Universidade do Minho
Escola de Engenharia

Ana Paula Fernandes Martins

Fraud Detection and Anti-Money Laundering applying Machine Learning techniques in Cryptocurrencies Transactional Graphs

February 2023

Fraud Detection and Anti-Money Laundering applying Machine Learning techniques in Cryptocurrencies Transactional Graphs

Ana Paula Fernandes Martins

UMinho | 2023

Universidade do Minho
Escola de Engenharia

Ana Paula Fernandes Martins

Fraud Detection and Anti-Money Laundering applying Machine Learning techniques in Cryptocurrencies Transactional Graphs

Intermediate Report Masters Dissertation
Integrated Master's in Engineering and Management of Information Systems

Work performed under the supervision of
Miguel Abrunhosa de Brito

February 2023

INDEX

LIST OF ABBREVIATIONS AND ACRONYMS

ML	Machine Learning
AI	Artificial Intelligence
AML	Anti-Money Laundering
GAD	Graph Anomaly Detection
LSTM	Long Short-Term Memory
SVM	Support Vector Machine
NLP	Natural Language Processing
UTXO	Unspent transaction output
GCN	Graph Convolutional Network
GNN	Graph neural Network
LR	Logistic Regression
MLP	Multilayer Perceptron
DSR	Design Science Research
RF	Random Forest

LIST OF FIGURES

1. INTRODUCTION

Cryptocurrencies have become increasingly popular in recent years, with some experts arguing that they may be among the most promising fintech innovations of the decade. Their use has been expanding and has been applied to a wide range of activities. Now, major companies are beginning to accept payments in cryptocurrencies as an alternative to traditional fiat currencies. This suggests that cryptocurrencies have the potential to fundamentally change market mechanisms and financial services (Eigelshoven et al., 2021).
As a result, research in this area is becoming increasingly important and has been receiving significant attention in recent years.
The problem and motivation that gave rise to this work, as well as the established objectives, will be discussed subsequently.

1.1 Problem and Motivation

Despite the decentralized nature of blockchain technology providing inherent security, the unregulated nature of the cryptocurrency market presents opportunities for criminals to exploit vulnerabilities and engage in illegal activities such as fraud and money laundering. With the growing adoption of cryptocurrencies, it is imperative for organizations to implement robust measures and strategies to detect and prevent such illicit activities.
This dissertation is sponsored by Uphold, a company that provides financial services through an online platform for the transaction of digital assets such as cryptocurrencies. This makes it a complex business with many players, where identifying anomalies and understanding user interactions is of great value. As a company that handles cryptocurrencies on a daily basis, Uphold is at risk of being used to commit illegal transactions which makes it vulnerable to illegal transactions that can compromise the trust of its users and put the company in a vulnerable position with regard to compliance with laws, regulations, and industry standards. The use of Anti-Money Laundering (AML) controls, which have been traditionally employed in the traditional financial system to detect money laundering transactions, is also being extended to the cryptocurrency market.

Though it is commonly believed that fraudulent transactions are easily distinguishable from normal transactions, this is not always the case. Criminals often mimic the behavior of legitimate transactions to conceal their activities, making them appear as normal transactions. Additionally, as criminals constantly devise new methods to carry out their illegal activities, it is essential to continuously develop and implement new and improved methods for detecting them.

Thus, the field of anti-money laundering (AML) is constantly evolving, with new methods and techniques being developed to stay ahead of financial criminals. One of the most promising areas of research in AML is the integration of machine learning (ML) algorithms to support traditional AML methods. ML techniques can be leveraged to extract patterns from data and identify suspicious activities, making it a valuable tool in this context.

These algorithms have the ability to adapt to changing market trends and improve the effectiveness of AML efforts. Despite the potential of this research, the application of ML in AML is still in its early stages, leaving ample room for further study and development.

The utilization of transactional data on blockchain technology provides new prospects for transaction network analysis. This can be achieved through the creation of visual representations of the transaction networks using transactional graphs. These graphs highlight valuable insights into the relationships between nodes, making them ideal for machine learning on graph data. However, traditional machine learning methods may have difficulty effectively handling the complex and non-linear connections present in graph data. Utilizing deep learning methods, such as neural networks, in combination with graph analysis can effectively identify complex relationships within the data and detect fraudulent activities in the cryptocurrency market. Additionally, deep learning methods can adapt and evolve with the market, making them an ideal choice for detecting fraud in a dynamic and ever-changing environment. In short, utilizing deep learning methods in combination with graph analysis is crucial for detecting fraudulent transactions in the cryptocurrency market, as it can effectively extract patterns and features from the data, even when they are hidden, and adapt to changing market trends.

1.2 Project Objectives

Considering the aforementioned problem, this dissertation aims to investigate the cryptocurrency market and detect patterns of fraudulent behavior using graph and ML techniques to gain insights that can be applied to the Uphold platform to detect and prevent illegal transactions, thus supporting compliance with regulations and risk management activities. Specifically, the objectives of this dissertation are to evaluate the feasibility of using ML and graph analysis to detect fraudulent transactions in cryptocurrencies and to compare the performance of traditional ML methods with that of graph-oriented deep learning methods to identify an effective mechanism for detecting and preventing fraud in transactions on the Uphold platform.

- To accomplish this goal, it is necessary to conduct research on key concepts related to this dissertation, such as fraud, and money laundering, and the measures companies must take to address them.
- Additionally, a literature review on the use of ML and graph analysis in combination with ML to detect fraudulent transactions must be conducted. The literature review should encompass the techniques and datasets that have been used in recent studies on these topics, as well as their conclusions and any insights that have been previously discovered about fraudulent behavior, such as anomalies and patterns that are commonly observed.
- Using the research as a foundation, the appropriate ML algorithms, and graph analysis techniques must be selected and applied to the transactional data.
- To do this, a representative dataset of Uphold's data must be obtained. The objective is to establish a baseline for comparison, and then apply the chosen method to the dataset.
- In the first phase, traditional ML methods will be used, followed by deep learning methods in the second phase. All tested techniques will then be evaluated and benchmarked to determine the most effective method in this context.
- Finally, there’s a possibility of applying the selected techniques to the real Uphold transactional data, to verify and compare its results with the test results.

1.3 Intermediate Report Objectives

In the initial stage, an intermediate report will be executed, which serves as this present document. In this phase, the defined objectives are the following:

- The problem and objectives for the dissertation must be clearly defined to accurately identify and articulate the issue that led to the research and to ensure that the formulated objectives are sufficient in addressing the problem.
- The state-of-the-art and the literature review must be relevant and appropriate to the current stage of development, and the identified problem ensuring that the review is comprehensive and effectively supports the research objectives.
- The research methodology, approach, or methods and techniques must be thoroughly explained, justified, and aligned with the research objectives, ensuring that the chosen design or approach is logically and appropriately suited to achieve the defined objectives.
- The identified research activities must be organized and planned effectively, ensuring that the plan is appropriate, feasible, and contains sufficient detail to effectively carry out the research and achieve the defined objectives.
- The potential risks that could impact the success of the research or hinder the satisfaction of its objectives must be identified, classified, and mitigation actions must be identified and evaluated for their effectiveness to minimize or eliminate those risks.
- The stages of the research must be aligned and consistent with the activities outlined in the initial work plan, ensuring that progress is being made as intended and that the research remains on track to achieve its defined objectives.

2. RESEARCH METHODOLOGY

In this dissertation, the investigation methodology employed is rooted in the principles of Design Science Research (DSR). DSR is an appropriate methodology for addressing practical issues and generating new knowledge and artifacts that can be utilized to enhance the design of existing systems or develop new ones. The ultimate outcome of DSR is frequently a new artifact, such as a tool, system, or method, that can be applied to improve the well-being of individuals or organizations. The specific problem addressed in this dissertation pertains to fraud detection in cryptocurrencies using graph techniques, with the research objective being to detect and prevent illicit transactions on the Uphold platform. The DSR methodology includes the following phases:

1. **Problem identification and formulation:** The initial phase of this research project involves the identification of a problem or opportunity that can be addressed through systematic investigation. This phase includes the definition of the problem, the comprehension of its context, and the formulation of clear and specific research objectives.
2. **Literature review:** In this phase, the researcher conducts a literature review in this phase to gain an understanding of existing solutions, identify gaps in current knowledge, and ensure that the proposed solution is novel. The literature review helps to establish the relevance and feasibility of the research and serves as a foundation for the design and execution of the research, providing a basis for comparison with the results obtained. It is an important step in the research process that helps to ensure the research is grounded in existing knowledge and contributes to the advancement of the field.
3. **Design and development:** This phase of the research process involves designing and developing a solution to the problem at hand. This includes the creation of a prototype or proof of concept, as well as the definition of the methods and techniques that will be utilized to evaluate the solution. The design and development of the solution is a critical step in the research process, as it helps to ensure that the proposed solution is practical, feasible, and effective.
4. **Evaluation:** This phase of the research process is dedicated to testing and evaluating the solution developed in the previous phase. The objective of this phase is to determine the effectiveness and potential impact of the solution, this can be achieved by conducting user

testing, experiments, or other forms of evaluation. The results of these tests and evaluations will provide valuable insight into the performance of the solution and will inform any necessary adjustments or improvements to the design.

5. **Dissemination and Implication:** This phase involves communicating the results of the research and its implications to the relevant stakeholders. This can include publishing the research in academic journals, presenting it at conferences, or creating a user manual.

2.1 Application of the Research Methodology

Following there's a description of this work's objectives that are related to each one of the DSR phases.

1. **Problem identification and formulation:** To ensure that the research is comprehensive and yields valuable and relevant outcomes, it is imperative to commence by clearly defining the problem that initiated the research and outlining the motivation for addressing it. Following that, is necessary to establish an objective, and a plan to solve the problem, which includes specific and measurable steps, should be laid out to achieve that objective. Furthermore, it is essential to conduct in-depth research on the context and related concepts of the topic to fully comprehend the field of exploration and be aware of the current state of the art. This involves understanding the concepts of cryptocurrencies, fraud, AML, and graph analysis.
2. **Literature review:** This phase implicates a thorough review of existing literature in the same field of experimentation to gain insight into previous solutions proposed and issues identified about the application of ML techniques and graph analysis in the field of AML. This includes identifying and describing relevant and related works that have been conducted in the past, as well as identifying areas that warrant further exploration in the current study.
3. **Design and development:** In this phase of the research, after conducting a thorough literature review, a methodology is proposed to attain the objectives of the study and to add to the existing body of knowledge on the utilization of ML techniques and graph analysis in the field of AML. This methodology includes an examination of the specific ML and graph analysis techniques that will be employed, the selection of an appropriate

dataset, and the establishment of a pipeline for implementation. Following the establishment of the methodology, it will be implemented.

4. **Evaluation:** The proposed methodology will undergo a thorough evaluation process, in which its results will be compared with those of a previous study that employed similar techniques. The comparison will be based on metrics that objectively measure the performance of the techniques on the identification of fraudulent transactions and will be used to determine the effectiveness of the proposed solution in comparison to the previous studies. This evaluation process aims to determine the superiority of the proposed methodology and its potential to contribute new knowledge to the field of AML.
5. **Dissemination and Implication:** In conclusion, the findings of the study will be presented in the form of a master's dissertation document at the university. Furthermore, an article featuring the key conclusions will be written and submitted for publication, as well as presented at an academic conference.

3. STATE OF ART

3.1 Cryptocurrencies

Cryptocurrencies are defined as a digital representation of value, issued by a private developer instead of a central bank, credit institute, or e-money institute and can be used as an "alternative form of money" (Bank, 2015).

Blockchain is the underlying technology behind cryptocurrencies. It is a decentralized, peer-to-peer network that utilizes various cryptographic techniques to create a secure, append-only database. The consensus protocol allows for the addition of verified and confirmed information to the blocks, which are then distributed to all participant nodes in the network (Tang et al., 2023). In this way, cryptocurrencies do not require a central authority to process and settle transactions, unlike traditional transactions. Instead, transactions are verified within a decentralized network through cryptographic techniques and the transaction data are then stored immutably on a distributed ledger (Gandal et al., 2018). This happens as the input of a signed transaction contains both the hash value of a previous transaction and an

index to locate the output since it is a reference to an output from a previous transaction. This guarantees that transactions are not easily changed or damaged (Tang et al., 2023).

Cryptocurrencies, due to their decentralized and peer-to-peer structure, offer several advantages over traditional monetary systems. These include lower transaction costs, increased transactional security, and privacy, and higher transaction efficiency (Rejeb et al., 2021). The transparency of cryptocurrency transactions is one of its key features, as all transactions are recorded on the publicly available blockchain (Trozze et al., 2022). Additionally, they offer a high degree of privacy as the transactions are not directly linked to a specific user or organization (Tang et al., 2023) being pseudo-anonymous in the sense that one might know the address of a user but cannot know exactly who he or she is (Yuan et al., 2018). Blockchain technology, by serving as a public ledger, enables the utilization of analytical tools for the purpose of searching for transactions associated with a specific address. This includes the capability to search and validate related transactions stored within a specific wallet address (Tang et al., 2023).

3.2 Fraud

Khrestina et al. (2017) define fraud as attempts to launder proceeds obtained illegally or first-party fraud or even fraudulent debit card charges. Zhou et al. (2021), on the other hand, define fraud generally as an illegal or criminal deception aimed at obtaining financial or personal benefits. As the adoption and usage of cryptocurrencies increases, it becomes crucial to comprehend and investigate the different types of fraud that occur within this system to combat and prevent it effectively.

While the immutability of transactions on the blockchain has the potential to greatly reduce fraud (Tang et al., 2023), the cryptocurrency market remains vulnerable to fraud and scams. This is due to its attractiveness to financial crime because of its lack of controls (Astrakhantseva et al., 2021) resulting from the pseudo-anonymity (Sabry et al., 2020), not existing a central third party to manage the cryptocurrency and regulate the market (Eigelshoven et al., 2021), and neither existing any jurisdiction to enforce the law (Astrakhantseva et al., 2021).

As well, according to Trautman (2014), criminals prefer virtual currencies because they offer the greatest degree of anonymity for both users and transactions, the ability to move illicit

proceeds quickly and confidently from one country to another, low volatility, widespread adoption, and trustworthiness.

Brenig et al. (2015) disagree with the notion that criminals prefer cryptocurrencies solely for their technical design features and instead argue that it is the contextual and transactional factors that make them attractive for money laundering. They found that instantaneous transactions, decentralization, and pseudo-anonymity are positive incentives for criminals, while limited acceptance and high price volatility are negative incentives.

The study by (Trozze et al., 2022) highlights that the cryptocurrency markets are vulnerable to various types of fraud and scams, which is a growing concern worldwide. The authors conducted a literature review to understand the types of fraud that have already been identified and those that might occur in the future. They identified 29 kinds of fraud schemes that might occur, such as Ponzi Schemes, phishing scams, and money laundering, and grouped them into four categories, including those considered harmful by experts, those in the public sector, those in the private sector, and those in the academic sector. The authors emphasized the need for further research to address the problems in the future.

3.2.1 Money Laundering

It's important to note that most of the frauds identified and studied are committed using cryptocurrencies as a financial instrument to conduct illegal activities (Brenig et al., 2015), i.e., cyber-criminals take advantage of the fragilities of the cryptocurrencies to commit their crimes as happens in money laundering, the process of conversion of illicit money which comes out of the crime which is then intermixed with licit money to make appear legitimate (Korejo et al., 2021) and the major type of fraud focused on in this work.

Law enforcement agencies globally recognize the use of cryptocurrencies in various criminal activities, mostly correlated with economic and financial spheres, in particular money laundering (Dyntu & Dykyi, 2018). The use of cryptocurrencies in money laundering activities allows criminal organizations to seamlessly transfer funds while evading detection by authorities, enabling them to conduct their illegal activities with greater success and less likelihood of being caught (Albrecht et al., 2019). This makes it increasingly difficult for law enforcement agencies to trace and investigate these illegal transactions (Mabunda, 2018).

Even though transactions made using cryptocurrencies are not directly linked to a specific user or organization, they can be traced through their associated addresses (Trozze et al., 2022). As a result, some users turn to alternatives such as Monero, which offers more privacy, or use "mixers" or "tumblers" to obscure the origin of their funds. These methods can form the basis for money-laundering schemes on the blockchain.

The largest case of money laundering to date is that of Liberty Reserve, which in 2013 was accused of laundering a staggering 6 billion US dollars. The organization conducted 55 million transactions using their virtual currency, "Liberty Dollars" (LD) but converted and stored the funds in fiat currency (USD) at the beginning and end of each transaction. This allowed them to evade detection and launder the funds using a virtual currency while still ultimately storing the funds in a traditional, government-backed currency (Dyntu & Dykyi, 2018).

3.2.2 Handling fraud

Given the growing threat of fraud, organizations, particularly those handling cryptocurrencies like Uphold, must take action to detect and prevent these deceitful activities. This can be accomplished by utilizing existing data and implementing effective controls, which will help to safeguard the organization and its customers against financial losses and other negative impacts of fraud.

Anomaly detection is a key technique for addressing fraud, as it involves identifying unusual items, events, or observations that raise suspicions by deviating significantly from most of the data. Fraud detection, which is a form of anomaly detection, is based on identifying anomalies and suspicious behavior in transactions and trade history.

However, criminals often attempt to conceal their fraudulent activities by mimicking normal transactional behavior (Lorenz et al., 2020), making it difficult to detect fraudulent activity as an anomaly. As fraudsters continue to develop new methods for committing fraud, it is essential to regularly improve these techniques to keep pace with their evolving tactics (Sabry et al., 2020). Additionally, due to the decentralized and transparent nature of blockchain technology, detecting fraudulent activity by rogue users is a challenging task to analyze, as it is hard to trace the generation of money and profits in such markets (Sureshbhai et al., 2020).

3.3 Anti-Money Laundering

Chen et al. (2018) emphasize that given the significant financial losses that result from money laundering, organizations must implement effective AML systems, controls, and practices to mitigate the risk of money laundering activities.

AML is a set of activities that involve the examination of customers' transactions by analysts and automated systems. The goal is to identify suspicious behaviors that may be linked to money laundering (Labanca et al., 2022).

AML solutions are a component of overall fraud control measures, and they help to automate and reduce the manual workload of the screening and checking process. AML solutions typically include a combination of software, databases, and analytical tools that can be used to monitor and analyze financial transactions for suspicious activity. (Chen et al., 2018). This includes monitoring transaction patterns, performing customer due diligence, and implementing procedures to detect and report suspicious activities. Such measures can help organizations identify and prevent money laundering and to comply with legal and regulatory requirements.

Traditionally, AML regulations require banks and other financial institutions to perform due diligence to monitor and prevent customers from performing transactions that may be involved in money laundering (Mabunda, 2018). This helps to ensure that banks are proactive in the fight against money laundering.

However, the decentralized and transparent nature of cryptocurrencies and blockchain technology makes it much harder to implement traditional AML practices in this case. The pseudonymous and decentralized nature of blockchain transactions makes it difficult to track the origin of the funds, making it harder to identify and prevent money laundering activities.

Traditional AML systems are designed to monitor unusual behaviors and detect potential money laundering activities. However, as money laundering techniques and financial crime are always evolving, the rules and algorithms used by these systems must be updated to ensure that they are able to capture these changes (Labanca et al., 2022). Additionally, these systems are based on pre-determined rules and are only able to detect known unusual behavior, which can result in false negatives if unknown methods of money laundering are used (Labanca et al., 2022). Furthermore, the use of specific static thresholds in these systems can result in a high number of false positives (Labanca et al., 2022).

Therefore, to effectively detect money laundering activities and comply with AML regulations, there is a need to develop new methods and technologies that can adapt to the ever-evolving nature of money laundering and reduce the number of false positives and negatives.

3.3.1 Machine Learning

ML plays a crucial role in the field of AML by helping to identify and prevent financial crimes. Overall, ML models can be used to identify unusual correlations and assess the effectiveness of money laundering detection algorithms, making them a valuable tool in the fight against financial crime. But it is crucial to understand the strengths and limitations of each algorithm when applied to the problem of detecting money laundering. This can help ensure that the appropriate model is used for a given task and that the results are interpreted correctly.

There are two main types of ML models: supervised and unsupervised. Supervised ML is trained on labeled data and is used to make predictions on new, unseen data. Unsupervised ML is trained on unlabeled data and is used to find patterns or structures in the data without any prior knowledge of the labels. Supervised learning algorithms, such as decision trees and random forests, can also be used to classify cryptocurrency transactions as suspicious or non-suspicious based on historical data. Unsupervised learning algorithms, such as k-means and neural networks, can be used to detect patterns in the data that may indicate money laundering activity.

Further, Han et al. (2020) suggest that using AI and graph learning can help enhance current AML solutions. These techniques can analyze financial data, identify patterns and behaviors that may indicate money laundering activities, and help to improve the accuracy and efficiency of AML systems (Sabry et al., 2020). ML can decrease the false-positive rate, while keeping the false-negative rate relatively low, to overcome the drawbacks of traditional AML systems (Vassallo et al., 2021). This can help to detect and prevent money laundering activities involving cryptocurrencies, which pose privacy and security threats.

The blockchain is an effective resource for this study as it allows for the public accessibility of transaction data. This facilitates the verification of the correlation between account operations and the identification of fraudulent activities (Ostapowicz & Żbikowski, 2019). Utilizing the publicly accessible data within the blockchain, researchers can gain valuable insights into the issue of money laundering.

When evaluating the performance of an ML model for AML, it is important to take into consideration metrics such as precision, recall, specificity, and F1-score. These metrics can provide insight into the model's ability to accurately identify money laundering. The choice of metrics used for evaluation depends on the model and whether it is in the training or testing phase. Accuracy is often the most highlighted metric by authors when evaluating the performance of their models.

In summary, AI and graph learning can improve AML systems by analyzing financial data and identifying money laundering patterns. ML can also decrease false positives and negatives to overcome the limitations of traditional AML systems and help detect and prevent money laundering activities involving cryptocurrencies.

3.4 Transactional Graphs

According to a study by Song et al. (2023) it was found that analyzing the information contained within the cryptocurrency transaction network can be an effective method for detecting illegal financial activities involving cryptocurrencies. This can include graph analysis and graph learning. While graph analysis refers to the process of studying and interpreting the relationships and connections between entities within a graph data structure, graph learning, on the other hand, refers to the application of machine learning techniques to graph data. Graph learning can learn the deep representation of node features directly in the topology of the network graph, which avoids time-consuming feature engineering and enhances the feature representation of nodes, making it beneficial for fraud detection in financial transactions (Li et al., 2022).

Graph Anomaly Detection (GAD) is a subfield of graph learning, as it utilizes graph representations and graph algorithms to identify anomalies in graph-structured data and aims to identify anomalous nodes, edges, or sub-graphs, expanding the concept of anomaly detection to graphs (Ma et al., 2021). Although, traditional anomaly detection techniques are not suitable for detecting anomalies in graph data due to the complex, high-dimensional nature of the data. As a result, researchers have turned to deep learning techniques, which have been shown to be effective in representing and recognizing patterns in graph data (Ma et al., 2021).

Graph learning can include various facets such as Node Classification, Graph Classification, Node Clustering, Link Prediction, and Influence Maximization (Daigavane et al., 2021). Graph learning and analysis can provide a deeper understanding of relationships and interactions within the cryptocurrency ecosystem.

Utilizing graph-based methods allows for the identification of patterns and connections that may not be apparent with traditional machine learning algorithms. These methods are typically based on or extend deep learning methods, as they can effectively transform graph data into vector representations for analysis and manipulation (Xia et al., 2021).

One type of graph used in this context is a transaction network graph, which is a vectorial space representation of the nodes and connections (edges) between them (Song et al., 2023). In the UTXO blockchain, such as with Bitcoin, these graphs can be divided into three types: address networks, transaction networks, and user networks. The graph-based representation enriches the data by incorporating the relations between transactions and facilitates data analysis related to link prediction, node classification, forensic investigation, and the relationship between entities (Tharani et al., 2021).

Graph embedding techniques, such as graph2vec or node2vec, have grown in popularity as a means for learning latent representations of vertices on large-scale networks. These techniques effectively reduce the data dimension of the transaction network and transform high-dimensional one-hot node vectors into dense low-dimensional node vectors (Lin et al., 2020).

Graph neural networks (GNNs) are a family of neural networks that can operate on graph-structured data, allowing for the extraction of high-level representations of graphs explicitly (Daigavane et al., 2021). This enables various graph analytic tasks such as classification, recommendation, and clustering to be performed effectively (Kim et al., 2023). The most explored technique in this context is the Graph Convolutional Network (GCN), a neural network that operates on local graph neighborhoods, involving a learnable kernel to output the node embeddings (Alarab et al., 2020b).

4. LITERATURE REVIEW

This literature review focuses on the use of machine learning algorithms for detecting fraud and money laundering. It also examines the application of graph learning and analysis in the

same field. The concepts discussed in the previous section are used to support and clarify the information presented in this review.

4.1 Machine Learning for Anti-Money Laundering

Chen et al. (2018) surveyed the literature on ML techniques for AML solutions in the suspicious transaction detection area and found that efficient data preparation, data transformation, and data analytics techniques are necessary for an AML solution to offer data quality assurance, high detection accuracy, scalability, and fast performance in suspicious transaction detection. They also found that various ML models, including fuzzy logic, Support Vector Machine (SVM), clustering, outlier detection, neural networks, genetic algorithms, Bayesian network, and sequence matching have been tested to better evaluate suspicious transactions in the banking sector.

Considering the security concerns surrounding cryptocurrencies, Choithani et al. (2022) conducted a study to evaluate the effectiveness of AI and ML algorithms in risk management within the banking sector and predict cryptocurrency behavior. Through their research, the authors found that utilizing SVM, Long Short-Term Memory (LSTM), and Artificial Neural Networks yielded superior results compared to other techniques examined.

Kamišalić et al. (2021) reviewed studies that examine the potential benefits of combining blockchain technology with data mining techniques for anomaly and fraud detection. They found that the most used methods were Gradient Boosting, SVM, and Random Forest (RF). Additionally, they identified a recent trend toward the use of Deep Learning (Neural Networks, and LSTM) for these applications, which corroborates the graph learning approach explored in this work.

Most of the research in this field utilizes machine learning models on the Elliptic dataset, which has been a valuable asset to the AML and research communities (Alarab & Prakoonwit, 2022) and was created by (Weber et al. (2019). The dataset, provided by cryptocurrency intelligence company Elliptic, is a graph network of Bitcoin transactions and is one of the largest labeled datasets available for any cryptocurrency. Weber et al. (2019) conducted a study on Anti-Money Laundering in Bitcoin. The study aimed to address the challenge of accurately identifying criminal activity in large, growing datasets of Bitcoin transactions while minimizing false positive rates. The researchers used traditional machine learning models (RF,

LR, and MLP) and compared their performance to that of GCN. The results showed that RF outperformed both LR and GCN, making it a more effective method for classifying Bitcoin transactions as illicit. However, the study highlights the potential of graph-based approaches, as they may offer valuable insights into this field. Alarab et al. (2020a) applied ensemble learning, a combination of various supervised learning techniques, to the Elliptic dataset and demonstrated that it outperforms other methods, including the one proposed by (Weber et al., 2019).

Ostapowicz & Żbikowski (2019) and Fawaz Ibrahim et al. (2021) proposed models to detect fraudulent accounts on the Ethereum network and classify them as "Fraud" or "Non-Fraud" using machine learning. Both studies found that the RF algorithm had the best results in terms of recall and false-positive rate. However, Ostapowicz & Żbikowski (2019) noted that the recall of 84.92% was not suitable for use in a real-world anti-fraud system.

The supervised methods have been widely explored in this research field as many authors have been proving that these have good results, particularly RF. However, the use of these methods is limited by the availability of large-scale labeled datasets (Labanca et al., 2022). i.e., when applying supervised methods, it's important to have historical data with the events accurately assigned, this means that the datasets need to be properly labeled (Zhang & Trubey, 2019).

The labeling process is costly and time-consuming, as it requires manual annotation of historical transactions, and the complexity of money laundering schemes makes it difficult to identify all the entities involved (Lorenz et al., 2020). Even the Elliptic dataset, which is considered one of the largest labeled datasets available, has only 2% of transactions classified as illicit and 21% as licit. Additionally, Alarab et al. (2020a) note that it is not possible to optimize recall and precision simultaneously, as this can lead to an increase in false positives. In light of these limitations, Lorenz et al. (2020) proposed using unsupervised methods and active learning to detect money laundering in datasets with few labels. By using a dataset with just a few hundred labels (5% of the total), the authors obtained similar performance as the best-supervised methods. Alarab & Prakoonwit (2022) also referred to this work and suggested that it could be applied to the Elliptic dataset.

Zhang & Trubey (2019) used ML and sampling schemes (to deal with unbalanced data) to detect money laundering activities in financial transactions and observed that artificial neural networks are the ones that perform better.

Labanca et al. (2022) proposed an active learning system for anomaly detection in transaction monitoring for money laundering detection in capital markets, which includes an unsupervised model for detecting known and unknown anomalous patterns. They found that the Isolation Forest algorithm was the best for this task and highlighted the importance of the unsupervised component in detecting novel patterns that the supervised models were unable to detect.

Due to the increasing amount of available data in the blockchain, also increases the difficulty in searching for preferent data and datasets (Tang et al., 2023). Chen et al. (2018) also pointed out that there is an extremely large volume of raw data that is highly imbalanced. This naturally happens because money laundering activities are rare events and anomalous transactions represent only a small fraction of the total number of transactions (Zhang & Trubey, 2019).

This presents a significant problem as it impacts the accuracy of the model, with a high accuracy not necessarily indicating that the model is useful. Therefore, in this case, it's more useful to analyze the trade-offs between recall and FNR.

To address this issue, Chen et al. (2018) suggested a clustering process to handle the large amounts of available data or use semi-supervised techniques. Yang et al. (2019) used a heuristic user address clustering method to cluster user pseudonyms and propose a user address update algorithm. They mined the user characteristics, classified different types of user features, and distinguished between normal users and abnormal user features. They used the Gaussian Mixture Model to cluster users and identify users with suspected anomalies based on the characteristics of the user classification. Therefore, the authors were able to verify the suspicious user and prove that there are abnormal users among suspicious users of the experiment.

As well, Monamo et al. (2017) used k-means clustering and a trimmed version of the algorithm to detect fraudulent activity in the Bitcoin transactions network. Both algorithms achieved optimal clustering. The trimmed algorithm was able to detect five of thirty well-known anomalies. Using k-d trees, they were able to detect two more anomalies. Based on the clustering labels for outliers, the authors employed some supervised classification models to understand the relation between the labels and predictor variables. RF achieved the best precision.

In conclusion, the literature reviewed suggests that traditional machine learning algorithms such as SVM, Random Forest, and Artificial Neural Networks have been extensively tested for fraud detection in the financial sector, especially for cryptocurrency transactions. However, these methods have limitations in the form of high costs for manual labeling, the difficulty of detecting rare events, and the challenge of balancing recall and false positive rates. On the other hand, unsupervised methods, and active learning, in combination with graph learning, have been proposed as alternative approaches to overcome these limitations. These approaches show promising results, as demonstrated by recent studies such as those by Lorenz et al. (2020) and Labanca et al. (2022). Nevertheless, more research is needed to explore the full potential of these methods in real-world anti-fraud systems.

4.2 Graph Learning and Analysis for Detecting Fraud and Money Laundering

In recent years, several techniques such as GCNs, DeepWalk, and node2vec have emerged as potential solutions for encoding topological structures from graphs into dense representations. These techniques aim to place nodes with similar neighborhoods close in the embedding space (Khazane et al., 2019). Despite their success, these techniques have limitations. For example, DeepWalk and node2vec only consider the structural information of the graph and do not incorporate temporal information, such as changes and evolution over time (Lin et al., 2020). This could be crucial for certain applications, making it a crucial aspect to consider in graph representation studies.

Motamed & Bahrak (2019) explored the properties of five different cryptocurrencies. They aimed to propose algorithms for extracting transaction-related features from Bitcoin-core and Ethereum. The authors generated three different types of graphs, including the Transaction graph, the Address graph, and the Money flow graph, to visualize the data. They found that the address graph and the transaction graph of Bitcoin, as well as the money flow transaction graph of Ethereum, were most suitable for visualizing anomalous transactions. They emphasized the importance of these visualizations for the future application of machine learning algorithms in detecting fraudulent transactions. Sait Canbaz et al. (2020) also acknowledged the significance of data visualization in fraud detection and proposed a web-based software tool for detecting financial fraud through graph representation.

Alarab et al. (2020b) proposed a novel approach that combined graph convolutional networks (GCN) with linear layers to perform node classification on the transaction graph. The approach utilized two sets of features, with the first set being node embeddings obtained from GCN and the second set obtained from the latent representation of a linearly transformed hidden layer from the original features. The results of the study indicated that the combination of features derived from GCN, and the latent representation of a linear layer improved the performance of the model in comparison to using graph convolutions alone.

Alarab & Prakoonwit (2022) addressed the limitations of previous studies that used GCN for detecting illicit transactions in the financial industry. These studies did not consider the temporal aspect of transactions, leading to limitations in the model's accuracy. The authors proposed a new model that combined LSTM and GCN and used active learning methods to label unlabeled data. The results showed that the new model outperformed previous studies, with an accuracy of 97.77% and an f1-score of 80%. The integration of LSTM was a key innovation in the study, as it considers the temporal sequence of transactions and improves the performance of the model.

Singh et al. (2021) aimed to tackle the issue of temporal biasing in the elliptic dataset by proposing a GNN-based adversarial loss architecture. The results showed that the GNN-based adversarial loss architecture was effective in addressing the problem of temporal biasing, leading to improved performance compared to previous studies.

Pocher et al. (2022) conducted a study comparing the performance of the GCN and GAT with that of the baseline approaches, such as RF proposed by Weber et al. (2019). Their results showed that GCN outperforms the baseline approaches, suggesting the potential value of GCN for this type of research.

Geng et al., (2022) introduced a graph attention mechanism, an improved version of the GCN, to enhance the traceback accuracy for identifying illegal behavior. The results of the study showed the potential of the graph attention mechanism as an effective approach for detecting illegal activities in blockchain networks. The findings of this study contribute to the growing body of literature on utilizing graph learning techniques for addressing security issues in blockchain systems.

Lin et al. (2020) developed a graph embedding method that integrates temporal and weighted information from financial transaction networks into node embeddings and applied it to an Ethereum network for phishing and non-phishing node classification. The study highlights the

importance of considering temporal and weighted information in financial transaction networks for fraud detection.

Caglayan & Bahtiyar (2022) proposed an approach to detect money laundering in financial transactions using node2vec, a graph embedding algorithm. The results were evaluated by calculating true positives, false positives, false negatives, and true negatives on a banking transactions dataset containing money laundering transactions. The study showed that node2vec provided better results in detecting money laundering compared to other methods. Similarly, in the study by Lopes et al. (2022), node2vec was used to obtain vector representations of nodes and edges of criminal networks. The combined use of node2vec and predictive ML methods were found to be effective in distinguishing between criminal, non-criminal, and mixed relationships in criminal networks. These studies highlight the potential of node2vec as a valuable tool in the analysis of financial and criminal transactions.

Further, Zhou et al. (2021) presented a distributed big data approach for detecting Internet financial fraud. The approach utilized node2vec to represent the topological features of the financial network graph as low-dimensional dense vectors. The resulting representation was then used to classify and predict financial fraud with the help of a deep neural network. The study showed that the proposed approach improved the efficiency of financial fraud detection with better precision, recall, F1-Score, and F2-Score. These findings suggest the potential of utilizing graph embedding and deep learning for effective and efficient Internet financial fraud detection.

Li et al. (2022) proposed a novel method called TA-Struc2Vec for detecting fraudulent users in financial transaction networks. The method considers both structural and transaction amount homogeneity in the network and employs ML classification algorithms to classify users. The proposed approach, TA-Struc2Vec, was shown to learn stronger structural homogeneity and further transaction amount homogeneity compared to previous methods. The results of the study indicate the potential of considering both structural and transaction amount information in detecting fraudulent users in financial transaction networks.

Yu et al. (2022) found that phishing nodes in Ethereum networks can be identified by their rapid transfer of funds obtained from multiple inputs. To detect such fraud, the authors proposed a GCN approach and evaluated its performance using transaction records to construct a labeled graph for classification. These findings suggest that graph-based methods

have the potential for detecting phishing fraud in cryptocurrency networks and may also be applicable to other types of fraud.

The reviewed literature underscores the importance of graph representation techniques in detecting fraudulent transactions, both financial and criminal. The use of GCN, DeepWalk, node2vec and their variants (e.g., GNN-based adversarial loss architecture, graph attention mechanism) has demonstrated potential in improving fraud detection accuracy. Integrating temporal and weighted information and combining graph representation with deep learning enhances the performance of fraud detection models. These findings provide a basis for future research and showcase the potential of graph representation in addressing security concerns in financial and criminal transactions.

5. METHOD

5.1 Dataset

Due to privacy concerns and feasibility problems, real Uphold data could not be used to develop the project. Then, one of the main points to develop the methodology was to find an online available dataset, that was viable to conduct the project.

The dataset chosen to implement the project is the Elliptic Dataset because, as described in the previous section of the literature review, it is the most explored dataset in detecting money laundering. Moreover, after an intensive search, this is the most complete dataset that allows for achieving the goals of this project since it is graph-based, which makes it adequate for its needs. Additionally, it corresponds to bitcoin transactions, which, is the cryptocurrency where most studies focus their attention, it’s the most widely known cryptocurrency, and it’s the most secure one.

A detailed description of the dataset follows.

This dataset is the world's largest labeled transaction dataset publicly available in any cryptocurrency and was created to motivate and enable the development of new techniques for detecting illicit cryptocurrency transactions.

The dataset is anonymized and corresponds to a transactional graph where a node represents a transaction, an edge can be viewed as a flow of transactions between one transaction and the other. Each transaction (node) has a label “licit”, “illicit” or “unknown”.

The graph is made of 203,769 nodes and 234,355 edges. Two percent (4,545) of the nodes are labeled class 1 (illicit). Twenty-one percent (42,019) are labeled class 2 (licit). The remaining transactions are labeled as unknown.

There is a time step associated with each node, representing a measure of the time when a transaction was broadcasted to the Bitcoin network. The time steps, running from 1 to 49, are evenly spaced with an interval of about two weeks. Each time step contains a single connected component of transactions that appeared on the blockchain within less than three hours between each other; there are no edges connecting the different time steps.

Each node has 166 features, with the first 94 features representing local information about the transaction – including the time step, number of inputs/outputs, transaction fee, output volume and aggregated figures such as average BTC received (spent) by the inputs/outputs and average number of incoming (outgoing) transactions associated with the inputs/outputs. And the remaining 72 features are aggregated features, obtained using transaction information one-hop backward/forward from the center node (giving the maximum, minimum, standard deviation, and correlation coefficients of the neighbor transactions for the same information data (number of inputs/outputs, transaction fee, etc.)).

Due to intellectual property issues, there is no exact description of all the features in the dataset, nor do we have the columns' names.

5.2 Tools

5.2.1 Python

The programming language to be utilized for this project is Python.

Python is a popular choice for machine learning due to its simplicity, readability, and a large community of developers. It has a wide range of libraries and frameworks that make it easy to implement various machine learning algorithms and techniques, such as scikit-learn, and TensorFlow. These libraries provide pre-built functions and tools for tasks such as data processing, model training, and evaluation, which makes it easy for developers to quickly

prototype and test their models. Overall, Python's ease of use, community support, and rich ecosystem of ML libraries make it an ideal choice for developing ML projects.

Following, there's a description of the main libraries used in this work.

Scikit-learn: This is a library in Python that provides many unsupervised and supervised learning algorithms, namely, Regression, Classification, and Clustering.

Pandas: This is a fast, powerful, flexible, and easy-to-use open-source data analysis and manipulation tool, built on top of the Python programming language.

Numpy: It's a Python library used for working with arrays.

Networkx: This is a package for the creation, manipulation, and study of the structure, dynamics, and functions of complex networks.

5.3 The exploitation of existing techniques

In view of the literature review conducted in the previous section, some concepts and techniques need to be further explored to understand how they work and how they can be applied to the current work. The literature review highligts the application of supervised machine learning algorithms, deep learning, node embedding, and graph neural network, namely GCN.

Graphs are different from normal data in the sense that it brings a lot of complexity, so the existing machine learning algorithms face various challenges. Also, the interconnected nature of nodes in a graph provides valuable information that can be leveraged for graph representation and machine learning.

5.3.1 Supervised Learning

Supervised learning on graphs is a type of machine learning where the goal is to learn a function that maps the graph's nodes or edges to a target label. The function is learned based on a labeled dataset, where the graph and the labels of certain nodes or edges are given.

5.3.2 Semi-Supervised Learning

Semi-supervised learning on graphs is a type of machine learning where the goal is to learn a function that maps the graph's nodes or edges to a target label. The function is learned based

on a partially labeled dataset, where the graph and the labels of only a subset of the nodes or edges are given. The remaining nodes or edges are unlabeled. The idea behind semi-supervised learning on graphs is that the labels of the nodes or edges are correlated with the graph structure and the relationships between the nodes.

In node classification, the goal is to predict the label of unlabeled nodes based on the labels of the labeled nodes and the graph structure.

5.3.3 The traditional strategy

When working with traditional machine learning methods on graph data, the input is typically a graph, and the goal is to extract features that describe the topological structure of the network around a specific node. This process is known as feature engineering. These topological features, such as the node's degree and centrality, can be combined with attribute-based information to train a classical ML model, such as an SVM and LR.

Figure 1 represents the traditional strategy when applying traditional ML strategies to graphs.

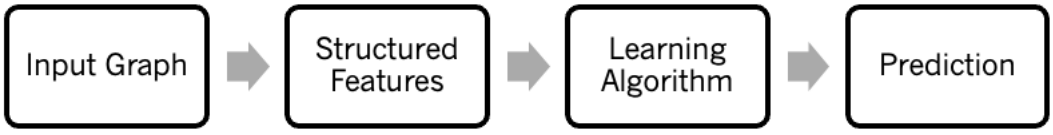

Figure 1: Traditional ML strategy for node classification

The process of feature engineering for graphs can be time-consuming and labor-intensive. Representation learning addresses this challenge by automatically extracting features from the graph data.

5.3.4 Node Embeddings

Node embedding is a specific type of representation learning that maps each node to a d-dimensional vector, also known as embedding. Node embedding is a technique used to characterize the features of a node in a graph, such as its degree, centrality, clustering coefficient, and graphlets.

One popular method for creating node embeddings is DeepWalk, which uses random walks to transform nodes in a network and visualize them in 2-dimensional space as vectors. This approach simplifies the feature engineering process and allows the model to learn the data structure more efficiently. DeepWalk is a graph embedding technique that is used to create node embeddings. It is based on the idea of using random walks to generate sequences of nodes and then using the sequences to train a Skip-Gram model, a neural network model commonly used for word embeddings. This allows DeepWalk to learn low-dimensional representations of the nodes in a graph that preserves the similarity between nodes. The resulting node embeddings can be used as input for various ML models.

Node2vec is an extension of the DeepWalk algorithm that aims to learn more expressive node embeddings by incorporating the idea of biased random walks. The idea behind node2vec is to explore the graph in different ways by controlling the probability of returning to the previous node, the probability of moving to a node connected to the current node, or the probability of moving to a node with a similar degree. This allows node2vec to learn embeddings that capture both the local and global structure of the graph. The embeddings generated by node2vec can be used as input for various ML models. Therefore, node2vec is a method to generate node embeddings that can be used in downstream ML tasks.

5.3.5 Deep Learning

Despite the efficacy of traditional machine learning techniques within the field of AML, the complexity of graph data has presented a number of challenges for these algorithms. To address this, GNNs have been developed as a specialized class of deep learning methods, specifically designed for the inference of graph-structured data. GNNs offer a direct application to graph data and provide a convenient means for performing node-level, edge-level, and graph-level predictions.

The utilization of GCNs is a prominent strategy for applying deep learning techniques to graph-structured data. GCNs are a specialized type of neural network that operates directly on graph data, leveraging the network structure to process the data. This allows GCNs to take advantage of the topological structure of the graph, incorporating information from both neighboring nodes and individual node features. As a result, GCNs are particularly well-suited for tasks such as node classification and link prediction, as they can effectively capture the

complex relationships present in graph-structured data. Furthermore, GCNs are a powerful tool for semi-supervised learning, a common task in graph-based problems.

it has been shown to be effective in identifying patterns of illicit activity, detecting suspicious transactions, and uncovering the underlying structure of the blockchain network in the context of cryptocurrency transaction analysis.

The fundamental inputs for a GCN are the graph structure and node features. The graph structure, which is representative of the connections between nodes in the graph, is commonly represented in the form of an adjacency matrix or an edge list. The node features, on the other hand, denote the characteristics associated with each node within the graph and are represented as a matrix, with each row corresponding to a node and each column corresponding to a feature.

One of the defining characteristics of GCNs is their hierarchical structure, which is composed of multiple layers, each with a unique function. These layers work in conjunction to process the graph data and produce the desired output for a given task. The number of layers in a GCN can be thought of as the maximum distance that the node features can travel within the graph. Commonly found layers in a GCN include the input layer, graph convolutional layers, non-linearity layer, and output layer. The number and configuration of these layers may vary depending on the specific task and the complexity of the graph data.

The essential concept behind the layers is that the output of one layer serves as the input for the next, allowing for gradual and thorough processing of the graph data. The l-th layer of the GCN takes the adjacency matrix A and the node embedding matrix $H^{(l)}$ as input, and uses a weight matrix $W^{(l)}$ to update the node embedding matrix to $H^{(l+1)}$ as output, using an activation function, typically ReLU (Weber et al., 2019).

5.4 Task

This section outlines the methodology and experimental setup proposed, which is based on the work of Weber et al. (2019) and Pocher et al. (2022). The study builds upon the work of Weber et al. (2019) and Pocher et al. (2022) and takes inspiration from the techniques and methods outlined in the previous section of the dissertation.

The primary objective is to replicate the results of Weber et al. (2019) and Pocher et al. (2022) in order to establish a benchmark against which to evaluate the results of this research.

To accomplish this, supervised methods for binary classification, such as RF, LR, and Multilayer Perceptron (MLP) will be initially employed. These methods are widely used in the field of AML but do not incorporate graph information. To address this limitation, it will also employ deep learning techniques, specifically GCN, as used by Weber et al. (2019) and Pocher et al. (2022) in its research.

The previous study conducted by Weber et al. (2019) found that the RF algorithm with the dataset features concatenated with the embeddings produced by GCN yielded the best results. However, Pocher et al. (2022) in their research discovered that GCN alone yielded the best results, without testing the node embeddings it produces with the supervised methods. This raises the question of whether GCN alone is the optimal solution or if the embeddings it produces play a crucial role in achieving superior performance. In this study, we propose to also investigate the use of a representation learning technique called node2vec to understand if there is another way to obtain node embeddings that can lead to improved performance of supervised models. This study aims to provide a deeper understanding of the potential of GCN and node embeddings in achieving superior performance.

5.4.1 Experimental Setup

As previously stated, this study is based on the Elliptic Dataset, as the Weber et al. (2019) and Pocher et al. (2022) studies. To ensure the comparability of results, similar experimental setups will be applied.

The dataset will be divided temporally 70:30 for training and testing purposes. Specifically, the first 34 timesteps will be used for training, while the remaining timesteps until 49 will be used for testing. This approach allows for a fair evaluation of the models by ensuring that the models are trained on a representative sample of the data and tested on unseen data.

In terms of preprocessing, the features were merged with the classes, class values were renamed to integer values and only the part of the dataset labeled licit or illicit was selected.

Different experimental setups will be tested to evaluate the performance of the models. The following steps will be followed:

1. Supervised models (RF, LR, and MLP) will be trained and tested using only the local features of the dataset.

2. The same supervised models will be trained and tested again, this time using all the features of the dataset.
3. The node2vec algorithm will be applied to obtain node embeddings and these embeddings will be combined with all the features of the dataset. The supervised models will be trained and tested again using this augmented dataset.
4. A GCN model will be trained and tested using the entire dataset.

This approach allows for a comprehensive evaluation of the models by testing them under different settings and comparing the results obtained with traditional supervised methods and deep learning techniques.

With these different setups, we aim to understand the impact of the use of different features, embeddings, and graph information on the performance of the models, and to establish a benchmark that can be used to evaluate the results of this research.

5.4.2 Evaluation Metrics

By using the same parameters as in previous studies of Weber et al. (2019) and Pocher et al. (2022), the evaluation of the solutions will be conducted in a consistent and comparable way. This will allow for an accurate comparison of the performance of the solutions to the established baselines. Benchmarking the performance of each approach studied in this work, will determine the effectiveness of the solutions, and identify areas for improvement.

The use of metrics such as Precision, Recall, F1-score, and the Micro Average Score will allow for a comprehensive evaluation of the solutions. Precision measures the proportion of true positives among the predicted positives, recall measures the proportion of true positives among the actual positives, and F1-score is a measure that combines precision and recall. Micro Average Score is a measure that calculates the average precision and recall for all the classes.

For each of the approaches outlined in the previous section, each supervised technique will be evaluated by making predictions on the test set and calculating performance metrics. The same process will be applied when evaluating the GCN.

6. PLAN

6.1 Risks

There are no relevant risks associated with this work.

6.2 Activities

This section contains the full list of activities to perform during this work.

Activity 1: Determining the research's focus, including the issue or opportunity it aims to address and its primary goals.

Activity 2: Conducting an in-depth study of the current knowledge and developments in the fields of Cryptocurrencies, Fraud detection, AML, and Graph Analysis.

Activity 3: Performing a thorough review of existing literature to examine the use of ML algorithms and graph analysis techniques in the detection and prevention of fraud and AML in the context of Cryptocurrencies.

Activity 4: Finding and selecting a suitable dataset that is representative of the area of interest.

Activity 5: Based on the findings from the literature review, propose a solution and develop a method for achieving the proposed solution.

Activity 6: Select the appropriate software or tools for handling the dataset, analyzing the data, and applying the chosen method.

Activity 7: Writing the intermediate report.

Activity 8: Carry out replications of the baseline methods from the literature review that used the same techniques and dataset as the proposed solution

Activity 9: Determining the node embeddings for the dataset, and then reproducing the baseline methods using these node embeddings to establish a fair comparison and evaluate the performance of the proposed solution.

Activity 10: Applying deep learning techniques to evaluate their performance.

Activity 11: Writing the final report.

Activity	Oct 2022	Nov 2022	Dec 2022	Jan 2023	Feb 2023	Mar 2023	Apr 2023	May 2023	Jun 2023
Activity 1	X								

Activity 2	X	X							
Activity 3		X	X	X					
Activity 4			X	X					
Activity 5				X					
Activity 6				X					
Activity 7	X	X	X	X	X				
Activity 8					X	X			
Activity 9						X			
Activity 9							X	X	
Activity 10	X	X	X	X	X	X	X	X	X

7. CONCLUSIONS

First, the problems that led to this dissertation were identified and are centered on the issue of fraud in cryptocurrency transactions, which Uphold must prevent to meet legal and industry standards. The research aims to enhance Uphold's fraud detection and prevention efforts and make a meaningful contribution to the field. The study of AML controls and ML in organizations, as well as the potential of graph analysis and deep learning, offers a rich and untapped area for exploration and growth in the field. The need for constant innovation and improvement in fraud detection methods due to the evolving nature of the problem was also acknowledged.

Subsequently, the objectives to respond to these problems and motivations were outlined corresponding to the investigation of the cryptocurrency market, detection of patterns of fraudulent behavior using graph and ML techniques, and the application of these insights to the Uphold platform to detect and prevent illegal transactions and support compliance with regulations and risk management activities.

Next, A comprehensive state of the art was conducted to address the key themes of fraud, money laundering, AML, graph learning, and analysis. A literature review was also carried out to understand the application of ML and graph learning and analysis in the context of AML.

The DSR methodology was chosen for this research due to its suitability for addressing practical issues and generating new knowledge and artifacts that can be used to improve existing systems or create new ones.

Finally, in accordance with the outlined objectives, a detailed plan of action was created including a timeline for the various tasks to be completed during the project's progression. Regarding the risks, there were no identified risks.

In conclusion, the objectives set forth for this intermediate report and the phase it represents have been successfully achieved. Additionally, the activities defined in the initial work plan to be carried out until this date were fully completed.

REFERENCES

Alarab, I., & Prakoonwit, S. (2022). Graph-Based LSTM for Anti-money Laundering: Experimenting Temporal Graph Convolutional Network with Bitcoin Data. *Neural Processing Letters*. https://doi.org/10.1007/s11063-022-10904-8

Alarab, I., Prakoonwit, S., & Nacer, M. I. (2020a). *Comparative Analysis Using Supervised Learning Methods for Anti-Money Laundering in Bitcoin*. https://doi.org/10.1145/3409073.3409078

Alarab, I., Prakoonwit, S., & Nacer, M. I. (2020b). *Competence of Graph Convolutional Networks for Anti-Money Laundering in Bitcoin Blockchain*. https://doi.org/10.1145/3409073.3409080

Albrecht, C., Duffin, K. M. K., Hawkins, S., & Morales Rocha, V. M. (2019). The use of cryptocurrencies in the money laundering process. *Journal of Money Laundering Control*, *22*(2), 210–216. https://doi.org/10.1108/JMLC-12-2017-0074/FULL/PDF

Astrakhantseva, I., Astrakhantsev, R., & Los, A. (2021). Cryptocurrency fraud schemes analysis. *SHS Web of Conferences*, *106*, 02001. https://doi.org/10.1051/shsconf/202110602001

Bank, E. C. (2015). Virtual currency schemes – a further analysis. In *European Central Bank* (Issue February).

Brenig, C., Accorsi, R., & Müller, G. (2015). Economic Analysis of Cryptocurrency Backed Money Laundering. *ECIS 2015 Completed Research Papers*. https://doi.org/10.18151/7217279

Caglayan, M., & Bahtiyar, S. (2022). Money Laundering Detection with Node2Vec. *Gazi University Journal of Science*, *35*(3), 854–873. https://doi.org/10.35378/GUJS.854725

Chen, Z., le Dinh, ·, Khoa, V., Ee, ·, Teoh, N., Nazir, · Amril, Ettikan, ·, Karuppiah, K., Kim, ·, & Lam, S. (2018). Machine learning techniques for anti-money laundering (AML) solutions in suspicious transaction detection: a review. *Knowl Inf Syst*, *57*, 245–285. https://doi.org/10.1007/s10115-017-1144-z

Choithani, T., Chowdhury, A., Patel, S., Patel, P., Patel, D., & Shah, M. (2022). A Comprehensive Study of Artificial Intelligence and Cybersecurity on Bitcoin, Crypto Currency and Banking System. *Annals of Data Science*, 1–33. https://doi.org/10.1007/S40745-022-00433-5/TABLES/2

Daigavane, A., Ravindran, B., & Aggarwal, G. (2021). Understanding Convolutions on Graphs. *Distill*, *6*(9), e32. https://doi.org/10.23915/DISTILL.00032

Dyntu, V., & Dykyi, O. (2018). CRYPTOCURRENCY IN THE SYSTEM OF MONEY LAUNDERING. *Baltic Journal of Economic Studies*, *4*(5), 75–81. https://doi.org/10.30525/2256-0742/2018-4-5-75-81

Eigelshoven, F., Parry, D., & Eigelshoven, F. (2021). *AIS Electronic Library (AISeL) Cryptocurrency Market Manipulation – A Systematic Literature Review Cryptocurrency Market Manipulation : A Systematic Literature Review*. 0–17.

Fawaz Ibrahim, R., Mohammad Elian, A., & Ababneh, M. (2021). *Illicit Account Detection in the Ethereum Blockchain Using Machine Learning*. https://doi.org/10.1109/ICIT52682.2021.9491653

Gandal, N., Hamrick, J. T., Moore, T., & Oberman, T. (2018). Price manipulation in the Bitcoin ecosystem. *Journal of Monetary Economics*, *95*, 86–96. https://doi.org/10.1016/j.jmoneco.2017.12.004

Geng, Z., Cao, Y., Li, J., & Han, Y. (2022). Novel blockchain transaction provenance model with graph attention mechanism. *Expert Systems with Applications*, *209*. https://doi.org/10.1016/J.ESWA.2022.118411

Han, J., Huang, Y., Liu, S., & Towey, K. (2020). Artificial intelligence for anti-money laundering: a review and extension. *Digital Finance*, *2*(3–4), 211–239. https://doi.org/10.1007/s42521-020-00023-1

Kamišalić, A., Kramberger, R., & Fister, I. (2021). Synergy of Blockchain Technology and Data Mining Techniques for Anomaly Detection. *Applied Sciences 2021, Vol. 11, Page 7987*, *11*(17), 7987. https://doi.org/10.3390/APP11177987

Khazane, A., Rider, J., Serpe, M., Gogoglou, A., Hines, K., Bruss, C. B., & Serpe, R. (2019). DeepTrax: Embedding Graphs of Financial Transactions. *Proceedings - 18th IEEE International Conference on Machine Learning and Applications, ICMLA 2019*, 126–133. https://doi.org/10.48550/arxiv.1907.07225

Khrestina, M. P., Dorofeev, D. I., Kachurina, P. A., Usubaliev, T. R., Dobrotvorskiy, A. S., Khrestina, M. P., Dorofeev, D. I., Kachurina, P. A., Usubaliev, T. R., & Dobrotvorskiy, A. S. (2017). Development of Algorithms for Searching, Analyzing and Detecting Fraudulent Activities in the Financial Sphere. In *European Research Studies Journal: Vol. XX*.

Kim, J., Lee, S., Kim, Y., Ahn, S., & Cho, S. (2023). Graph Learning-Based Blockchain Phishing Account Detection with a Heterogeneous Transaction Graph. *Sensors 2023, Vol. 23, Page 463*, *23*(1), 463. https://doi.org/10.3390/S23010463

Korejo, M. S., Rajamanickam, R., & Muhamad, M. H. (2021). The concept of money laundering: a quest for legal definition. *Journal of Money Laundering Control*, *24*(4), 725–736. https://doi.org/10.1108/JMLC-05-2020-0045/FULL/PDF

Labanca, D., Primerano, L., Markland-Montgomery, M., Polino, M., Carminati, M., & Zanero, S. (2022). Amaretto: An Active Learning Framework for Money Laundering Detection. *IEEE Access*, *10*, 41720–41739. https://doi.org/10.1109/ACCESS.2022.3167699

Li, R., Liu, Z., Ma, Y., Yang, D., & Sun, S. (2022). Internet Financial Fraud Detection Based on Graph Learning; Internet Financial Fraud Detection Based on Graph Learning. *IEEE Transactions on Computational Social Systems*, *PP*. https://doi.org/10.1109/TCSS.2022.3189368

Lin, D., Wu, J., Yuan, Q., & Zheng, Z. (2020). T-EDGE: Temporal WEighted MultiDiGraph Embedding for Ethereum Transaction Network Analysis. *Frontiers in Physics*, *8*. https://doi.org/10.3389/FPHY.2020.00204

Lopes, D. D., Cunha, B. R. da, Martins, A. F., Gonçalves, S., Lenzi, E. K., Hanley, Q. S., Perc, M., & Ribeiro, H. v. (2022). Machine learning partners in criminal networks. *Scientific Reports 2022 12:1*, *12*(1), 1–9. https://doi.org/10.1038/s41598-022-20025-w

Lorenz, J., Inês Silva, M., Aparício, D., João Tiago Ascensão, F., Pedro Bizarro, F., & Tiago Ascensão, J. (2020). *Machine learning methods to detect money laundering in the Bitcoin blockchain in the presence of label scarcity*. https://doi.org/10.1145/3383455

Ma, X., Wu, J., Member, S., Xue, S., Yang, J., Zhou Quan Sheng, C. Z., Xiong, H., & Akoglu, L. (2021). *A Comprehensive Survey on Graph Anomaly Detection with Deep Learning*. 1. https://doi.org/10.1109/TKDE.2021.3118815

Mabunda, S. (2018). Cryptocurrency: The New Face of Cyber Money Laundering. *2018 International Conference on Advances in Big Data, Computing and Data Communication Systems, IcABCD 2018*. https://doi.org/10.1109/ICABCD.2018.8465467

Monamo, P. M., Marivate, V., & Twala, B. (2017). *A Multifaceted Approach to Bitcoin Fraud Detection: Global and Local Outliers*. 188–194. https://doi.org/10.1109/ICMLA.2016.0039

Motamed, A. P., & Bahrak, B. (2019). Quantitative analysis of cryptocurrencies transaction graph. *Applied Network Science*, *4*(1). https://doi.org/10.1007/S41109-019-0249-6

Ostapowicz, M., & Żbikowski, K. (2019). Detecting Fraudulent Accounts on Blockchain: A Supervised Approach. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *11881 LNCS*, 18–31. https://doi.org/10.48550/arxiv.1908.07886

Pocher, N., Zichichi, M., Merizzi, F., Shafiq, M. Z., & Ferretti, S. (2022). *Detecting Anomalous Cryptocurrency Transactions: an AML/CFT Application of Machine Learning-based Forensics*.

Rejeb, A., Rejeb, K., & Keogh, J. G. (2021). Cryptocurrencies in Modern Finance: A Literature Review. *ETIKONOMI*, *20*(1), 93–118. https://doi.org/10.15408/ETK.V20I1.16911

Sabry, F., Labda, W., Erbad, A., & Malluhi, Q. (2020). Cryptocurrencies and artificial intelligence: Challenges and opportunities. *IEEE Access*, *8*, 175840–175858. https://doi.org/10.1109/ACCESS.2020.3025211

Sait Canbaz, Y., Doğrusöz, U., Çeliksoy, M., Güngör Finance Turkcell Ödeme Hizmetleri AŞ, F., & Kurban, K. (2020). *Hydra: detecting fraud in financial transactions via graph based representation and visual analysis; Hydra: detecting fraud in financial transactions via graph based representation and visual analysis*. https://doi.org/10.1109/ISMSIT50672.2020.9255191

Singh, A., Gupta, A., Wadhwa, H., Asthana, S., & Arora, A. (2021). Temporal Debiasing using Adversarial Loss based GNN architecture for Crypto Fraud Detection. *Proceedings - 20th IEEE International Conference on Machine Learning and Applications, ICMLA 2021*, 391–396. https://doi.org/10.1109/ICMLA52953.2021.00067

Song, W., Zhang, W., Wang, J., Zhai, L., Jiang, P., Huang, S., & Li, B. (2023). *Blockchain Data Analysis from the Perspective of Complex Networks: Overview* (Vol. 28, Issue 1). https://www.webofscience.com

Sureshbhai, P. N., Bhattacharya, P., & Tanwar, S. (2020). KaRuNa: A blockchain-based sentiment analysis framework for fraud cryptocurrency schemes. *2020 IEEE International Conference on Communications Workshops, ICC Workshops 2020 - Proceedings*. https://doi.org/10.1109/ICCWORKSHOPS49005.2020.9145151

Tang, J., Lu, X., Xiang, Y., Shi, C., & Gu, J. (2023). Blockchain search engine: Its current research status and future prospect in Internet of Things network. *Future Generation Computer Systems*, *138*, 120–141. https://doi.org/10.1016/J.FUTURE.2022.08.008

Tharani, J. S., Charles, E. Y. A., Hou, Z., Palaniswami, M., & Muthukkumarasamy, V. (2021). Graph based visualisation techniques for analysis of blockchain transactions. *Proceedings - Conference on Local Computer Networks, LCN*, *2021-October*, 427–430. https://doi.org/10.1109/LCN52139.2021.9524878

Trautman, L. (2014). *Virtual Currencies Bitcoin & What Now After Liberty Reserve, Silk Road, and Mt. Gox?* (Vol. 20). http://scholarship.richmond.edu/jolt/vol20/iss4/3

Trozze, A., Kamps, J., Akartuna, E. A., Hetzel, F. J., Kleinberg, B., Davies, T., & Johnson, S. D. (2022). Cryptocurrencies and future financial crime. *Crime Science*, *11*(1), 1–35. https://doi.org/10.1186/S40163-021-00163-8/TABLES/8

Vassallo, D., Vella, V., & Ellul, J. (2021). Application of Gradient Boosting Algorithms for Anti-money Laundering in Cryptocurrencies. *SN Computer Science*, *2*, 143. https://doi.org/10.1007/s42979-021-00558-z

Weber, M., Domeniconi, G., Chen, J., Karl Weidele, D. I., Bellei Elliptic, C., Robinson Elliptic, T., Leiserson, C. E., Bellei, C., & Robinson, T. (2019). *Anti-Money Laundering in Bitcoin: Experimenting with Graph Convolutional Networks for Financial Forensics*.

Xia, F., Sun, K., Yu, S., Aziz, A., Wan, L., Pan, S., & Liu, H. (2021). Graph Learning: A Survey. *IEEE Transactions on Artificial Intelligence*, *2*(02), 109–127. https://doi.org/10.1109/TAI.2021.3076021

Yang, L., Dong, X., Xing, S., Zheng, J., Gu, X., & Song, X. (2019). An abnormal transaction detection mechanim on bitcoin. *Proceedings - 2019 International Conference on Networking and Network Applications, NaNA 2019*, 452–457. https://doi.org/10.1109/NANA.2019.00083

Yu, T., Chen, X., Xu, Z., & Xu, J. (2022). *MP-GCN: A Phishing Nodes Detection Approach via Graph Convolution Network for Ethereum*. https://doi.org/10.3390/app12147294

Yuan, Y., Member, S., & Wang, F.-Y. (2018). Blockchain and Cryptocurrencies: Model, Techniques, and Applications; Blockchain and Cryptocurrencies: Model, Techniques, and Applications. *SYSTEMS*, *48*(9). https://doi.org/10.1109/TSMC.2018.2854904

Zhang, Y., & Trubey, P. (2019). Machine Learning and Sampling Scheme: An Empirical Study of Money Laundering Detection. *Computational Economics*, *54*(3), 1043–1063. https://doi.org/10.1007/S10614-018-9864-Z

Zhou, H., Sun, G., Fu, S., Wang, L., Hu, J., & Gao, Y. (2021). Internet Financial Fraud Detection Based on a Distributed Big Data Approach with Node2vec. *IEEE Access*, *9*, 43378–43386. https://doi.org/10.1109/ACCESS.2021.3062467