score	Score based on linked partner or associate entities
Watchlist Name	Indicator of match in external watchlists
PEP Flag	Politically Exposed Person flag
Main Participant Flag	Indicates if the client is the main case participant
Open Legal Case	Flag for an open legal or administrative process
Fraud detection	Indicator of whether a dossier was created
State (UF)	Client's federative unit of residence
Relationship Start Date	Date of relationship start with the institution
Analyst Review Notes	Free-text notes from analyst review
Analyst Decision	Final analyst decision: REPORT or DISMISS

To ensure a robust and unbiased evaluation, the dataset was partitioned using a stratified hold-out strategy. First, 30% of the full dataset was segregated as a final, untouched test set, reserved exclusively for assessing the performance of the optimized model. To preserve the data's underlying structure in this initial split, a multi-column stratification strategy was employed. Unlike standard approaches that stratify only on the target variable (`Report`/`Dismiss`), our method also stratified on quintiles of key financial indicators: declared income and transaction amount. This advanced stratification ensures that the training and test sets are representative not only of the class distribution but also of the financial profiles of the alerts, a critical factor in AML contexts.

These financial attributes were selected due to their central role in money laundering schemes, as noted in regulatory guidelines. This relevance was later reinforced by SHAP results, which identified income and transaction amount as among the most influential features in the model's decision-making process. The remaining 70% subset was used for training and validation. Preprocessing followed a modular pipeline architecture to ensure reproducibility and guard against data leakage:

- **Imputation:** To handle missing values while preserving data integrity—a crucial step in financial fraud detection workflows [Moepya et al. 2016]—selected numerical attributes were manually imputed with the median *prior* to the preprocessing pipeline.
- **Normalization:** To ensure that monetary and score-based attributes (e.g., income, transaction amount, participant score) contribute proportionally to distance-based learning algorithms and gradient-based optimizers, numerical features were standardized using `StandardScaler`, which centers the data around zero with unit variance.
- **Categorical Encoding:** Categorical attributes such as product type, state (UF), and watchlist flags were encoded using `OneHotEncoder`. This transformation avoids imposing ordinal relationships where none exist, ensuring that each category is treated independently by the model.
- **Feature Engineering:** Domain-specific attributes were created to enhance the model's expressiveness. These include relationship duration (in years), a binary flag indicating whether the client appears on a watchlist, and disaggregated indicators of alerted product types. Such transformations allow the model to capture relevant patterns that would otherwise remain hidden in raw categorical fields [Domingos 2012].
- **Pipeline Construction:** To ensure reproducibility and prevent discrepancies between training and inference stages, all preprocessing steps—imputation, normalization, and categorical encoding—were encapsulated using `scikit-learn`.

Finally, to assess potential multicollinearity between input features, we computed the Pearson correlation matrix over the numerical attributes from the training set. All variable pairs exhibited low to moderate correlations, with no $|\rho|$ exceeding 0.8.

To select a baseline model, a comparative evaluation was conducted involving three distinct types of classifiers, each chosen to address different aspects of the Anti-Money Laundering (AML) problem. We selected `Logistic Regression` as a classic linear model to establish a strong, interpretable benchmark, a crucial aspect in regulated financial contexts. To capture the complex, non-linear interactions often present in financial fraud data, we also included two ensemble models: `Random Forest` and `Gradient Boosting`. The latter is a tree-based model consistently reported as state-of-the-art for structured data, often outperforming even deep learning models on typical tabular datasets [Grinsztajn et al. 2022]. This combination allowed us to measure the performance gains offered by complex models against a simple, transparent baseline.

The data in Table 2 shows that all three models performed significantly better than a random classifier, with Gradient Boosting achieving the highest Area Under the Curve (AUC) of 0.733, indicating slightly superior discriminatory power. The quantitative results in Table 2 detail this performance. Although all models showed similar metrics, Gradient Boosting achieved the highest accuracy and recall (both 0.74). The F1-score, which balances precision and recall, was similar for GBoost and Random Forest (0.69).

Based on these results, Gradient Boosting was selected as the baseline model for the subsequent stages of the study. Its balanced performance, particularly leading in the recall metric at this initial phase, made it the most suitable candidate aligned with the project's objective. To ensure methodological rigor, we separated the test set (30% of the original data) before any modeling decision, preserving an untouched subset for final evaluation. All preprocessing, feature engineering, model selection, and threshold tuning were conducted exclusively on the training partition. This approach mitigates the risk of information leakage and optimistic bias. We also avoided synthetic oversampling and applied stratified validation to account for class imbalance. Cross-validation was used during hyperparameter tuning, and feature importance was later analyzed to avoid reliance on spurious correlations. Although the labels were assigned by experienced human analysts, potential labeling biases remain a limitation, as further discussed in the related section.

Using GBoost as the baseline, we compared three state-of-the-art boosting algorithms—`XGBoost`, `LightGBM`, and `CatBoost`—using the same stratified training and test sets, a fixed threshold of 0.5, and metrics including accuracy, precision, recall, and F1-score. As shown in Table 3, `XGBoost` achieved the highest Recall and F1-score, making it the selected candidate model for the final comparison against `GBoost`.

Table 2. Performance comparison of candidate classifiers on the validation set. Gradient Boosting (GBoost) slightly outperformed Random Forest and Logistic Regression across most metrics, supporting its selection as the final model for deployment in the Sonora analyst.

Model	Accuracy	Precision	Recall	F1-score	AUC
GBoost	0.74	0.71	0.74	0.69	0.733
Random Forest	0.72	0.69	0.72	0.69	0.714
Logistic Reg.	0.73	0.69	0.73	0.66	0.716

Table 3. Comparison of advanced boosting classifiers to select a challenger model for the final evaluation. `XGBoost` was chosen for its superior Recall and F1-score.

Model	Accuracy	Precision	Recall	F1-score
XGBoost	75.54	74.94	75.54	75.18
LightGBM	75.22	74.41	75.22	74.77
CatBoost	75.42	74.84	75.42	75.11

To ensure the final comparison was between the best possible versions of our two finalist models, both the baseline (`GBoost`) and the challenger (`XGBoost`) were optimized. We employed `Optuna`, a **hyperparameter optimization** framework based on Bayesian sampling. This process efficiently tuned key parameters for each model—such as learning rate, maximum tree depth, and number of estimators—ensuring both were performing at their peak before the final threshold analysis. All optimization runs were seeded to ensure reproducibility [Akiba et al. 2019].

4.2. Model Comparison and Threshold Analysis

Having selected `GBoost` as the baseline model and `XGBoost` as the strongest candidate—and after tuning both using the `Optuna` framework as described previously—we conducted the final comparative analysis. This evaluation assessed performance across multiple decision thresholds to determine which model better aligns with the high-recall requirement of AML workflows.

Results show that `GBoost` consistently achieves higher recall than `XGBoost` at all thresholds tested, supporting the project's priority of minimizing the risk of undetected suspicious cases. Considering the regulatory importance of reducing false negatives in AML, and in alignment with this work's primary objective, the `GBoost` model with a 0.10 threshold was selected for deployment.

Table 4. Threshold-based comparison between GBoost and XGBoost models on the test set (values in %).

Threshold	**GBoost**		**XGBoost**	
	Recall	Precision	Recall	Precision
0.10	96.56	35.68	85.90	39.40
0.20	81.31	42.36	71.48	44.13
0.30	65.90	45.68	58.85	47.87
0.40	46.23	53.01	47.21	50.44
0.50	30.49	59.24	37.05	53.05

4.3. Explainability

Explainability is central to Sonora's design, supporting analyst trust, operational efficiency, and regulatory auditability. The system integrates SHAP to provide two perspectives: a global view of the model's overall behavior and local, instance-level justifications for each prediction. As shown in Figure 2, declared income, transaction amount, and specific product categories are among the most influential features, validating the stratification choices made during dataset preparation.

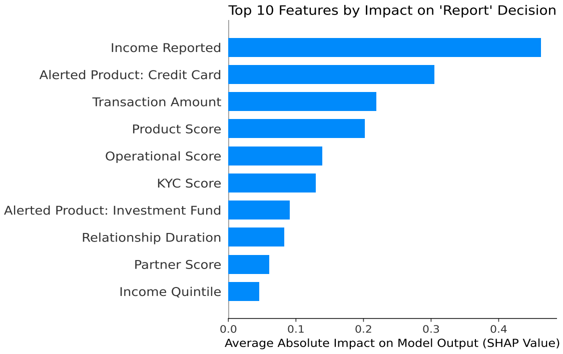


Figure 2. Top 10 features with the highest impact, according to SHAP values, in the final Gradient Boosting model.

While the global view is essential for model validation, analysts also require clear justifications for individual alerts. To address this, Sonora generates concise summaries that include the predicted class, associated probabilities, and the top three contributing factors. For example:

> Model predicted ‘Report’. Probabilities: Report (85.3%), Dismiss (14.7%). Key factors: high transaction amount (+45.2%), low reported income (+28.9%), and alerted product “Investment Fund” (+15.1%).

Restricting explanations to the top three factors ensures clarity and operational usability, while the complete set of SHAP values remains available internally for audits or more in-depth investigations.

5. Comparative Analysis

We selected a diverse set of recent studies that apply artificial intelligence to different stages of the Anti-Money Laundering (AML) investigative workflow. Instead of focusing on a single task—such as transaction scoring or post-alert triage—the selected work reflects varied strategies that span from suspicious activity detection to analyst decision support. Our comparative criteria also emphasized methodological diversity. The studies include graph-based models, gradient boosting, and deep neural networks, paired with a range of explainability techniques such as SHAP, embeddings, and symbolic rules. A common thread among them is the acknowledgment of regulatory rule-based logic, whether as a basis for filtering alerts or as an interpretability mechanism.

Although most of the selected works are grounded in real-world institutional data, we deliberately included two exceptions. Assumpção et al. [Assumpção et al. 2023] was considered for its innovative use of Graph Neural Networks, even though it does not incorporate predefined compliance rules. Konstantinidis and Gegov (2024), while using synthetic data, contribute a structured approach that combines deep learning and SHAP explanations—highlighting the trade-offs between experimental design and operational realism.

Table 5 summarizes the qualitative characteristics of the selected studies, including data type (whether the models were trained on real, synthetic, or mixed datasets), rule integration (whether regulatory rules were incorporated into the workflow), model architecture, and the type of explainability technique employed. Table 6 compiles the reported recall and accuracy values for each model. We focus on these two metrics to highlight the trade-offs between sensitivity to suspicious activity (recall) and overall classification performance (accuracy). As shown, the selected approaches differ not only in algorithmic design but also in assumptions about data availability and institutional context. For example, works that rely on graph structures [Assumpção et al. 2023] or synthetic datasets [Konstantinidis and Gegov 2024] may face integration barriers in regulated environments. Others, like Thanathamathee et al. (2024) and Tertychnyi et al. (2022), focus on explainability through SHAP while preserving compatibility with real data pipelines.

Sonora was designed for direct deployment within operational AML workflows. It consumes structured alerts already filtered by institutional rules, relies on widely available AML features (e.g., declared income, transaction value, product type), and integrates instance-level SHAP explanations to assist analysts and auditors. This makes it particularly suited for adoption in compliance-driven settings, balancing model performance, transparency, and regulatory alignment.

Table 5. Summary of qualitative attributes of selected AML-related studies, including the nature of training data, incorporation of regulatory rules, model architecture, and type of explainability technique applied.

Study	Data	Rules	Model	XAI
Sonora	Real	Yes	G-Boost	SHAP
Assumpção	Real	No	GNN	Embeddings
Tertychnyi	Real	Yes	CatBoost	SHAP
Thanathamathee	Real	Yes	XGBoost	SHAP
Chatzimparmpas	Real	Yes	Hybrid	Heuristics
Konstantinidis	Synthetic	Yes	DNN	SHAP

6. Contributions and Applicability

This work presents a practical, explainable, and high-recall solution for Anti-Money Laundering (AML) monitoring based on real-world institutional data. The contributions are primarily methodological and applied, aligning with operational demands in regulated financial environments.

Table 6. Reported recall and accuracy values from recent AML research works, used to compare overall classification performance and sensitivity to suspicious activity, contextualizing Sonora's results.

Study	Recall	Accuracy
Sonora	0.96	0.50
Assumpção	0.94	0.89
Tertychnyi	0.91	0.87
Thanathamathee	0.78	0.77
Chatzimparmpas	0.82	0.79
Konstantinidis	0.99	0.91

6.1. Main Contributions

- **Real-world deployment context:** The proposed model was trained on operational alerts from a real financial institution, labeled by experienced human analysts and filtered by existing compliance rules. This guarantees ecological validity and direct relevance to actual AML workflows.
- **End-to-end pipeline with explainability:** Sonora integrates a Gradient Boosting classifier, a modular and reproducible preprocessing pipeline, and SHAP-based, instance-level explanations. This creates a deployable architecture that balances predictive performance, regulatory transparency, and operational interpretability.
- **Recall-oriented optimization:** The modeling approach included extensive experiments with train/test splits and threshold calibration, explicitly prioritizing recall for the "Report" class—a critical factor for regulatory compliance and institutional risk mitigation in AML.
- **Human-centric interpretability:** SHAP explanations are delivered in a format accessible to compliance analysts and suitable for audit documentation, fostering transparency and supporting human-in-the-loop decision-making.

- **Comparative validation:** The solution underwent comprehensive comparative analysis against recent AML research, demonstrating strong performance in terms of recall, explainability, and operational readiness in regulated environments.

6.2. Replicability and Adaptation

Although Sonora was trained on proprietary data from a Brazilian financial institution, its architecture was designed with portability in mind. The model operates on a set of features that are widely available in AML systems, allowing for straightforward replication in other organizational contexts.

- **Feature structure alignment:** The dataset should include structured alerts with key AML attributes, such as income, transaction value, relationship duration, product type, and internal risk indicators (e.g., flags, watchlists, scores).
- **Reusability of the preprocessing pipeline:** The preprocessing pipeline leverages established libraries such as scikit-learn, enabling adaptation to datasets with similar structure using standard normalization, encoding, and imputation techniques.
- **Model retraining and threshold calibration:** While the overall architecture and SHAP-based explanation layer can be reused, retraining the classifier on local data is recommended to account for institutional patterns and regulatory contexts. Threshold optimization procedures can be tuned according to the risk appetite and compliance priorities of each institution.
- **Explainability portability:** SHAP explanations remain interpretable across institutions as long as feature semantics—that is, the correspondence of feature meaning—is preserved, enabling consistent analyst support and compliance reporting.

6.3. External Dataset Validation

To assess robustness and generalization, Sonora was applied to the public UCI Default of Credit Card Clients dataset [UCI Machine Learning Repository 2009], adapted to match the input format expected by Sonora through engineered risk scores and compliance indicators. This adaptation process replicated the institution's fixed rules for operational and product risk—such as frequent high-value payments, consistent credit limit overuse, large one-time bills, cash advance patterns, and KYC-related risk thresholds—ensuring alignment with the original compliance logic.

Although performance decreased due to domain shift, the model maintained high recall (93.4%) for the `Report` class, supporting its design as a high-sensitivity solution. These results highlight both Sonora's adaptability and the need for external benchmarks more closely aligned to operational AML data.

6.4. Regulatory and Compliance Alignment

Brazilian AML regulations require analyst accountability, auditability, and decision justification. Sonora addresses these requirements by generating SHAP-based, instance-level explanations that can be directly integrated into audit trails and compliance reports, strengthening governance and oversight while complementing existing rule-based workflows. Internationally, the FATF Recommendations and the EU Artificial Intelligence Act [Parliament and Council 2024] emphasize that high-risk AI systems must be explainable, fair, and traceable. Sonora meets these principles through a human-in-the-loop architecture, interpretable risk scoring, and explicit avoidance of black-box models in compliance-critical contexts.

7. Limitations

This study used proprietary data from a Brazilian financial institution, aligned with its specific risk patterns, product structures, and compliance workflows. Due to confidentiality and regulatory restrictions, the dataset cannot be publicly released, limiting reproducibility. Furthermore, the model was trained on human-assigned labels (`Report`/`Dismiss`) that reflect perceived regulatory risk rather than confirmed illicit activity, introducing potential bias from analyst subjectivity or biases and institutional policies.

8. Conclusion

This study introduced Sonora, an autonomous analyst designed to enhance Anti-Money Laundering (AML) monitoring through explainable machine learning. Built using real-world institutional data and integrated with SHAP explanations, Sonora prioritizes high recall to reduce the likelihood of missing suspicious activities, while maintaining interpretability and compatibility with existing rule-based workflows. The system achieved strong performance (recall of 0.96 for the `Report` class) without requiring synthetic resampling or deep modifications to the original data. This result highlights the effectiveness of threshold tuning and structured data preparation in addressing the strict regulatory demands of high-risk financial environments. Furthermore, by leveraging commonly available AML features and offering per-instance explanations, Sonora supports analysts in both operational triage and audit documentation. Despite these strengths, the study presents limitations. Its generalizability across different financial institutions remains untested, and it does not currently incorporate external behavioral signals or anomaly detection strategies.

Future research should validate Sonora across institutions, evaluate additional explainability techniques such as LIME [Silva et al. 2023], explore hybrid models combining supervised and unsupervised learning, and assess long-term impact on analyst workflows and regulatory outcomes. In addition, experiments using publicly available or synthetic datasets—such as PaySim— should be conducted to assess the generalizability of the approach and support independent replication by the research community. Future work should also investigate threshold calibration strategies tailored to different institutional risk tolerance profiles and regulatory interpretations, as the optimal trade-off between recall and precision may vary significantly across compliance environments.

References

Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2623–2631. ACM.

Assumpção, C., Batista, J., and Finger, M. (2023). Delator: Dynamic embedding learning for anti-money laundering using transaction networks. *Expert Systems with Applications*, 213:119041.

Banco Central do Brasil (2020). Circular nº 3.978, de 23 de janeiro de 2020. Dispõe sobre controles internos a serem adotados pelas instituições financeiras. Brasília, DF. Accessed: June 1, 2025.

Brasil (1998). Lei nº 9.613, de 3 de março de 1998. Dispõe sobre os crimes de lavagem de dinheiro e bens. Planalto – Presidência da República. Accessed: June 1, 2025.

Chen, Z., Khoa, L. D. V., Teoh, E. N., Nazir, A., Karuppiah, E. K., and Lam, K. S. (2018). Machine learning techniques for aml solutions in suspicious transaction detection: A review. *Knowledge and Information Systems*, 57(2):313–339.

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, 55(10):78–87.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232.

GAFI (2012). Padrões internacionais de combate à lavagem de dinheiro e ao financiamento do terrorismo e da proliferação: As recomendações do GAFI. Technical report, GAFI, Paris, France. Official translation by COAF. Accessed: June 1, 2025.

Grinsztajn, L., Oyallon, E., and Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data?

Han, J., Cai, Y., and Xue, Y. (2020). Artificial intelligence for anti-money laundering: A review and extension. *Digital Finance*, 2(4):211–239.

Jullum, M., Øystein Huseby, Løland, A., Espe, N. V., and Bjørkevoll, V. H. (2020). Detecting money laundering transactions with machine learning. *Journal of Money Laundering Control*, 23(1):173–186.

Konstantinidis, S. and Gegov, D. (2024). Interpretable deep learning for AML: A case study with SHAP and dnns. *Journal of Financial Crime Analytics*, 4(1):35–50.

Lundberg, S. M., Erion, G. G., and Lee, S.-I. (2020). Consistent individualized feature attribution for tree ensembles. *Nature Machine Intelligence*, 2(1):56–67.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In ., editor, *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, pages 419–428, Red Hook, NY, USA. Curran Associates, Inc. Accessed: June 1, 2025.

Moepya, S. O., Akhoury, S. S., and Nelwamondo, F. V. (2016). Measuring the impact of imputation in financial fraud. In *2016 Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference*, pages 1–6. IEEE.

Parliament, E. and Council (2024). Regulation (eu) 2024/1689 of 13 june 2024 laying down harmonised rules on artificial intelligence (ai act). Official Journal of the European Union. Accessed: June 1, 2025.

Silva, R. M., Sbrana, A., Castro, P. A. L., and Soma, N. Y. (2023). Developing and assessing a human-understandable metric for evaluating local interpretable model-agnostic explanations. *International Journal of Intelligent Engineering and Systems. DOI: 10.22266/ijies2023.0831.26*, 16:318–332.

UCI Machine Learning Repository (2009). Default of credit card clients dataset. `https://www.kaggle.com/datasets/uciml/default-of-credit-card-clients-dataset`. Dataset accessed in July 2025.