2023 IEEE International Conference on e-Business Engineering (ICEBE) | 979-8-3503-2555-3/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICEBE59045.2023.00028

Enhancing Anti-Money Laundering: Development of a Synthetic Transaction Monitoring Dataset

Berkan Oztas
Computing and Informatics
Bournemouth University
Bournemouth, United Kingdom
boztsa@bournemouth.ac.uk

Deniz Cetinkaya
Computing and Informatics
Bournemouth University
Bournemouth, United Kingdom
dcetinkaya@bournemouth.ac.uk

Festus Adedoyin
Computing and Informatics
Bournemouth University
Bournemouth, United Kingdom
fadedoyin@bournemouth.ac.uk

Marcin Budka
Computing and Informatics
Bournemouth University
Bournemouth, United Kingdom
mbudka@bournemouth.ac.uk

Huseyin Dogan
Computing and Informatics
Bournemouth University
Bournemouth, United Kingdom
hdogan@bournemouth.ac.uk

Gokhan Aksu
AML Transaction Monitoring
Danske Bank Group
London, United Kingdom
gokhan.aksu@uk.danskebank.com

***Abstract*—Money laundering remains a continuous global problem, necessitating the development of new enhanced transaction monitoring methods. Current anti-money laundering (AML) procedures within the industry are inefficient, and access to transaction monitoring data is limited due to legal and privacy constraints, with available data lacking true labels and diversity. This study presents a new AML transaction generator and uses it to create a dataset called SAML-D. The SAML-D dataset contains 12 features and 28 typologies, expanding beyond the existing datasets by incorporating a wider range of typologies, geographic locations, high-risk countries, and high-risk payment types. The typologies are created based on existing datasets, the literature, and semi-structured interviews with AML specialists. Additionally, machine learning experiments are conducted to present the applicability of the dataset within the field of AML and results are compared to an existing dataset. The primary purpose of the generator and dataset is to provide researchers with an additional resource to evaluate their models and facilitate comparative analysis of their results, potentially assisting the development of more advanced and capable transaction monitoring methods.**

***Index Terms*—Anti-Money Laundering (AML), Transaction Monitoring, Synthetic Dataset, Machine Learning, Artificial Intelligence**

I. Introduction

Money laundering is a continuous global problem and refers to the process of converting criminally attained money to appear as arising from legitimate sources [1]. Money launders often require the use of banks to launder their funds, leaving complex transactional and behavioral patterns. The anti-money laundering (AML) unit in banks is required to conduct transaction monitoring to detect and report suspicious activities to the authorities [2]. Currently, most financial institutions utilize rules-based techniques to monitor transactions. Rules-based approaches lead to high false positive rates of 95% [3], creating redundant investigations and increasing operational costs for banks. Given the challenges associated with rules-based approaches, there has been an increasing amount of attention in developing machine learning approaches to reduce false positives and achieve more effective results [4].

Although, machine learning approaches are showing promising results, accessing data is challenging. Real financial transaction data consisting of money laundering behaviors is not generally available, due to legal and privacy reasons [5] [6]. In cases where anonymized real data is available, the labeling of money laundering transactions is limited due to a lack of ground truths, and laundering transactions often going undetected [7]. The availability of synthetic money laundering data is limited with these datasets often having significant shortcomings, such as the absence of critical features or a lack of diverse money laundering typologies.

This paper introduces a novel synthetic AML transaction generator. The dataset is in tabular format and includes various features of 'suspicious' and 'normal' transactions conducted by different entities. The key contributions of this study involve an extension of both normal and suspicious typologies in the dataset, improving upon existing synthetic alternatives. Including new features such as geographic locations and high-risk countries adds a layer of complexity and brings a greater degree of realism to the dataset. A current challenge in the domain is comparing the results of different studies and machine learning algorithms as experiments are conducted using different datasets [8]. Our dataset can address these challenges by serving as a benchmark dataset for researchers, enabling comparison and consequently supporting more meaningful analysis. Additionally, we conducted a set of preliminary experiments using simple machine learning algorithms, demonstrating the utility of our data and establishing a comparison point.

II. Related Work

AMLSim project introduces a two-component method for generating synthetic banking transaction data using a multi-agent-based simulator [9]. The 'Transaction Graph Genera-

979-8-3503-2555-3/23/$31.00 ©2023 IEEE
DOI 10.1109/ICEBE59045.2023.00028

tor' creates accounts and attributes based on an input CSV account file and establishes basic interactions between them. Suspicious transactions are then added using another CSV alert parameter file. The 'Transaction Simulator' uses a multi-agent program to simulate transactions within the network. The simulator mimics real-world transactions, aiming for an accurate portrayal of real-life financial behaviors. Moreover, the number of transactions can be modified during generation. Several datasets with varying amounts of transactions are provided. In this paper, we focus on the analysis of the '100Kvertices-10Medges' dataset. To our knowledge, AMLSim is the only synthetic dataset utilized to asses new transaction monitoring approaches in the current literature [10] [11]. However, there are limitations to AMLSim such as its singular transaction type.

The IT-AML dataset [12] is a synthetic representation of financial transactions created via a multi-agent virtual world model. The dataset is not derived from anonymized real-world individuals, rather it represents an entirely synthetic construct where various entities interact in ways that mimic real-world financial transactions. The dataset includes a collection of good and bad actors, with the latter attempting to launder money following one of the eight incorporated typologies: Fan-Out, Fan-In, Cycle, Bipartite, Stack, Random, Scatter Gather, and Gather Scatter. The typologies from AMLSim are used. The IT-AML dataset expands on the AMLSim dataset by incorporating more transaction types and typologies, making it a valuable AML dataset. A selection of datasets is provided, with varying numbers of transactions and illicit activities included. In this paper we utilized the HI-Small_Trans.csv set.

Money Laundering Data Production (MLDP) is a dataset created during a Master's project that simulates financial transactions [13], incorporating suspicious transactions based on traditional stages of money laundering. Despite its attention to the key money laundering stages, MLDP includes a limited number of transactions and lacks detail about the different types of money laundering methods, providing room for improvement in future synthetic datasets.

Other transactional datasets have been created for the purpose of fraud detection. Although these datasets were not designed for AML purposes, they offer valuable insights such as the size and features of the datasets, due to the domain's similarities. PaySim [14] takes a different approach by utilizing a simulator to generate synthetic financial datasets based on a month-long sample of real mobile money transactions. The method focuses on replicating the statistical patterns observed in the original dataset. The resulting dataset is vast, containing millions of transactions and unique sender/receiver IDs. The Credit Card Fraud Detection dataset [15] is a real dataset from 2013 and focuses on credit card transactions made by European cardholders. The data consists of a low percentage of fraudulent transactions and provides a large number of features for each transaction. However, 28 out of the 31 features are PCA-transformed because of confidentiality issues (V1–V28), while the others are Time, Amount, and Class.

III. Methodology

The framework employed in creating the synthetic anti-money laundering dataset was constructed to ensure well-represented money laundering typologies, interactions between agents (bank accounts), and realistic banking transactions. This section presents the adopted approaches and highlights how they interact to create the dataset. The production of the "normal" and "suspicious" transactions involves two methods as shown in Fig 1; the agent-based approach [16] and the typology-based approach [17]. When creating normal transactions, firstly normal transaction typologies are identified and created. Then a set of regular bank accounts are generated. For each regular bank account, typologies are randomly assigned to generate the different types of normal transactions. This allows for each regular bank account to acquire its own unique combination of normal typologies to have more complex and diverse transaction behaviors. Generating fraudulent accounts that carry out "normal" and "suspicious" transactions consist of two steps. The first step involves creating money laundering typologies. Then "suspicious" transactions with many sender and receiver bank accounts are generated by replicating the various money laundering typologies. Each typology is considered to be its own category and therefore created separately. This allows for each typology to be significantly represented in the generated dataset [18]. The next step starts by identifying all the unique sender and receiver accounts that conduct suspicious transactions. Then for each unique account, multiple normal typologies are randomly assigned to generate the different types of normal transactions. This allows for suspicious accounts to have normal transactions with normal accounts who have not been involved in suspicious activities previously. Once these two datasets are generated, they are combined to create the synthetic anti-money laundering dataset which includes "normal" and "suspicious" transactions.

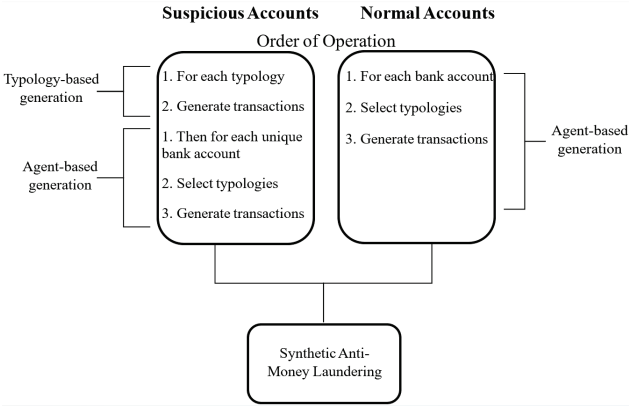

Fig. 1. Generation process of the 'Normal' and 'Suspicious' accounts and transactions

Generating the various suspicious and normal typologies was a complex procedure, involving multiple phases. The first phase consisted of an in-depth evaluation of the academic

literature [19] [20] [18], AML domain [21] [22], and existing datasets [9] to identify common and emerging typologies that could be included in our dataset. Further evaluation was done through semi-structured interviews with 8 AML experts. The interviews lasted approximately 60 minutes and were conducted online. Once all the interviews were complete the data was transcribed and anonymized for data analysis to then take place. Semi-structured interviews were selected as they allow for an in-depth understanding of the participant's experiences, opinions, and knowledge related to the AML subject [23]. This helped derive a better understanding of several money laundering typologies such as, difficult to detect typologies using rules-based methods, typologies leading to high false positives using rules-based methods, and high-risk payment types. Based on these evaluations, many typologies were chosen to be included in the dataset. Then using Python each typology was created. A combination of both simple and complex typologies was generated to incorporate the various levels of strategies used in the real world. The suspicious and normal typologies were all ran and tested individually to ensure a desired output was being generated.

To represent essential characteristics of banking transactions and include relevant attributes for money laundering transactions, the following includes some of the features comprised in the synthetic dataset; transaction type, transaction amount, sender and receiver bank locations, amongst others. The features were generated through probabilistic modeling to warrant randomness and to reflect the behaviors of real-world banking transactions. Features were chosen precisely through analyzing the current AML literature [24], existing money laundering datasets [25] [26], and attaining specialist input. Assessing the features of a real dataset allowed to identify crucial features to help identify suspicious transactions that banks have access to.

IV. Dataset Description

A total of 12 features are contained in the dataset which were chosen due to their associations with anti-money laundering transactions. Fig2, presents a preview of the generated transactions. The synthetic dataset contains a comprehensive perspective of transactions across multiple banks, focusing on the United Kingdom, as opposed to creating a singular bank viewpoint of the transactions. This broader view of transaction flows enables the measurement of performance enhancements in transaction monitoring, highlighting the potential if banks were to share data. This could be valuable to encourage data sharing amongst financial institutions in the future [27]. Creating an individual bank's viewpoint of transactions can be achieved by slightly modifying the generator.

The dataset includes 'Time' and 'Date' features denoting transactional chronology essential to identifying money laundering techniques. 'Sender' and 'Receiver' account details, together with time and date, uncover behavioral patterns and complex banking connections. The 'Amount' feature presents the transaction amounts and can highlight potentially suspicious activities through unusual transaction values. 'Payment Type' represents various transaction methods, each carrying distinct risk levels and regulations. The payment types included in our dataset are; credit card, debit card, cash, automated clearing house (ACH) transfers, cross-border, and cheque. Geographic context, inserted through 'Sender Bank Location' and 'Receiver Bank Location', identifies high-risk regions. High-risk countries such as Mexico, Turkey, Morocco, and the United Arab Emirates are included [21] [22]. 'Payment Currency' and 'Receiver Currency' align with location features, with mismatched instances adding complexity. Finally, the binary 'Is Suspicious' feature differentiates between normal and suspicious transactions, with 'Type' further classifying the typologies, providing deeper insights into prevalent or high-risk transactional typologies.

A. Typologies

A total of 28 typologies are included in the dataset, split between 11 normal and 17 suspicious. The typologies were chosen based on existing datasets [9], money laundering literature [28], and through semi-structured interviews with AML specialists. The typologies can be expressed using graphical networks to visualize the structure and flow of the transactions. There are 15 different graphical network structures for the 28 typologies. Fig 3, presents three out of the 15 different types of structures that the various typologies adopt. Multiple typologies have the same structure to increase complexity, however, the parameters can diverge significantly (i.e. transaction amounts, duration, receiver location, and entities involved). The typologies with the same structure will have different parameter values that also overlap with each other to increase intricacy, mimic real-world situations, and make detection harder. This section discusses some of the typologies included in our dataset, the ones that are selected from the literature, and the typologies produced by the authors.

The AMLSim dataset by IBM proposes 6 normal typologies and 8 AML typologies [9]. However, in their simulated dataset, only the following 3 out of the 8 AML typologies are included;

Time	Date	Sender_account	Receiver_account	Amount	Payment_currency	Received_currency	Sender_bank_location	Receiver_bank_location	Payment_type	Is_Suspicious	Type
11:08:46	2023-07-27	4526129299	6923954937	3446.97	UK pounds	UK pounds	UK	UK	Credit card	0	Normal_Fan_In
13:16:05	2022-12-27	3800564863	4957380430	5969.64	UK pounds	UK pounds	UK	UK	Credit card	0	Normal_Fan_Out
15:16:00	2023-05-25	6804140050	1708350588	3292.48	US dollar	UK pounds	USA	UK	Cross-border	0	Normal_Fan_Out
07:02:55	2023-06-27	6877419076	6895020634	4058.76	US dollar	UK pounds	USA	UK	Cross-border	0	Normal_Fan_Out
09:31:48	2022-11-20	1203205349	4634880202	5599.60	UK pounds	UK pounds	UK	UK	Debit card	0	Normal_Fan_In

Fig. 2. Preview of the generated transactions with features

Fan-Out, Fan-In, and Cycle. All the proposed normal typologies are included. In our dataset generation method, we included the following suspicious typologies proposed by IBM; Fan- Out, Fan-In, Cycle, Bipartite, Stacked Bipartite, Scatter-Gather, and Gather-Scatter. The normal typologies adopted from IBM in our dataset are Single Transaction, Fan-Out, Fan-In, Mutual, Forward, and Periodical. The typologies in our dataset adopt the structure of the proposed IBM typologies but increase complexity and randomness by having varying parameter ranges and varying numbers of entities involved each time the typology is generated. The authors further developed the Fan-In and Fan-Out typologies by increasing the transaction layers, thus producing the Layered Fan-In and Layered Fan-Out typologies. These typologies represent the layering process in money laundering [29]. One generation of the Layered Fan-In typologies structure is shown in Fig 3, illustrating multiple sender accounts transacting with a fewer number of accounts, who subsequently transact with a single receiver. The Layered Fan-Out typologies mirror this structure but with transactions flowing in the reverse direction. The typologies can accommodate a varying number of accounts within each layer, and the transaction values can increase or decrease by margins of 10% to 20% across each layer depending on the typology.

During the semi-structured interviews, the participants were asked about typologies that are difficult to detect and scenarios that generate high false positives currently in the industry using existing transaction monitoring methods. High-risk transaction types and their characteristics were also discussed, with emphasis put on cash transactions and high-risk geography by the interviewees [30]. Typologies were chosen and created using the results from interviews and the academic literature. The following completes the suspicious typologies included in our dataset; Structuring, Smurfing, Over-Invoicing, Deposit-Send, Cash Withdrawal, Single Large Transaction, Behaviour Change 1, and Behavioural Change 2. The newly incorporated normal typologies consist of Cash Withdrawal, Cash Deposit, Small Fan-out, Mutual Plus, and Normal Group. During the interviews, the Smurfing [31] and Structuring [32] typologies were frequently mentioned which are considered the most prevalent techniques in the literature concerning the placement stage of money laundering [20] [33]. The suspicious Cash Withdrawal typology was created in response to the interviews. It was stated that preventing activities like forced sexual servitude poses a challenge in AML, given the difficulty in detecting low-value withdrawals [34]. The Deposit-Send typology is considered suspicious due to the rapid movement of funds and potentially facilitating terrorism finance [35]. This typology refers to a situation where an account first deposits cash into the bank and then within a short period of time sends it to another account. The transaction amount is generally below the reporting threshold limit, with the second transaction having an increased chance of being sent to a high-risk country. Another challenge institutions currently face, identified during the interviews, is detecting and monitoring changes in the behavior of customers. Therefore, Behavioural Change 1, Behavioural Change 2, and Normal Group typologies were created. The Normal Group typology entails an account (main account) that regularly transacts with another group of accounts. The group of accounts is split into two, core accounts and regular accounts. The main account engages in transactions with the core accounts more frequently than the regular accounts. The Behavioural Change 1 and 2 typologies adopt the same structure as the Normal Group typology. However, in Behavioural Change 1, the main account deviates from its usual patterns and transacts with new accounts. In contrast, under the Behavioural Change 2 typology, the main account transacts with new accounts in high-risk locations.

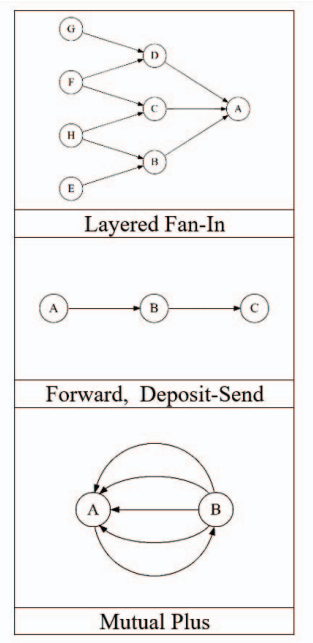

Fig. 3. Graphical structures of different typologies

V. Results

We used the proposed generator and typologies to create a synthetic transactional dataset for transaction monitoring. The parameters of the dataset were drawn from various sources to enhance credibility and practicality. The knowledge and experience of an AML specialist's input was the main source, which played a fundamental role in deciding the parameters for the normal and suspicious typologies. The criteria for setting the parameters were also determined from assessing the AML literature [36] and existing transaction datasets [9] [12]. In this section, we compare our dataset to publicly available AML datasets and perform statistical analysis.

A. Comparison

This section presents the analysis and comparison of datasets specifically designed for money laundering research. These datasets comprise of AMLSim, IT-AML, MLDP, and

TABLE I
COMPARISON OF DIFFERENT TRANSACTION DATASETS

Dataset Name	No. of Features	No. of Transactions	No. of Accounts	No. of SAR transactions	Types of Transactions
AMLSim	8	12 476 012	100 000	17052 (0.137%)	1 - Transfer
PaySim	11	6 362 620	6 353 307	8213 (0.129%)	5 - Cash-in, Cash-out, Debit, Payment, Transfer
IT-AML	11	5 078 345	515 000	5177 (0.102%)	7 - Credit Card, ACH, Wire, Cheque, Cash, Bitcoin, Reinvestment
Credit Card Fraud Detection	31 (anonymized)	284 807	N/a	484 (0.172%)	N/a
MLDP	7	2340	2340	1399 (60%)	2 - Transfer, Cash-in
SAML-D	12	9 411 384	749 507	11658 (0.124%)	6 - Credit Card, Debit Card, ACH, Cheque, Cash, Cross border

TABLE II
COMPARATIVE ANALYSIS OF KEY CHARACTERISTICS ACROSS TRANSACTION DATASETS

	AMLSim	PaySim	IT-AML	Credit Card Fraud Detection	MLDP	SAML-D
Suspicious Typology Explanation	Yes (3)	No	Yes (8)	No	No	Yes (17)
Normal Typology Explanation	Yes (6)	No	Yes (6)	No	No	Yes (11)
Model Multiple Currencies	No	No	Yes	N/A	No	Yes
Model Geographic Locations	No	No	No	N/A	No	Yes
Labelled Typologies	Yes	No	No	No	No	Yes

our newly proposed Synthetic Anti-Money Laundering Dataset (SAML-D). Tables I and II provide a comparison of the datasets.

The AMLSim, IT-AML, and SAML-D datasets all include a practical and large amount of transactions. They incorporate a realistic ratio of money laundering to normal transactions, producing an imbalanced dataset that represents real-world circumstances accurately. Additionally, these datasets provide flexibility by allowing for an adjustable volume of transactions during the generation phase. Conversely, the MLDP dataset offers a comparatively limited scope with only 2340 transactions. Moreover, due to an unrealistic proportion of money laundering to normal transactions, the MLDP dataset deviates from what is typically considered as real-world contexts. The IT-AML and SAML-D datasets both include various types of transactions, hence offering a more complete perspective than other datasets in comparison. For instance, AMLSim and MLDP include one and two transaction types respectively, presenting a more simplified view of transaction dynamics. The SAML-D dataset includes 28 typologies offering a richer and more nuanced scope compared to the 14 in IT-AML, 9 in AMLSim, and 3 in MLDP. The explanations for each typology are available in the SAML-D, IT-AML, and AMLSim datasets, facilitating a deeper understanding of the data. Whereas, in the MLDP dataset no explanation of the types of money laundering activities is given, though they are labeled within the data. Likewise, the SALM-D dataset labels each typology, enhancing the interpretability of the dataset meaning a higher accuracy and performance can be reached. The typology label feature is not present in either AMLSim or IT-AML. A unique characteristic of the IT-AML and SAML-D datasets is the incorporation of transactions involving multiple currencies, which adds a layer of complexity and a higher degree of realism to these datasets. The SAML-D dataset further enhances its value and realism through the addition of geographic locations, even featuring high-risk countries. This feature aligns the dataset more closely with features often employed in the industry.

Overall, our proposed dataset, SAML-D, provides a detailed and robust resource for transaction monitoring in the field of AML. SAML-D contributes to the AML domain by introducing innovative features and typologies facilitating effective and complex analysis of transaction monitoring approaches.

B. Experiment

In this section we conduct basic machine learning experiments on our newly developed dataset, SAML-D, and the AMLSim dataset, to detect suspicious transactions. The AMLSim dataset serves as our comparison point due to its utilization within the AML literature. We aim to establish the SAML-D dataset's suitability, purpose, and applicability within the field of AML.

The machine learning algorithms chosen for our experiments comprise various approaches to better capture the dataset's dynamics. These include Support Vector Machines (SVM), a distance-based model adept at handling complex

class boundaries; Naïve Bayes (NB), a probabilistic model, Decision Trees, a tree-based model ideal for understanding feature importance and their interplay; and Random Forest, an ensemble model. These methods were chosen as they are some of the most utilized techniques in the literature [8] [37].

Data analysis was the first step to get a better understanding of the data, identify the datatypes for each feature, and check for missing values. Next, the data was pre-processed, transforming it into a suitable format for the chosen machine learning algorithms. Preprocessing involved converting categorical features, such as payment type, into numerical form by the use of one-hot encoding and label encoding. This allowed the algorithms to interpret the information more effectively [37]. The date feature was split into year, month, and day, while the is fraud feature in AMLSim was converted from a boolean datatype to an integer. After redundant columns were dropped, both datasets were standardized rescaling the numerical variables to a comparable scale. This was especially important for the transaction amount feature, given its excessive range. Also, some algorithms like SVM are sensitive to the scale of the input features and therefore perform better on standardized data.

We adopted a 70-15-15 stratified train-validation-test split for our experiments. The 70% training data enabled to effectively fit the model to the data, while the 15% validation data was utilized for preliminary performance evaluation. Finally, the 15% test data was used to evaluate the models. This approach helped prevent model overfitting and enabled generalized results. Due to the constraint of a single GPU, our experiments were conducted on a representative subset of the datasets.

The evaluation metrics used to assess the performance of our models include the True Positive Rate (TPR), True Negative Rate (TNR), False Positive Rate (FPR), False Negative Rate (FNR), and Area Under the Curve (AUC) score. These metrics are widely used in the AML literature and provide valuable insights into the model's performance [38]. The TPR measures the proportion of actual suspicious transactions that are correctly identified. Assessing the TPR is crucial, as missing suspicious transactions can result in significant consequences for banks in the form of fines and potential legal action by authorities. The TNR indicates the percentage of normal transactions that are correctly labeled as normal. The evaluation metrics also assessed the model's error. The FPR quantifies the transactions that are incorrectly labeled as suspicious by the model, which can lead to high operational costs and wasted resources for banks. On the contrary, the FNR presents the rate of suspicious transactions that are mislabelled as normal and go undetected by the institution. A high FNR can result in reputational damage and regulatory fines. The AUC score represents the model's overall performance and efficiency in classifying between normal and money laundering transactions. In summary, the chosen evaluation methods provide both an overview of the model's performance and a nuanced analysis of its specific elements.

All the model's performances had a lower TPR and higher

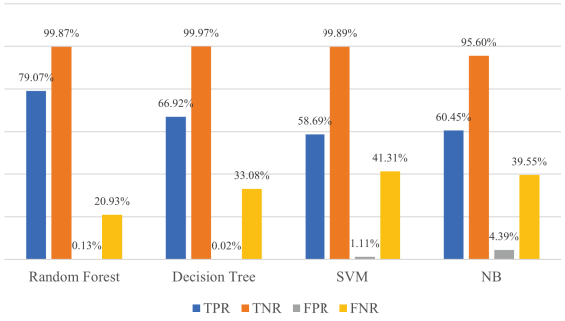

Fig. 4. Experiments results for the SAML-D dataset

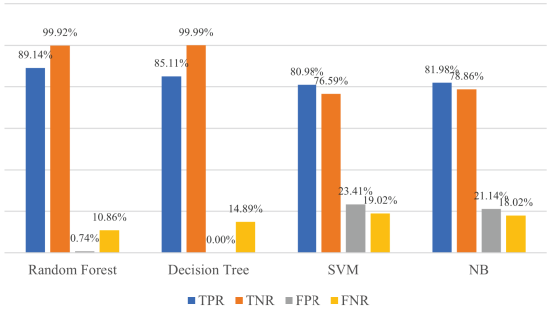

Fig. 5. Experiments results for the AMLSim dataset

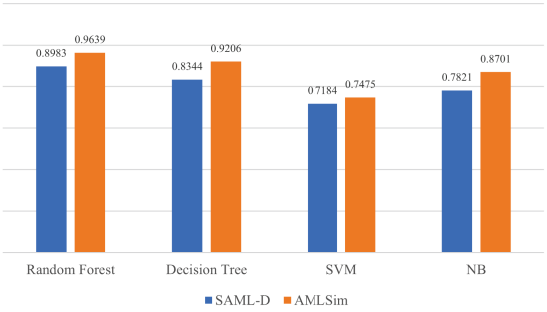

Fig. 6. Comparison of AUC score for the SAML-D and AMLSim datasets

FNR on the SAML-D dataset compared to the AMLSim dataset, as shown in Fig 4 and Fig 5. The results indicate that these models encountered challenges in detecting money laundering transactions in the SAML-D dataset, implying a higher level of complexity or deception in the money laundering structures and patterns. Despite the lower FPR observed in the AMLSim dataset, the models successfully detected most of the money laundering transactions. This capability is crucial and of greater importance for financial institutions than having a model with a higher FPR, as failure to detect suspicious behaviors could potentially lead to fines and reputational damage. The AUC score presents a similar trend with the SAML-D dataset attaining a lower score for

every model, shown in Fig 6. This emphasizes the difficulty these machine learning models had in detecting suspicious transactions in the SAML-D dataset. Our findings demonstrate that while these machine learning models perform well on the AMLSim dataset their performance weakens on the SAML dataset. This highlights the need for sophisticated and robust models. Potential approaches could involve feature engineering to produce more predictive attributes and exploring more advanced or tailored machine learning models.

VI. Conclusion

In conclusion, this research addresses the challenge of accessing AML transaction monitoring data, which is typically unavailable due to legal and privacy constraints, or limited in terms of true labels and diversity. Our novel synthetic AML transaction generator provides a valuable resource to advance transaction monitoring in the field of AML. Using the generator a dataset called SAML-D was created. The primary purpose of the generator and dataset is to provide researchers with an additional resource to evaluate their models and facilitate a comparative analysis of their results.

The SAML-D dataset contains 12 features and 28 typologies, including both 'normal' and 'suspicious' entities. The typologies were chosen and created based on existing datasets, the literature, and 8 semi-structured interviews with AML specialists. Our approach brings significant enhancements compared to other synthetic datasets such as AMLSim, IT-AML, and MLDP. Key enhancements include the introduction of geographic locations and the attention given to high-risk payment types, giving the dataset a high degree of realism that reflects the complexities seen in real-world industry situations. Additionally, the SAML-D dataset includes a vaster selection of typologies offering a richer and more nuanced scope compared to the other datasets.

In testing the SAML-D dataset against the more established AMLSim dataset through machine learning experiments, we found that models had more difficulty identifying suspicious transactions within the SAML-D dataset. This implies a higher level of complexity and difficulty in the money laundering patterns and structures, making SAML-D a valuable resource for future studies in this domain.

Despite its contribution, it is vital to recognize that the SAML-D dataset has limitations. As it is a synthetic dataset, it will not fully capture the intricacy and unpredictability of real-world transactions. Some parameters, while based on informed estimations and expert consultations, may not embody all real-world scenarios. Moreover, the typologies included will not encapsulate all possible money laundering strategies, particularly due to the criminal's ever-changing techniques. However, the SAML-D dataset offers a resource for researchers and practitioners to conduct experiments and compare their findings, potentially assisting the development of more advanced and capable transaction monitoring methods.

As for future work, we aim to conduct experiments with more complex machine learning algorithms on the SAML-D dataset, leading to further improvements in detecting suspicious activities.

References

[1] M. S. Korejo, R. Rajamanickam, and M. H. Md. Said, "The concept of money laundering: a quest for legal definition," Journal of Money Laundering Control, vol. 24, no. 4, pp. 725-736, 2021.

[2] M. A. Naheem, "Money laundering: A primer for banking staff," International Journal of Disclosure and Governance, vol. 13, no. 2, pp.135-56, 2016.

[3] E. Eifrem, "How graph technology can map patterns to mitigate money-laundering risk," Computer Fraud & Security, vol. 2019, no. 10, pp.6-8, 2019.

[4] A. I. Canhoto, "Leveraging machine learning in the global fight against money laundering and terrorism financing: An affordances perspective," Journal of Business Research, pp. 441-452, 2021.

[5] M. Jullum, A. Løland, R. B. Huseby, G. Ånonsen, and J. Lorentzen, "Detecting money laundering transactions with machine learning," Journal of Money Laundering Control, vol. 23, no. 1, pp. 173-186, 2020.

[6] X. Cheng et al., "Combating emerging financial risks in the big data era: A perspective review," Fundamental Research, vol. 1, no. 5, pp. 595-606, 2021.

[7] Europol, "From Suspicion to Action: Converting financial intelligence into greater operational impact," 2017. Available: https://www.europol.europa.eu/cms/sites/default/files/documents/ql-01-17-932-en-c_pf_final.pdf.

[8] B. Oztas, D. Cetinkaya, F. Adedoyin, and M. Budka, "Enhancing Transaction Monitoring Controls to Detect Money Laundering Using Machine Learning," in 2022 IEEE International Conference on eBusiness Engineering (ICEBE), pp. 26–28, 2022.

[9] T. Suzumura and H. Kanezashi, "Anti-Money Laundering Datasets: InPlusLab Anti-Money Laundering DataDatasets," 2021. [Online]. Available: http://github.com/IBM/AMLSim/.

[10] A. Tundis, S. Nemalikanti, and M. Mühlhäuser, "Fighting Organized Crime by Automatically Detecting Money Laundering-Related Financial Transactions," in Proceedings of the 16th International Conference on Availability, Reliability and Security, Vienna, Austria, 2021, Art. No. 38.

[11] M. Shokry, A. Ehab, M. A. Rizka, and N. M. Labib, "Counter terrorism finance by detecting money laundering hidden networks using unsupervised machine learning algorithm," International Conference on ICT, Society & Human Beings, pp. 89-97, 2020.

[12] E. Altman et al., "Realistic Synthetic Financial Transactions for Anti-Money Laundering Models," arXiv preprint arXiv:2306.16424, 2023.

[13] M. Mahootiha, "Money laundering data," 2020. [Online]. Available: https://www.kaggle.com/datasets/maryam1212/money-laundering-data.

[14] E. A. Lopez-Rojas and S. Axelsson, "Money laundering detection using synthetic data," in Annual Workshop of the Swedish Artificial Intelligence Society (SAIS), Linköping University Electronic Press, Linköping University, 2012.

[15] Y. A. Le Borgne, W. Siblini, B. Lebichot, and G. Bontempi, Reproducible Machine Learning for Credit Card Fraud Detection - Practical Handbook, Université Libre de Bruxelles, 2022. [Online]. Available: https://github.com/Fraud-Detection-Handbook/fraud-detection-handbook.

[16] S. H. Chen and R. Venkatachalam, "Agent-based modelling as a foundation for big data," Journal of Economic Methodology, vol. 24, no. 4, pp. 362–383, 2017.

[17] D. Valbuena, P. H. Verburg, and A. K. Bregt, "A method to define a typology for agent-based analysis in regional land-use research," Agriculture, Ecosystems & Environment, vol. 128, no. 1, pp. 27–36, 2008.

[18] K. Plaksiy, A. Nikiforov, and N. Miloslavskaya, "Applying big data technologies to detect cases of money laundering and counter financing of terrorism," in 2018 6th International Conference on Future Internet of Things and Cloud Workshops (FiCloudW), pp. 70–77, IEEE, 2018.

[19] Ping. He, "A typological study on money laundering," Journal of Money Laundering Control, vol. 13, no. 1, pp. 15–32, 2010.

[20] S. M. Irwin, A. Raymond Choo, and L. Liu, "Modelling of money laundering and terrorism financing typologies," Journal of Money Laundering Control, vol. 15, no. 3, pp. 316–335, 2012.

[21] "National risk assessment of money laundering and terrorist financing 2020," UK Government, 2020. [Online]. Available: https://www.gov.uk/government/publications/national-risk-assessment-of-money-laundering-and-terrorist-financing-2020
[22] "High-risk and other monitored jurisdictions - June 2023," Financial Action Task Force (FATF), 2023. [Online]. Available: https://www.fatf-gafi.org/en/publications/High-risk-and-other-monitored-jurisdictions/Increased-monitoring-june-2023.html
[23] M. C. Johnson, and S. R. Kessler, "The art and science of semi-structured interviewing: A comprehensive guide for researchers," Qualitative Research Journal, vol. 21, pp. 131–147, 2019.
[24] J. de J. Rocha-Salazar, M. J. Segovia-Vargas, and M. del M. Camacho-Miñano, "Money laundering and terrorism financing detection using neural networks and an abnormality indicator," Expert Systems with Applications, vol. 169, 114470, 2021.
[25] R. Desrousseaux, G. Bernard and J. -J. Mariage, "Profiling Money Laundering with Neural Networks: a Case Study on Environmental Crime Detection," in 2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI), Washington, DC, USA, 2021, pp. 364–369.
[26] Z. Rouhollahi, A. Beheshti, S. Mousaeirad, and S. R. Goluguri, "Towards Proactive Financial Crime and Fraud Detection through Artificial Intelligence and RegTech Technologies," in The 23rd International Conference on Information Integration and Web Intelligence, Linz, Austria, 2022, pp. 538—546.
[27] M. Betron, "The state of anti-fraud and AML measures in the banking industry," Computer Fraud & Security, vol. 2012, no. 5, pp. 5–7, 2012.
[28] J. Simser, "Money laundering: emerging threats and trends," Journal of Money Laundering Control, vol. 16, no. 1, pp. 41–54, 2013.
[29] H. Heinrich-Böll-Stiftung and R. Schönenberg, Eds., Transnational Organized Crime: Analyses of a Global Challenge to Democracy, Transcript Verlag, 2013. [Online]. Available: http://www.jstor.org/stable/j.ctv1fxh0d
[30] M. Riccardi and M. Levi, "Cash, Crime and Anti-Money Laundering," in The Palgrave Handbook of Criminal and Terrorism Financing Law, C. King, C. Walker, and J. Gurulé, Eds. Cham: Palgrave Macmillan, 2018.
[31] M. Starnini et al., "Smurf-Based Anti-money Laundering in Time-Evolving Transaction Networks," in Machine Learning and Knowledge Discovery in Databases. Applied Data Science Track, vol. 12978, Y. Dong, N. Kourtellis, B. Hammer, and J. A. Lozano, Eds. Cham: Springer, 2021.
[32] M. M. El-Banna, M. H. Khafagy and H. M. El Kadi, "Smurf Detector: a Detection technique of criminal entities involved in Money Laundering," in 2020 International Conference on Innovative Trends in Communication and Computer Engineering (ITCE), Aswan, Egypt, 2020, pp. 64–71.
[33] B. Unger and E. M. Busuioc, "The Scale and Impacts of Money Laundering," Edward Elgar, Cheltenham, UK; Northampton, MA, 2007. [Online]. Available: http://www.loc.gov/catdir/toc/ecip074/2006035573.html
[34] J. McDowell and G. Novis, "The consequences of money laundering and financial crime," Econ. Perspect., An Electron. J. U.S. Dep. State, vol. 6, no. 2, May 2001.
[35] S. Gao, D. Xu, H. Wang and Y. Wang, "Intelligent Anti-Money Laundering System," in 2006 IEEE Int. Conf. Serv. Oper. Logist., Informatics, Shanghai, China, 2006, pp. 851-856.
[36] M. A. Naheem, "Money laundering: A primer for banking staff," Int. J. Disclos. Gov., vol. 13, no. 2, pp. 135-156, May 2016.
[37] R. M. Suresh and R. Padmajavalli, "An overview of data preprocessing in data and web usage mining," in 2006 1st Int. Conf. Digit. Inform. Manage., pp. 193-198, IEEE, 2006.
[38] A. A. S. Alsuwailem and A. K. J. Saudagar, "Anti-money laundering systems: a systematic literature review," J. Money Laund. Control, vol. 23, no. 4, pp. 833-848, 2020.