Journal of Cases on Information Technology
Volume 25 • Issue 1

Intelligent Anti-Money Laundering Fraud Control Using Graph-Based Machine Learning Model for the Financial Domain

Atif Usman, Department of Computer Science and Information Technology, Virtual University of Pakistan, Pakistan
https://orcid.org/0000-0002-8130-7957

Nasir Naveed, Department of Computer Science and Information Technology, Virtual University of Pakistan, Pakistan

Saima Munawar, Department of Computer Science and Information Technology, Virtual University of Pakistan, Pakistan*
https://orcid.org/0000-0002-2446-3670

ABSTRACT

Financial domains are suffering from organized fraudulent activities that are inflicting the world on a larger scale. Basel Anti-Money Laundering (AML) index enlists 146 countries, which are impacted by criminal acts like money laundering, and represents the country's risk level with a notable deteriorating trend over the last five years. Despite AML being a substantially focused area, only a fraction of such activities has been prevented. Because financial data related to this field is concealed, access is limited and protected by regulatory authorities. This paper aims to study a graph-based machine-learning model to identify fraudulent transactions using the financial domain's synthetic dataset (100K nodes, 5.3M edges). Graph-based machine learning with financial datasets resulted in promising 77-79% accuracy with a limited feature set. Even better results can be achieved by enriching the feature vector. This exploration further leads to pattern detection in the graph, which is a step toward AML detection.

KEYWORDS

Anti-Money Laundering, Machine Learning, Networks, Semi-Supervised Learning, Tensorflow, Transactions

1. INTRODUCTION

Financial domains travail from organized fraudulent activities, which in turn affects the economy of the organization as well as the national level (Truman & Reuter, 2004). These activities involve the financial sector as a medium to transfer funds, but the factual magnitude of money laundering is uncertain and even unknown in most cases. Every country is forced by law in this global fight against money laundering. UNO2 estimates the total amount of funds circulated under money laundering cover in a meta-analysis on drugs and crimes, about 2.7% of global GDP, i.e., 1.6 trillion USD (Pietschmann & Walker, 2011).

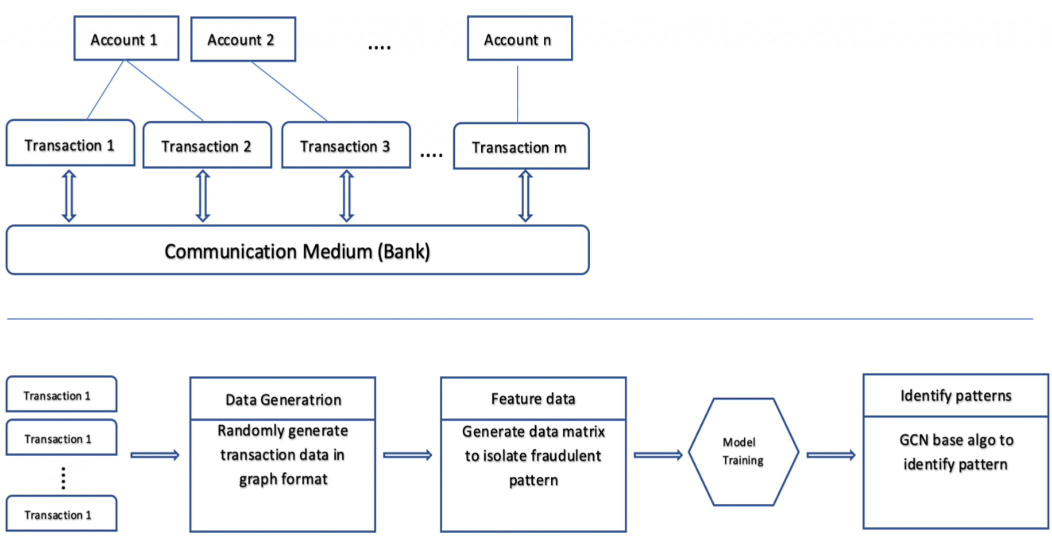
DOI: 10.4018/JCIT.316665
*Corresponding Author

Another analysis by Walkers in the late '90s mentioned 2.85 trillion USD involved in money laundering worldwide. Money laundering is a method of obtaining money generated from criminal activity and masking these funds that appear to be clean and originated from a legitimate source, i.e., the process of illegally gaining assets appears legal (IMF, 2020). The most common criminal activities that the world faces nowadays are terrorist funding, drug smuggling, and human trafficking.

Figure 1 illustrates the concept of transactions performed by agents on different bank accounts and these transactions performed using cash, cheque, electronic banking (ATM) or mobile banking uses a banking system for fulfilment. Theoretically, these transactions can be used intelligently to train a machine learning model which used advanced learning capabilities to identify and even predict potential money laundering attempts. Basel AML index (Manning, Wong, &Jevtovic, 2020) shows the list of countries that are heavily victimized by this act and caused severe damage to economies. The annual report from the UN States department "The International Narcotics Control Strategy Report" (INCSR) estimates illicit financial flows put over $10bn in Pakistan (Rates, Guides, Center, Clinton, & Hotline, 2017). This shows the vast radius and substantial impact these fraudulent activities are generating on societies and economies over the globe (Omar, Johari, & Arshad, 2014), which makes this area fundamentally essential to detect at as early a stage as possible (Schott, 2006).

Act against Money laundering becomes obligatory for any country because of the impact that it can create e.g. government drop control over economic policies, elevating conceivable failure in the banking system, and small to medium-scale businesses (Qureshi, 2017). Another adverse effect it can cause is in the private sector. The inflow of out-sized capital provides an ability to the companies where marketing penetration can be achieved by lowering the cost prices of goods at a significant scale and gaining a competitive advantage over others where there is no one to match their prices. Such consequence hits adverse and damaging impacts on small to medium-scaled businesses. In the same way, the large inflow of money for a short time in a bank can cause liquidity problems, leading to bankruptcy. Further on, it has various other impacts on economic policies, the reputation of financial institutions, the corruption perception index (CPI), and social effects. Most of the work done to handle fraudulent scenarios within financial domains are alert systems focused on linear methods and designed with threshold rules to work with suspicious transactions. This means that the system generally ingests marketing data transactions and applies defined rules to mark them as faulty or fraudulent transactions. Following problems usually appear in such AML systems.

Figure 1. The overall architecture of the system outlining the scope and concept



- Always keep up the system's rules up to date.
- Continuous updates in rules by data analysis (Alexandre & Balsa, 2015),(Gao & Ye, 2007).
- Counter the number of false alerts from millions of transactions

These are the elements triggering the need to have a state-of-the-art AML tool (Deng, Joseph, Sudjianto, & Wu, 2009) discussed the statistical method of typical fraud detection) but very little literature is available addressing this topic. On a broader level, currently, two Machine learning-base Methods are available, which can be used in AML, i.e., supervised, and unsupervised learning. Graph network and graphical data analytics have recently been raised as important tools and have been used in many domains, but this method has not been published or industrially implemented in the financial domain. A different tactic is used to handle this problem, instead of a linear approach, the graph-base model is to represent the data and used a semi-supervised machine learning model to solve the problem and demonstrate it. This approach will allow using N number of features to isolate the fraudulent activity and report it.

1.1 Research Question

How can a graph-based machine learning model achieve an effective and timely reaction to a criminal financial investigation?

The paper is organized as: Section 2 describes the money laundering concept, the AML process, and its importance to curbing at the national stage. It further discusses previous relative works and techniques used to tackle money-laundering cases. Section 3 explains the data and methodology used. Finally, the results, discussion, and conclusion are presented in sections 4th, 5th& 6th.

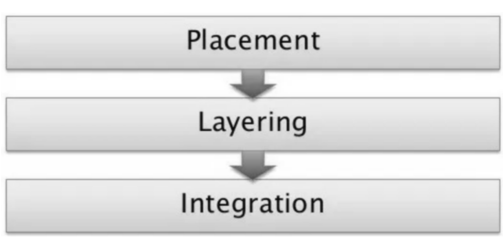
2. REVIEW OF LITERATURE

The main objective is to analyze different machine learning methodologies and approaches that have been used in the past in the anti-money laundering context. Hence this research includes (Sect 2.1) appropriate knowledge of anti-money laundering, (Sect 2.2) Categorization, and (Sect 2.3) Research methodology and protocol.

2.1. Anti-Money Laundering

Money Laundry has been declared an illegal activity by international financial and legal authorities because it is tightly linked with various crimes of serious nature like funding to criminal organizations e.g., terrorists who operate over money laundering networks. Problem with this area of work other than regulatory and compliance (which restricts the access of data), there is a variety of ways within financial and commercial activities where fraud can occur; cash, digital money, credit card payments, offshore real-estate, and wire transfers are few major means to Money laundering (States, 2009) (IMF, 2020). Each method of money laundering brings a distinct level of complexity associated with it (A. D. Bank, 2003)(Unger & Van der Linde, 2013). Although money laundering is diverse and complicated, it generally contains three phases: placement, layering, and integration (Choo, 2015) as shown in Figure 2.

Figure 2. Stages in the money laundering process



Placement: It is Illegal proceeds from illicit activities that are placed into financial systems. The money laundering process starts with placement where the outflow of funds is being made to foreign banks (Gee, 2014), in a way that seems legit and does not catch governing national authorities' attention.

Layering: There are various layers of multiple transactions, and wire transfers to conceal the illegal proceeds and complicate the tracing of funds. Term layering emphasizes the action of forking one significant amount to many small layers of transactions. This part of the process has become instantly fast since wire transfers; digital transactions are introduced. By which multiple transactions can hit in a short time interval (Banks, 2017).

Integration: Integration of funds into financial systems through legitimate forms. This is the stage where money is mixed back or added to the additional fund, and once the process reached this step, it is almost for authorities to trace back.

AML area a job to prevent criminals from flowing illegal funds through the financial system. The challenge that AML has introduced from the regulatory compliance authorities to the financial institutions is a responsibility to certify Know Your Customer (KYC) standards, monitor financial activities, act against the accounts believed sceptical, and generate well-timed Suspicious Activities Reports (SARs) which can be submitted to law enforcement and regulatory agencies.

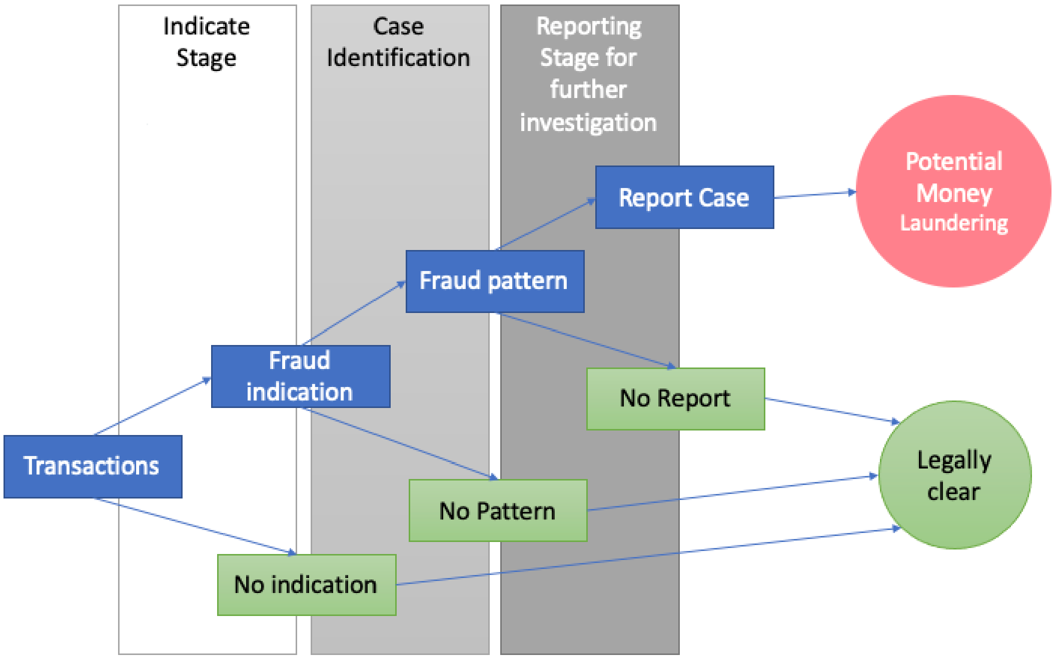
Much legislation in the past decade has been formulated against money laundering. But demands law enforcement authorities to compel these on common ground, i.e., arrest any kind of money launderers irrespective of their power, e.g., business, political or govt officials. The selective approach while enforcing AML policies keeps the corruption and money laundering seed alive and growing. Some prominent legislation made against money laundering are (Commission, 2018): AML Ordinance 2007, AML Act 2010, Amendments to AML Act 2010, AML Regulations 2010 and AML Rules 2010. The need of the time is to control money laundering activities at as early stage as possible to prevent further damages, and this Global need to comply with AML standards runs at a huge cost (tens of billions of dollars) every year. According to a survey by Klynveld Peat Marwick Goerdeler (KPMG) 2014, this cost has grown by 15% since 2004. Compliance with the AML act and standards also comes with severe penalties: In 2012, Hongkong and Shanghai Banking Corporation Limited (HSBC)(Shelley, 2014) was fined $1.9 billion when found guilty of maintaining laundry money and accepting transactions used in drug traffic. Standard Chartered (Kittrie, 2016) was fined $340 million once found guilty of money laundering and facilitating illegal transactions (Kossovsky, Greenberg, &Brandegee, 2012). Liberty reserve bank was closed by US Federal authorities once it proved involved in $6 billion (Rothchild, 2016) laundering.

In 2009 an active learning procedure was proposed by Deng (Deng et al., 2009) that works through the sequential design method. His research is focused on both synthetic data and transaction data from financial institutions. A way to identify suspiciously was proposed by Liu et al. (Liu, Zhang, & Zeng, 2008). The advantages and disadvantages to use synthetic data were discussed by Lopez (E. A. Lopez-Rojas &Axelsson, 2012) because either there is not enough data or financial institutes are reluctant/bounded not the share data. Work has been done in this field, but the most commonly available solutions are copyrighted to the technology/service provider and not easy to regulate. Some available solutions are designed to work w.r.t a specific strategy (Grint & O'Driscoll, 2017).

This research aims to provide a methodology to help detect and report fraudulent activity. And in association with methods, it is essential to describe how the AML process works within an organization (Figure 3). In general, AML procedures require a compliance department to certify compliance control and training of employees. Knowing your customers is the crucial step in the AML procedure; it should maintain customers' profiles from inception. Surveillance over activities in the account within defined constraints, spot account and identified transactions for a manual review and finally generate a proper Suspicious activity report for regulatory authorities to take further legal actions against the suspects.

Monitoring typically has three levels: 1) The trigger stage, where transactions pass through the system based on a set of rules. At this level, a thorough investigation is performed before providing

Figure 3. The Typical Process of monitoring, investigation, and reporting in banks



a detailed report and all possible proof to internal authorities. Trigger an alert if the transaction falls under fraudulent criteria and pushes for manual investigation. 2) The reported transactions of the suspect are aggregated into cases. The report usually includes the transaction itself, recent bank transaction history, account details of involved parties, etc. 3) At his stage, a detailed manual investigation is performed with additional background information. After having sufficient detail, the case is marked as potential money laundering.

Account activities are monitored by Transaction monitoring systems that are mainly rules-based engines configured for escalation on a defined threshold of volume and frequency of transactions. The fundamental problem with the Rule-base method is that they are incapable of detecting new or different money laundering typologies and can produce false positive (FP) results. FP is a standard transaction that is incorrectly detected as money laundering when it is not or vice versa. As an example, scenario, a transaction monitoring system in a bank is configured to mark all transactions above USD 10,000, near missing threshold like USD 9,500, and a series of distributed transactions ($2,000 x 5) within a 24-hour window crossing the threshold figure. On those flagged transactions analyst takes an initial review and decides whether to escalate to the compliance department. There is AML trained specialist conducts a thorough analysis in coordination with the compliance team and further decides if a suspicious activity report (SAR) needs to be filed or not and suspends the suspected account accordingly. In this case, it is also possible that many groups are operating more petite than the threshold level of the transaction and working for one brain. Many issues can be overcome to resolve the challenging techniques.

It looks simple, but if it adds one more dimension to this forensic scenario, e.g., the account owner's profile, there is more to dig into. In case of a nearly missed or distributed transaction, an analyst at the initial review pulls up a customer's profile (KYC) and finds out that the nearly missed transaction belongs to an owner who has political exposure, and the amount is credited to a foreign account in European Union (EU). He further checks more transaction activity and cases to comply for further review. This forensic behaviour is immensely complex because there are other dimensions besides KYC, e.g., geographic dynamics, Circling, transaction type, properties, etc. FPs and threats are probable to fines from the regulation group on top. Second, such a solution is complicated to scale concerning time and effort. Third, the deplorable group are still winning, and they are well ahead in refining, masking, and layering the accounts to get them denser. A refined AML solution should require overcoming these challenges.

2.2. Categorization

Three main categories are identified against the research question.

2.2.1 Machine Learning

Machine learning is used to classify, cluster, and identify/predict patterns in data (Jiawei Han, Kamber, & Pei, 2012). Supervised learning (i.e., for labelled data) and unsupervised learning (i.e., for unlabeled data) are the main two classifications of machine learning techniques. Supervised learning algorithms demand more data management to find familiarizes between input data and its output. On the other hand, unsupervised learning identifies hidden patterns internal to input data exclusive of labelled responses. In situations where acquiring some concrete labelled training sets is infeasible or very costly. A small set of labelled training datasets can be used with semi-supervised learning, which refers to transductive learning to infer correct labels for unlabeled data. Semi-supervised learning is the most feasible approach to follow in each scenario. It can be of great practical value, as it aims to find graphical data patterns using feature sets (i.e., input) and the model learns to capture the similar (i.e., outcome).

2.2.2 Data Analytics

Data analytic techniques (Elgendy&Elragal, 2014) are used by companies on large datasets to understand their businesses better, customers, build strategies, develop products and detect anomalies in the industries.

2.2.3 Multi-Technology Approach

There are two phases of the intelligent methods (Tai & Kan, 2019) approach in multi-technology. Machine learning techniques are used with data analytics methodologies to gain the positives of both technologies and bring the most effective result.

2.3. Research Methodology and Protocol

Research methodology starts with the identification of scientific databases which host articles, research context, journals, etc. Five major online scientific databases were selected: Springer, ACM Digital Library, Taylor & Francis DL, IEEE Xplore, and Elsevier-Science Direct. The literature search process starts with the definition of the criteria to enlist the articles inclusion or exclusion protocols from the analysis that the article must (1) be published in a computer science discipline, (2) be in content type, conference paper, journal, article or chapter (2) be written in English, (3) be published between 2017 and Jan-2021, and (4) have its text search string available in at least one of the aforementioned five databases.

To achieve a comprehensive search strategy, that tried multiple strings using Boolean expressions to combine the terms, but to keep the research focused, used "anti AND money AND laundering" as a final term to continue. Almost 267 papers in total met the inclusion criteria. After that, further pruned the articles by "reading titles, Abstract Reading and quick Skimming through Introduction", eliminating 211 unrelated articles, journal papers, white papers, and chapters. The remaining 56 papers were taken to the next level of "Reading the full Article." This process further excluded another group of 22 unrelated papers and systematic reviews; the remaining 34 articles were taken to the final analysis of the literature review as indicated in Table 1 and Table 2.

Table 1. Search term and inclusion criteria results

Search Criteria	ACM	IEEE	Elsevier	T&F	Springer
Query String: “Anti+ Money+ Laundering”	66	68	502	721	2049
Period: 2017 –Jan 2021	41	25	229	243	978
Discipline: Computer Science			74	6	124
Filter Conference OR Journal OR Article	41	24	59	1	71
Filter relevant abstract	9	16	10	1	20
Total consideration	56				

Table 2. Statistics of selected research work

Sr	Category	Total	Research identification
1	Machine Learning	10	(Feng et al., 2019); (Tertychnyi, Slobozhan, Ollikainen, & Dumas, 2020); (Guevara, Garcia-Bedoya, & Granados, 2020); (Mocko & Ševcech, 2018); (Seifi & Ekhveh, 2019); (D. Bank, 2020); (Pambudi, Hidayah, & Fauziati, 2019); (Pelckmans, 2020); (Luna, Palshikar, Apte, & Bhattacharya, 2018); (Chen, 2020)
2	Data Analytics	10	(Adedoyin, Kapetanakis, Samakovitis, & Petridis, 2017); (Khanuja & Adane, 2018); (Hanbar, Shukla, Modi, & Vyjayanthi, 2019); (Plaksiy, Nikiforov, & Miloslavskaya, 2019); (Jin & Qu, 2018); (Hamid, 2017); (Li, Cao, Qiu, Zhao, & Zheng, 2017); (Prakash, Apoorva, Amulya, Kavya, & KN, 2019); (Plaksiy, Nikiforov, & Miloslavskaya, 2018); (El-Banna, Khafagy, & El Kadi, 2020); (El-Banna et al., 2020)
3	Multi-technology Approach	3	(Jonker, Habeck, Park, Jordens, & van Schaik, 2017); (Tai & Kan, 2019); (Camino, State, Montero, & Valtchev, 2017)
4	Eliminated	33	Systematic reviews, Process studies, Case studies, and Industrial surveys have been eliminated.

Table 3 presents research studies and the details of different evaluation measures taken to evaluate the outcome of their studies. The studies are divided into three main categories. (1) Machine learning; includes all those studies that used machine learning approaches in one way or another as an evaluation measure. (2) Analytics; include all those studies that used big data analytics as an alternative to finding the anomalies in the datasets. (3) Multi-technology approach; include those research papers which used both machine learning and data analytics to gain the advantage of both technologies.

AMLs area a job is to prevent criminals from flowing illegal funds through the financial system. AML has introduced from the regulatory compliance authorities to the financial institutions a responsibility to certify Know Your Customer (KYC) standards, monitor financial activities, act against the accounts believed sceptical, and generate well-timed Suspicious Activities Reports (SARs) which can be submitted to law enforcement and regulatory agencies. SAR is the report that contains information about bank accounts and transactions potentially involved in fraudulent activity. All involved transactions in the topology are flagged as SAR and listed in the report for further investigation. The need of the time is to control money laundering activities at as early stage as possible to prevent additional damages, and this Global need to comply with AML standards runs at a considerable cost (tens of billions of dollars) every year. This review aims to provide a survey with studies that happened between 2017-2021 around the AML process and their measures to solve it in different spaces.

(Feng et al., 2019) proposed a multilayered approach with two separate machine learning models; Model 1 is trained on the training set generated by the simulator while model 2 is trained on collected fraudulent transaction datasets and emphasized to connect both models serially to achieve more accurate results. A similar kind of evolution technique was used by (Tertychnyi et al., 2020) where the output of

Table 3. Mapping of dataset and evaluation measures

Category	Reference	Availability of data labels	Dataset	Types of evaluation measures
Machine Learning	(Feng et al., 2019)	S	SY	SVM, LR, MLP, Multilayer approach
	(Tertychnyi et al., 2020)	S	RW	LR, GB, data im-balancing
	(Guevara et al., 2020)	US	RW	One Class SVM, DT (IF)
	(Mocko&Ševcech, 2018)	S	SY	LR, RF, SVM, DT, accuracy and performance comparison
	(Seifi&Ekhveh, 2019)	US	SY	NLP Methods: Word2Vec, Doc2Vec, Clustering algorithm – K-Means
	(D. Bank, 2020)	S	RW	TransE, ComplEX algorithm comparison for performance and accuracy
	(Pambudi et al., 2019)	S	SY	SVM, Tuned SVM, Data imbalancing using RUS
	(Pelckmans, 2020)	US	SY	K-Means, FADO: algorithm comparisons
	(Luna et al., 2018)	US	SY	RSS, FastVOA, LOF comparison
	(Chen, 2020)	S	RW	SVM (Risk base Fn), DT(RT, RF, c4.5)
Analytics	(Adedoyin et al., 2017)	DA	SY	GA, K-Nearest neighbour
	(Khanuja&Adane, 2018)	DA	SY	Rule-base Bayesian Classification algorithm
	(Jin& Qu, 2018)	DA	SY	Loop discovery, De-looping the data, Capital flow in the hierarchy, Merge similar nodes algorithm
	(Hamid, 2017)	DA	RW	Conceptual physical Currency prediction system
	(Li et al., 2017)	DA	RW	Spark GraphX, Lauvain algorithm
	(Prakash et al., 2019)	DA	SY	Hash-based Algorithms
	(Plaksiy et al., 2018)	DA	SY	algorithm to detect and Typology Mutation
	(El-Banna et al., 2020)	DA	RW	Clustering, PageRank, Neo4j
	(He & Qu, 2019)	DA	SY	Presented algorithm to capture Periodic behaviour
Multi-Technology	(Jonker et al., 2017)	S + DA	RW	SSC, PageRank, ShortestPath, Egonet data analytical algorithm used with SVM(RBF), BIRCH, MCL
	(Tai & Kan, 2019)	S + DA	RW	SVM, Analytics
	(Camino et al., 2017)	S + DA	SY	OCSVM, IF, GMM

S, Supervised; U, Unsupervised; DA, Data analytics;
SY, Synthetic datasets; RW, Real-world datasets;
SVM, Support vector machine; LR, Logical regression; MLP, Multilayer perception; GB, Gradient boost; DT, decision tree; IF, Isolation forest;RF, Random forest; OCSVM, One-class SVM; RBF, Radial base function; BIRCH, Balanced iteration reducing and clustering using hierarchies; MCL, Markov cluster algorithm; GMM, gaussian mixture model

layer 1 will become the input of the Model at layer 2; in addition to that, a data sampling technique is also used to sample data before it inputs to training models, but no considerable performance improvement is recorded. Some studies like (Guevara et al., 2020) (Luna et al., 2018) have used unsupervised learning, on the other hand, to identify hidden patterns internal to input data exclusive of labelled responses. In situations where acquiring some concrete labelled training sets is infeasible or very costly. A small set of labelled training datasets can be used with semi/Unsupervised learning that refers to transductive learning to infer correct labels for unlabeled data. Two of the clustering algorithm are discussed by (Pelckmans, 2020) and summarized that both could be most profound in different scenarios. A mixed technology utilization technique has been observed in (Jonker et al., 2017), (Camino et al., 2017) work, where both machine learning and data analytics are used to gain the advantage of both technologies. Other analytical approaches filter, refine, and aggregate the enormous dataset to achieve the best possible truth results. This High-rank dataset is then used as input to classifiers. While (Tai & Kan, 2019) work in inverse order, their classifiers are used before the data analytics algorithm. During the last 5 years, there are various studies in which big data analytics has been significantly used as an alternative to machine learning classifications, with a point of view that data analytics can bring more accuracy to the desired results. A fraudulent account detection study (Adedoyin et al., 2017) used analytics to detect suspicious accounts. Similarly (Li et al., 2017) showed analytic considerable results on real-world banking datasets while using AML patterns to identify the novel community. Analytic has also been successfully used in the detection and mutation of case topologies (Plaksiy et al., 2018). Money laundering topologies using a graphing approach are considered and analytic and coding techniques are then performed to generate similar variants of the sub-graph.

3. MATERIAL AND METHODS

Even before the programmable computers were introduced, people wondered if these machines can become faster and more intelligent than humans (Lovelace, 1842). Today's era of information bombardment and data is vast that the human brain could easily take decades to comprehend the information. Artificial Intelligence (AI) is thriving because of various practical applications and research areas. Challenge in AI is to address and solve problems that people can perform quickly, but usually, it is hard for them to describe like image recognition or pattern generation. AI adheres to a solution that allows computers to learn from the experience and environment in terms of a hierarchy of concepts. These hierarchical concepts leverage computers to learn complex scenarios. If it can draw a graph showing these concepts and how they are built in a hierarchy on top of each other, it will form a deep graph with multiple layers, and that is why this concept is called deep learning (Bengio, 2009).

Formally Deep learning(Bengio, 2009) is a subset of machine learning and a function of Artificial intelligence (AI) that can simulate the working of the human brain in processing data, identifying/detecting objects and making decisions. Supervised and unsupervised learning are the two main algorithm segregationists in Machine learning. As the name suggests, supervised learning algorithms used the labelled data set to train and define patterns based on it. While unsupervised learning doesn't predict accurate output, it inspects input data and generates patterns based on inspection to describe hidden structures of unlabeled data. There is one more function named Semi-Supervised learning, which attempts to take a middle ground by using both labelled and unlabeled data for training.

The problem to be solved in digital payments is to detect the potential fraud in the system related to a graph network. Employing machine learning algorithms for this purpose, which can process all transactions happening in the system, align patterns and detect potential fraud cases. Graphical data analytics has raised an important tool in the recent past, and AML is also a cash flow relationship between the parties, formulating a network of nodes and edges. Where a node/vertex represents an account, and an edge represents any correspondence between two accounts as nodes. These correspondences/links can be a single transaction or a group of transactions volume within a given time.

Further on, this method can also be attached to public sources to enrich its information for decision-making. In traditional rule-based systems, various techniques are used to identify the entity, and algorithms like cycle detection, degree of distribution, and ego-net are used to determine the anomalies. (Jingguang Han et al., 2018) produced a work that used natural language processing (NLP) for unstructured data combined from multiple sources and produces a suspicious level score. This work is done on data streams like news, financial reporting, social media, Twitter, etc. This work is convincing but needs a more refined graph analytical method to work effectively on a larger scale. Whereas Graph learning is used deep neural networks and emerged as a promising technique to solve complex patterns and focus while solving challenges in the AML area.

AMLSim is a purpose build multi-agent-based simulation of financial transactions with an ability to indulge money laundering patterns, topologies, and account behaviour. It can make a replica of real-world scenarios (perform both legal and illegal activities by virtual agents) – where each agent represents a bank account. Data simulation is two steps. First, a graph network of accounts is created using – the NetworkX library – accounts are connected through transactions. In the second step, the time series of transactions is appended using PaySim (E. Lopez-Rojas, Elmir, &Axelsson, 2016), i.e., transaction timestamps, accounts, and transactions are precisely synchronously generated. Three input files are needed for simulation, as shown in Figure 4. As an output, the transaction log is generated containing money laundering patterns and topologies.

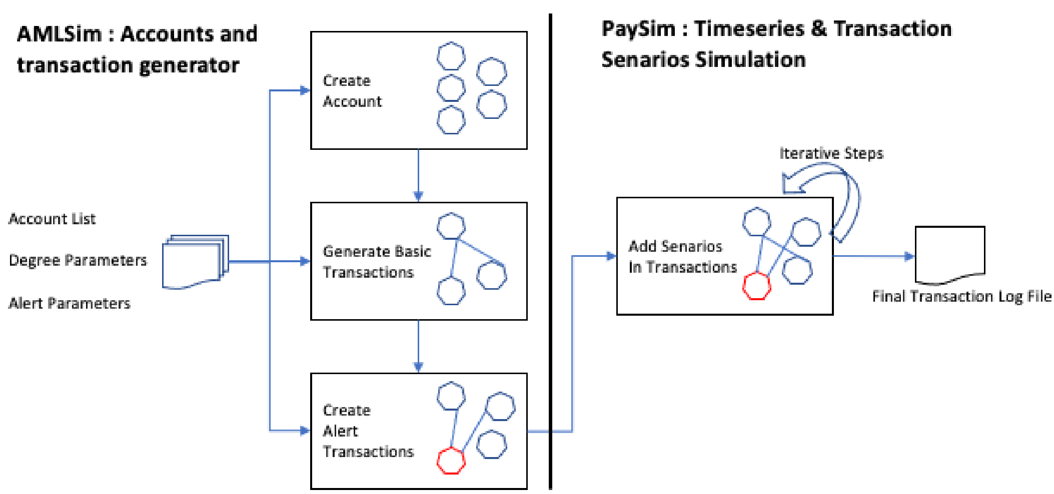
A semi-supervised machine learning method is proposed for identifying suspicious transactions and accounts in terms of money laundering. This work will open further new dimensions in solving AML challenges and reducing manual work.

- Graph analytical model is used directly on the transactions containing legitimate and illegal scenarios (Deng et al., 2009).
- The Standard supervised AML method demands and assumed that the experts mark suspicious activity data with some label to help the learning process. In contrast, the semi-supervised method provides the power of neural networks and their prediction capability in an unforeseen scenario.
- AML literature (Ngai, Hu, Wong, Chen, & Sun, 2011) is very limited and restricted in data sources and modelling frameworks; the authors simulated customer data, transaction information, and history.

3.1. Data Sources and Refinement

Data related to this field is very much restricted to access to relevant information because customer financial data is protected by law and regulation authorities. An alternate way to this privacy situation is to generate a synthetic dataset using a simulator. Synthetic data simulation is an urbane substitute to represent real-world scenarios and behavioural data to examine (E. Lopez-Rojas et al., 2016). Simulation brings an advantage in the sense to adhere to different conditions and scenarios with the possibility to validate the model's output. PaySim and AMLSim simulators by IBM & MIT are used to generate the required patterns of data (Borrajo, Veloso, & Shah, 2018)to support research work. It provides the behaviour of bank transactions, where X amount transfers from one account to

Figure 4. The Logical structure of data processing



another with single and multiple agents/accounts. The data generated represents a real-world banking transaction pattern, where agents conducted activities, and from those activities, various activities are regular, and some can be wicked.

Data generating steps are:

- Generate a list of accounts with a defined degree of distribution.
- Generate transactions based on distribution and real-world dynamics.
- Generate transaction data in graph dataset format for the listed accounts.

This procedure allows for generating dynamic graph data on a massive scale with a list of semi-realistic suspicious activities. The simulator is undoubtedly helpful for research purposes (scalability, graph data analytics, etc.), but meaningful measures must require real data.

During this experiment, the AML graph dataset was generated with a limited configuration, i.e., 100K unique nodes corresponding to 5.3M edges based on the configured degree of accounts/agents. Each node/vertex represents an account with attributes like account number, type, owner, creation date, etc. The edges represent the activity between the vertexes with the attribute's transaction ID, mode of transfer, and amount date (Table 4). The data is labelled sparsely with flagged transactions based on volume, velocity, and SARs patterns.

4. RESULT

Machine and deep learning produce remarkable contributions and developed amazing results on data, especially prediction models related to unstructured data such as images, voices, and videos. Deep learning graph-structured data is a field that is still maturing, covering the scalability and complexity challenges.

In the recent past significance of this area has become one of the hot topics for researchers. In recent times graph data modelling got a lot of focus as an emerging area of research, and some brilliant contributions have been added, one of which is graph neural network models. To demonstrate work, that selected graph convolution networks (GCN) (Kipf& Welling, 2016)(Figure 5) method with semi-supervised learning and classification.

AML network always forms a high-density, high-dimensional graph involving many nodes (accounts) and their edges (connections/transactions), where one node can represent one account or a full pattern itself (*i.e., a node representing a graph within a graph*) same way edges can be one transaction or aggregation of multiple transactions.

A common architecture has been used in almost all graph neural network models; referring to Graph Convolutional Networks, it provides a tendency to learn the features of a graph

$\mathcal{G} = \left(\mathcal{V},\, \mathcal{E}\right)$; with N vertices $v_i \in \mathcal{V}$ and edges $(v_i, v_j) \in \mathcal{E}$, by taking inputs

Table 4. Transaction record structure after final compilation

TX_ID	SENDER_A	RECEIVER	TX_TYPE	TX_AMOUNT	TIMESTAMP	IS_FRAUD	ALERT_ID
326	5148	2083	TRANSFER	19.79	0	FALSE	-1
327	4519	2765	TRANSFER	531.78	0	FALSE	-1
328	4721	9971	TRANSFER	190.0	0	FALSE	-1

- An adjacency matrix $A \in \mathbb{R}^{NxN}$ generated from the input graph object, it can be binary or weighted with a degree matrix $D_{ij} = \sum_i A_{ij}$
- A feature matrix $X = \left(\mathcal{N} \times \mathcal{F}\right)$; a feature vector x_i for each node i (where N is the number of nodes/vertices and F is the feature vector). This model also provides node-level outputs that can be further modelled using some pooling operations (Duvenaud et al., 2015).

Eventually, the neural network layer can be written as

$$H^{l+1} = f(H^{l}\mathrm{A});$$

and this whole architecture may be summarized in the following function

$$f(H^{l}\mathrm{A}) = \sigma(\hat{A}H^{l}W^{l})$$

Where Â is the normalized adjacency matrix (the square matrix is used to represent a finite graph), H^l contains graph vertices in the l^{th} layer, W^l is a feature matrix, and σ represents activation function (nonlinear) like Rectified linear unit (ReLU).

It's a pictorial depiction of multilayer GCN (Kipf & Welling, 2016) for semi-supervised learning. C represents the input channel, F is the feature vector map, and labels are donated by Y.

To experiment with this model, that used AMLsim for synthetic dataset creation as Appendix1 (Table 8), consists of 100 K accounts classified into k classes (*in the experimental case, these are two*). A graph of 100K vertexes with 5.4 M edges is formed (Table 5). Each node represents an account along with its attributes and each edge is a transaction between two accounts and their attributes. The data is sparsely labelled and distributed into three subsets. A subset of 20K nodes and edges-related areas is used as a training set, 15K nodes for validation, and around 50K nodes data are used as a test set.

A similar experiment has been taken (Table 6), where the author has chosen four different ML techniques such as Decision tree (DT); Conditional inference tree (CT); Random forest (RF); Neural network (NN) to run on the same simulated dataset (Visser &Yazdiha, 2020). All these techniques are popular classifiers and can be linearized into decisions. But need continuous improvement to handle new inventions and detection of newly emerged patterns, as shown in Table 6.

The learning, evaluation, and prediction of this semi-supervised prediction model depend on the feature vectors associated with nodes. Each node in the dataset is described by a 0/1 valued feature vector indicating the presence/absence of the corresponding feature. That helps the model predict

Figure 5. Graph Convolution Network

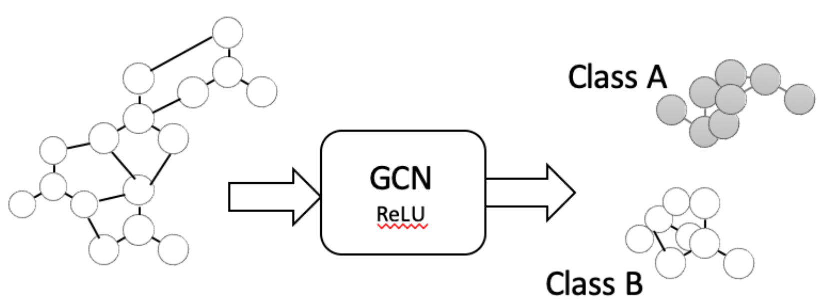


Table 5. Dataset Statistics

Dataset	AML
Type	Financial
Nodes	99,336
Edges	5,385,128
Classes	2
Features	Degree matrix
Label rate	0.151

Table 6. F1-score and accuracy comparison between different machine learning models

Metric	DT	CT	RF	NN
F1-Score	0.423	0.205	0.524	0.414
	0.374	0.364	0.693	0.479
	0.352	0.187	0.512	0.354
	0.439	0.337	0.619	0.459
Accuracy	0.637	0.557	0.678	0.693
	0.616	0.613	0.767	0.742
	0.608	0.552	0.673	0.664
	0.643	0.602	0.725	0.693

the target vertex's suspiciousness and identify the potentially bad transactions associated with it. Feature extraction is a vital component in graph learning and requires a significant level of AML domain knowledge. For this experiment, used a degree matrix as a feature vector. A degree matrix is a diagonal matrix that contains information about the degree of each vertex—that is, the number of edges attached to each vertex. The multi-class and feature vectors give more flexibility to the model, as the key discriminators may not be the same when attempting to predict suspicious nodes and associated activities. It has executed the model in many cycles of time on the given datasets, and the average of the results gathered is as follows (Figure 6).

4.1 AML Test Result

Graph learning on Synthetic AML dataset (100 K nodes, 5.4 M edges)

Workstation: Intel(R) 2.20 GHz, core i7, 16 GB DDR3 1600 MHz memory

The term *Accuracy* (Table 7) indicates the overall performance of ML algorithms (e.g., GCN) by specifying the percentage of instances that conform to rules based on a feature matrix, where one instance contains one specific condition and corresponding system state.

The enrichment of the feature matrix impacts the accuracy. For this experiment, used a degree matrix as a feature matrix W. Here it showed the results with 100K nodes, but in real-time scenarios, it can be millions/billions of nodes, and that amount of scale requires time and resources appropriately to train the model. The model's efficiency is very important because financial institutes/banks manage millions of transactions every hour and need to identify suspicious patterns and activities as quickly as possible. The empirical effort shows that graph-based machine learning with a financial dataset is promising, and significant results can be achieved with an enriched feature matrix.

Figure 6. Machine learning GCN Model Execution

```
Epoch: 0016 train_loss= 0.69411 train_acc= 0.76365 val_loss= 0.69463 val_acc= 0.72971 time= 0.12093
Epoch: 0017 train_loss= 0.69435 train_acc= 0.76210 val_loss= 0.69458 val_acc= 0.72538 time= 0.11671
Epoch: 0018 train_loss= 0.69431 train_acc= 0.75230 val_loss= 0.69426 val_acc= 0.71965 time= 0.12277
Epoch: 0019 train_loss= 0.69399 train_acc= 0.75205 val_loss= 0.69407 val_acc= 0.71418 time= 0.11722
Epoch: 0020 train_loss= 0.69375 train_acc= 0.76435 val_loss= 0.69418 val_acc= 0.71651 time= 0.11567
Epoch: 0021 train_loss= 0.69385 train_acc= 0.76335 val_loss= 0.69433 val_acc= 0.72058 time= 0.12251
Epoch: 0022 train_loss= 0.69400 train_acc= 0.75930 val_loss= 0.69426 val_acc= 0.72511 time= 0.12846
Epoch: 0023 train_loss= 0.69392 train_acc= 0.75365 val_loss= 0.69411 val_acc= 0.72331 time= 0.11751
Epoch: 0024 train_loss= 0.69378 train_acc= 0.75890 val_loss= 0.69409 val_acc= 0.72238 time= 0.13151
Epoch: 0025 train_loss= 0.69372 train_acc= 0.75765 val_loss= 0.69417 val_acc= 0.71971 time= 0.11744
Epoch: 0026 train_loss= 0.69373 train_acc= 0.77000 val_loss= 0.69420 val_acc= 0.72025 time= 0.11710
Epoch: 0027 train_loss= 0.69378 train_acc= 0.76075 val_loss= 0.69416 val_acc= 0.71818 time= 0.12539
Epoch: 0028 train_loss= 0.69372 train_acc= 0.75930 val_loss= 0.69414 val_acc= 0.71858 time= 0.11595
Epoch: 0029 train_loss= 0.69366 train_acc= 0.76645 val_loss= 0.69418 val_acc= 0.71931 time= 0.12472
Early stopping...
Optimization Finished!
Test set results: cost= 0.69400 accuracy= 0.79101 time= 0.04625
```

Table 7. The output of test executions

Test	Accuracy	Time
1	0.79368	0.64259
2	0.79341	0.64247
3	0.77961	0.64877
4	0.78993	0.63951
5	0.79771	0.64674
6	0.79385	0.65057
7	0.78964	0.64416
8	0.79762	0.64578
9	0.79963	0.64281
10	0.79112	0.63975

5. DISCUSSION

There are multiple types of technologies; two are widely used with various further evaluation measures to solve different anti-money laundering scenarios. Table 1 shows the details of the models and sub-evaluation measures used. Big data analysis and machine learning both have their advantages and gaps. The authors observed that both machine learning and data analytics were implemented to gain the advantage of both technologies. On the other hand, Machine learning in many studies implemented a multilayered approach with different evaluation and categorization measures. The most categorization method used is SVM and LR DT. All these algorithms are either linear or tree base methods. Underneath the dataset in approximately all research used is graph data. Either dataset is formulated or converted in the graph format to gain the advantage of relation, connection, and clustering. A comparison is taken between the popular machine learning evaluation measures (Table 6 & 7) with graph base ML approach.

During this experiment degree of the nodes in the adjacency matrix was used as a node feature vector. Discovered comparatively better results, but significant improvements can be achieved by enriching node feature vector with mode details like introducing KYC - customer profiling, tracking record of customer status in the society, criminal history record, engagement in politics, etc.

5.1 Limitations

The core limitation in this field of study is the availability of the real financial dataset and limited literature discussing real-time scenarios, bank account pro-filings, and solutions.

5.2 Research Question

How can a graph-based machine learning model achieve an effective and timely reaction to a criminal financial investigation?

5.3 Answer

Fifty (50+) pieces of research were carried out between 2017- (and Jan) 2021 based on the inclusion and exclusion criteria mentioned. Results showed that various methods are used with machine learning with data analytics. Linear regression, Rule-based regression, Vector machine and decision trees base classification measures. But none of the considerable input found on graph-structured machine learning modelling, which in theory handles N number of feature set on the node, and N number of nodes can be effectively used. A node can be an account with a weight of transactions generated to and from this account and this can even be a sub-graph that represents a specific pattern and can be learned by a graph-based machine learning model. This research appended GCN machine learning algorithms and showed the significance of results in comparison to other classifiers and a substantial potential for future research in this direction.

6. CONCLUSION

The motivation behind this study was to find an effective and flexible way to solve the anti-money laundering problem. Money laundering is a challenge to society's well-being. This organized crime affects countries' economies and produces a negative image of financial institutes at an international level, e.g., inclusion in the list of countries with weak financial action task force (FATF) compliances. It is the urge research community, financial institutes, and regulatory authorities to work side by side to employ machine learning technologies to fight against money laundering. This empirical study has reviewed the AML process and current widely used classification methods experimented on over a graph-based machine learning model for financial institutes and shared preliminary work using the synthetic dataset to identify fraudulent activities. The results are promising, and future research is encouraged to move further in this direction e.g., another discrete approach that becomes possible with graph structures is to train the graph-based model for topological patterns detection, i.e., using sub-graph (real reported case patterns) as nodes. The present research aimed to explore artificial intelligence approaches that are used to solve anti-money laundering problems and weigh graph-based machine learning algorithms to solve a similar problem.

ACKNOWLEDGMENT

The author would like to acknowledge supervisors for their support, suggestions, and reviews while conducting this study and thankful to them for the successful completion of this part of the research.

CONFLICT OF INTEREST

The authors of this publication declare there is no conflict of interest.

FUNDING AGENCY

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

REFERENCES

Adedoyin, A., Kapetanakis, S., Samakovitis, G., & Petridis, M. (2017). *Predicting fraud in mobile money transfer using case-based reasoning.* Paper presented at the International Conference on Innovative Techniques and Applications of Artificial Intelligence. doi:10.1007/978-3-319-71078-5_28

Alexandre, C., & Balsa, J. (2015). *Client profiling for an anti-money laundering system.* arXiv preprint arXiv:1510.00878.

Bank, A. D. (2003). *Countering Money Laundering in the Asian and Pacific Region*. Academic Press.

Bank, D. (2020). *Translation Embeddings for Knowledge Graph Completion in Consumer Banking Sector.* Paper presented at the Artificial Intelligence. IJCAI 2019 International Workshops, Macao, China.

Banks, J. (2017). *Online gambling and crime: Causes, controls and controversies*. Routledge. doi:10.1057/978-1-137-57994-2

Bengio, Y. (2009). *Learning deep architectures for AI*. Now Publishers Inc. doi:10.1561/9781601982957

Borrajo, D., Veloso, M., & Shah, S. (2018). Simulating and classifying behavior in adversarial environments based on action-state traces: An application to money laundering. *arXiv preprint arXiv:2011.01826*.

Camino, R. D., State, R., Montero, L., & Valtchev, P. (2017). *Finding suspicious activities in financial transactions and distributed ledgers.* Paper presented at the 2017 IEEE International Conference on Data Mining Workshops (ICDMW). doi:10.1109/ICDMW.2017.109

Chen, T.-H. (2020). Do you know your customer? Bank risk assessment based on machine learning. *Applied Soft Computing*, *86*, 105779. doi:10.1016/j.asoc.2019.105779

Choo, K.-K. R. (2015). *Cryptocurrency and virtual currency: Corruption and money laundering/terrorism financing risks? In Handbook of digital currency*. Elsevier. doi:10.1016/B978-0-12-802117-0.00015-1

Comission, U. S. E. (2018). *Anti-Money Laundering (AML)*. Source Tool for Broker-Dealers.

Deng, X., Joseph, V. R., Sudjianto, A., & Wu, C. J. (2009). Active learning through sequential design, with applications to detection of money laundering. *Journal of the American Statistical Association*, *104*(487), 969–981. doi:10.1198/jasa.2009.ap07625

Duvenaud, D. K., Maclaurin, D., Iparraguirre, J., Bombarell, R., Hirzel, T., Aspuru-Guzik, A., & Adams, R. P. (2015). *Convolutional networks on graphs for learning molecular fingerprints.* Paper presented at the Advances in neural information processing systems.

El-Banna, M. M., Khafagy, M. H., & El Kadi, H. M. (2020). *Smurf Detector: a Detection technique of criminal entities involved in Money Laundering.* Paper presented at the 2020 International Conference on Innovative Trends in Communication and Computer Engineering (ITCE). doi:10.1109/ITCE48509.2020.9047811

Elgendy, N., & Elragal, A. (2014). *Big data analytics: a literature review paper.* Paper presented at the Industrial conference on data mining. doi:10.1007/978-3-319-08976-8_16

Feng, Y., Li, C., Wang, Y., Wang, J., Zhang, G., Xing, C., . . . Lian, Z. (2019). *Anti-money Laundering (AML) Research: A System for Identification and Multi-classification.* Paper presented at the International Conference on Web Information Systems and Applications. doi:10.1007/978-3-030-30952-7_19

Gao, Z., & Ye, M. (2007). A framework for data mining-based anti-money laundering research. *Journal of Money Laundering Control*.

Gee, S. (2014). *Fraud and fraud detection: A data analytics approach*. John Wiley & Sons. doi:10.1002/9781118936764

Grint, R., & O'Driscoll. (2017). *New technologies and anti-money laundering compliance*. PA Consulting Group

Guevara, J., Garcia-Bedoya, O., & Granados, O. (2020). *Machine Learning Methodologies Against Money Laundering in Non-Banking Correspondents.* Paper presented at the International Conference on Applied Informatics. doi:10.1007/978-3-030-61702-8_6

Hamid, O. H. (2017). *Breaking through opacity: A context-aware data-driven conceptual design for a predictive anti money laundering system.* Paper presented at the 2017 9th IEEE-GCC Conference and Exhibition (GCCCE). doi:10.1109/IEEEGCC.2017.8448084

Han, J., Barman, U., Hayes, J., Du, J., Burgin, E., & Wan, D. (2018). Nextgen aml: Distributed deep learning based language technologies to augment anti money laundering investigation. *Proceedings of ACL 2018, System Demonstrations*. doi:10.18653/v1/P18-4007

Han, J., Kamber, M., & Pei, J. (2012). Data mining: Concepts and techniques. Morgan Kaufman Publishers.

Hanbar, H., Shukla, V., Modi, C., & Vyjayanthi, C. (2019). *Optimizing e-KYC Process Using Distributed Ledger Technology and Smart Contracts.* Paper presented at the International Conference on Computational Intelligence, Security and Internet of Things.

He, S., & Qu, Z. (2019). Research on the Periodical Behavior Discovery of Funds in Anti-money Laundering Investigation. *Proceedings of the 2019 11th International Conference on Machine Learning and Computing*. doi:10.1145/3318299.3318356

IMF. (2020). Anti-Money Laundering/Combating the Financing of Terrorism (AML/CFT) (Vol. 04 01). International Monetary Fund.

Jin, Y., & Qu, Z. (2018). *Research on Anti-Money Laundering Hierarchical Model.* Paper presented at the 2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS). doi:10.1109/ICSESS.2018.8663895

Jonker, C., Habeck, T., Park, Y., Jordens, F., & van Schaik, R. (2017). *Graph Analytics for Real-Time Scoring of Cross-Channel Transactional Fraud.* Paper presented at the Financial Cryptography and Data Security: 20th International Conference, FC 2016, Christ Church, Barbados.

Khanuja, H. K., & Adane, D. (2018). *Detection of suspicious transactions with database forensics and theory of evidence.* Paper presented at the International Symposium on Security in Computing and Communication.

Kipf, T. N., & Welling, M. (2016). *Semi-supervised classification with graph convolutional networks.* arXiv preprint arXiv:1609.02907.

Kittrie, O. F. (2016). *Lawfare: Law as a weapon of war*. Oxford University Press. doi:10.1093/acprof:oso/9780190263577.001.0001

Kossovsky, N., Greenberg, M. D., & Brandegee, R. C. (2012). *Reputation, stock price, and you: Why the market rewards some companies and punishes others*. Springer. doi:10.1007/978-1-4302-4891-0

Li, X., Cao, X., Qiu, X., Zhao, J., & Zheng, J. (2017). *Intelligent anti-money laundering solution based upon novel community detection in massive transaction networks on spark.* Paper presented at the 2017 fifth international conference on advanced cloud and big data (CBD). doi:10.1109/CBD.2017.38

Liu, X., Zhang, P., & Zeng, D. (2008). *Sequence matching for suspicious activity detection in anti-money laundering*. Paper presented at the International Conference on Intelligence and Security Informatics. doi:10.1007/978-3-540-69304-8_6

Lopez-Rojas, E., Elmir, A., & Axelsson, S. (2016). *PaySim: A financial mobile money simulator for fraud detection.* Paper presented at the 28th European Modeling and Simulation Symposium, EMSS, Larnaca.

Lopez-Rojas, E. A., & Axelsson, S. (2012). *Money laundering detection using synthetic data.* Paper presented at the Annual workshop of the Swedish Artificial Intelligence Society (SAIS).

Lovelace, A. (1842). *Sketch of the analytical engine invented by Charles Babbage*. Academic Press.

Luna, D. K., Palshikar, G. K., Apte, M., & Bhattacharya, A. (2018). Finding shell company accounts using anomaly detection. *Proceedings of the ACM India Joint International Conference on Data Science and Management of Data*. doi:10.1145/3152494.3152519

Manning, M., Wong, G. T., & Jevtovic, N. (2020). Investigating the relationships between FATF recommendation compliance, regulatory affiliations and the Basel Anti-Money Laundering Index. *Security Journal*.

Mocko, M., & Ševcech, J. (2018). *Simulation of Bank Transaction Data.* Paper presented at the International Workshop on Multi-Agent Systems and Agent-Based Simulation.

Ngai, E. W., Hu, Y., Wong, Y. H., Chen, Y., & Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision Support Systems*, *50*(3), 559–569. doi:10.1016/j.dss.2010.08.006

Omar, N., Johari, Z. A., & Arshad, R. (2014). Money laundering–FATF special recommendation VIII: A review of evaluation reports. *Procedia: Social and Behavioral Sciences*, *145*, 211–225. doi:10.1016/j.sbspro.2014.06.029

Pambudi, B. N., Hidayah, I., & Fauziati, S. (2019). *Improving Money Laundering Detection Using Optimized Support Vector Machine.* Paper presented at the 2019 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI). doi:10.1109/ISRITI48646.2019.9034655

Pelckmans, K. (2020). Monitoring High-Frequency Data Streams in FinTech: FADO Versus $ K $ K-Means. *IEEE Intelligent Systems*, *35*(2), 36–42. doi:10.1109/MIS.2020.2977012

Pietschmann, T., & Walker, J. (2011). *Estimating illicit financial flows resulting from drug tracking and other transnational organized crimes*. United Nations.

Plaksiy, K., Nikiforov, A., & Miloslavskaya, N. (2018). *Applying big data technologies to detect cases of money laundering and counter financing of terrorism.* Paper presented at the 2018 6th International Conference on Future Internet of Things and Cloud Workshops (FiCloudW). doi:10.1109/W-FiCloud.2018.00017

Plaksiy, K., Nikiforov, A., & Miloslavskaya, N. (2019). *Big Data Analytics for Financial Crime Typologies.* Paper presented at the International Conference on Big Data Innovations and Applications. doi:10.1007/978-3-030-27355-2_13

Prakash, A., Apoorva, S., Amulya, K., Kavya, T., & KN, P. K. (2019). *Proposal of expert system to predict financial frauds using data mining.* Paper presented at the 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC).

Qureshi, W. A. (2017). An Overview of Money Laundering in Pakistan and Worldwide: Causes, Methods, and Socioeconomic Effects. *U. Bologna L. Rev.*, *2*, 302.

Rates, F. P. D., Guides, C. C., Center, U. D., Clinton, S. H. R., & Hotline, I. G. (2017). *Money Laundering and Financial Crimes*. US Department of State.

Rothchild, J. A. (2016). *Research Handbook on Electronic Commerce Law*. Edward Elgar Publishing. doi:10.4337/9781783479924

Schott, P. A. (2006). *Reference guide to anti-money laundering and combating the financing of terrorism*. The World Bank.

Seifi, S. T., & Ekhveh, A. A. (2019). *Representing Unequal Data Series in Vector Space with Its Application in Bank Customer Clustering.* Paper presented at the International Congress on High-Performance Computing and Big Data Analysis.

Shelley, L. I. (2014). *Dirty entanglements: Corruption, crime, and terrorism*. Cambridge University Press. doi:10.1017/CBO9781139059039

Tai, C.-H., & Kan, T.-J. (2019). *Identifying Money Laundering Accounts.* Paper presented at the 2019 International Conference on System Science and Engineering (ICSSE). doi:10.1109/ICSSE.2019.8823264

Tertychnyi, P., Slobozhan, I., Ollikainen, M., & Dumas, M. (2020). *Scalable and Imbalance-Resistant Machine Learning Models for Anti-money Laundering: A Two-Layered Approach.* Paper presented at the International Workshop on Enterprise Applications, Markets and Services in the Finance Industry. doi:10.1007/978-3-030-64466-6_3

Truman, E. M., & Reuter, P. (2004). *Chasing Dirty Money: The Fight Against Anti-Money Laundering*. Academic Press.

Unger, B., & Van der Linde, D. (2013). *Research handbook on money laundering*. Edward Elgar Publishing. doi:10.4337/9780857934000

Visser, F., &Yazdiha, A. (2020). *Detection of Money Laundering Transaction Network Structures and Typologies using Machine Learning Techniques*. Academic Press.

APPENDIX

Table 8. Synthetic AML dataset by AMLsim

step	type	amount	nameOrig	oldbalanceOrg	newbalanceOrig	nameDest	oldbalanceDest	newbalanceDest	isFraud
1	CHECK	78.0	1873	0.0	0.0	87708	0.0	78.0	0
1	DEPOSIT	78.0	1873	0.0	0.0	99423	0.0	78.0	0
1	CHECK	199.05	388	0.0	0.0	28098	0.0	199.05	0
1	CASH-OUT	52.7	0	0.0	0.0	388	0.0	52.7	0
1	WIRE	150.38	1013	0.0	0.0	48144	0.0	150.38	0
1	CREDIT	134.02	356	0.0	0.0	99858	0.0	134.02	0
1	CHECK	173.25	835	0.0	0.0	9834	0.0	173.25	0
1	DEPOSIT	98.95	1821	0.0	0.0	99517	0.0	98.95	0
1	CHECK	98.95	1821	0.0	0.0	25150	0.0	98.95	0
1	DEPOSIT	112.08	600	0.0	0.0	84551	0.0	112.08	0
1	CREDIT	94.38	1632	0.0	0.0	79879	0.0	94.38	0
1	CREDIT	94.38	1632	0.0	0.0	19064	0.0	94.38	0
1	CREDIT	108.05	1207	0.0	0.0	48531	0.0	108.05	0
1	CHECK	185.99	407	0.0	0.0	32437	0.0	185.99	0
1	DEPOSIT	78.29	1518	0.0	0.0	48120	0.0	78.29	0
1	DEPOSIT	78.29	1518	0.0	0.0	79979	0.0	78.29	0
1	CREDIT	146.72	443	0.0	0.0	95615	0.0	146.72	0
1	CASH-IN	82.21	443	0.0	0.0	0	0.0	82.21	0
1	CHECK	191.63	979	0.0	0.0	66904	0.0	191.63	0
1	WIRE	182.92	772	0.0	0.0	37449	0.0	182.92	0
1	DEPOSIT	75.4	1580	0.0	0.0	29429	0.0	75.4	0
1	CREDIT	75.4	1580	0.0	0.0	99996	0.0	75.4	0
1	CHECK	158.61	975	0.0	0.0	38980	0.0	158.61	0
1	CHECK	138.42	1489	0.0	0.0	99995	0.0	138.42	0
1	DEPOSIT	62.04	1810	0.0	0.0	88353	0.0	62.04	0
1	CREDIT	169.89	584	0.0	0.0	39499	0.0	169.89	0
1	CREDIT	124.9	618	0.0	0.0	39940	0.0	124.9	0
1	DEPOSIT	98.83	1517	0.0	0.0	99056	0.0	98.83	0
1	DEPOSIT	98.83	1517	0.0	0.0	38049	0.0	98.83	0
1	CHECK	106.87	1442	0.0	0.0	89545	0.0	106.87	0
1	CHECK	71.9	1669	0.0	0.0	69940	0.0	71.9	0
1	DEPOSIT	71.9	1669	0.0	0.0	39791	0.0	71.9	0
1	CASH-OUT	53.82	0	0.0	0.0	1669	0.0	53.82	0
1	CHECK	115.94	1438	0.0	0.0	18423	0.0	115.94	0
1	WIRE	98.1	1734	0.0	0.0	38097	0.0	98.1	0
1	CHECK	98.1	1734	0.0	0.0	30930	0.0	98.1	0
1	CREDIT	115.13	431	0.0	0.0	27896	0.0	115.13	0
1	DEPOSIT	99.37	1878	0.0	0.0	39286	0.0	99.37	0
1	CREDIT	99.37	1878	0.0	0.0	40554	0.0	99.37	0

Atif Usman received a Master's degree in business and information technology management from International Islamic University, Pakistan, in 2005. Currently a MS student in the Department of Computer Sciences at VU, Pakistan, with a focused area of research in AI applications and Machine learning. In parallel, he has worked extensively in various areas of industry, and currently appointed at a senior position "Senior Technology Advisor" at Den Norke Bank, Norway.

Nasir Naveed has been working as an Associate Professor in the department of Computer Science and Information Technology, Virtual University of Pakistan, Lahore. He graduated from the Institute of Web Science and Technologies, University of Koblenz, Germany. Currently, he is investigating the use of machine learning techniques for exploiting the text contents for search in Linked Open Data and developing scalable methods for Big Data analysis and Data Science. In past, he worked on the temporal analysis of social media contents in popular social networks using machine learning and information retrieval techniques.

Saima Munawar has been working as an Assistant Professor in the department of computer science and Information Technology, at the Virtual University of Pakistan, Lahore. She graduated from the National College of Business Administration and Economics, Lahore. She has several publications on various technical subjects in international and national HEC-recognized journals. Her current research interests are Artificial Intelligence and Education Technology.