MBA Thesis

IU University of Applied Sciences
Study program: M.B.A Big Data Management (90 ECTS)

Enhancing Traditional Methods of Identifying Beneficial Owners with Big Data and Machine Learning

Neelesh Khantwal
Enrolment Number: 32209508
Detmolder Strasse 6
Berlin

Supervisor: Prof. Dr. Tobias Broweleit
Date of submission: 08 September 2024

Table of Contents

Tables And Figures

Acknowledgement

I would like to express my gratitude towards my supervisor, Prof. Dr. Tobias Broweleit, for his support, advice and valuable feedback throughout this project. I am grateful to get an opportunity to work under his supervision during this phase.

I thank IU International University of Applied Sciences for providing this platform to work on a subject I have been passionate for. Additionally, the focus on personalized learning provided by the university, as well as the learning resources offered to the students, has really helped my throughout my academic journey. I would like to thank the teachers and other faculty members of the university for the guidance and learnings provided through lectures and courses.

I am grateful to my family who gave me their confidence and support to take on this academic project and trusted me at its completion.

Finally, I thank my friends, near and far, who have been a moral support through this journey.

ABSTRACT

Identifying Beneficial Owners (BO) for business accounts is very important for Anti-money laundering (AML) and Know Your Customer (KYC) regulations, especially in businesses operating in the financial sector. Finding the BO of a business is as cumbersome as ever before due to complex structures of ownership, manual procedures, and disparate data sources. This dissertation, therefore, seeks to examine the potential for the use of big data and machine learning in augmenting effectiveness and efficiency in the process of identifying beneficial owners in business accounts in the United Kingdom (UK) regulatory environment.

The current paper first looks into the existing regulatory requirements and the drawbacks associated with current identification methods. The second part explores the potential of applying big data technologies and machine learning algorithms to overcome these challenges. Data from public registries, social media, and regulatory databases are pulled into Apache Hadoop, Spark, and other open-source big data platforms. The approach for detecting patterns or anomalies that may signal beneficial ownership includes machine learning models using both supervised and unsupervised learning techniques, natural language processing, and network analysis. The results from these will clearly show that big data and machine learning will enhance the present process of identification with more accurate and timely insight compared to the existing traditional methods. Case studies in the practical showcase are going to confirm this kind of effectiveness in situations on network analysis and anomaly detection. The findings in this research provide a robust structure upon which financial institutions can augment their compliance and risk management practices.

CHAPTER 1: INTRODUCTION

In financial industry, beneficial ownership refers to the natural person, who ultimately owns or controls a legal entity, or the one who effectively owns or controls a legal entity that owns or controls a further entity (Mugarura, 2017). Identifying beneficial ownership is central to fighting money laundering and terrorism financing as it provides information of the individuals behind transactions and entities (Chitimira & Munedzi, 2023). The natural person who ultimately owns or controls the legal entity is the beneficial owner, who should not remain unknown in financial dealings (Chitimira & Munedzi, 2023).

Customer Due Diligence (CDD) is especially important in the context of Anti-money Laundering (AML) and Counter Financing for Terrorism (CFT) where they provide the names and dates of birth on beneficial owners (Koker, 2006). As per Chitimira & Munedzi (2023) CDD is vital in this sector because it requires financial institutions to establish an ability to:

- Clearly identify their customers;
- Identify the beneficial owner behind the customer; &
- Develop a solid understanding of the nature and purpose of each customer relationship.

With proper knowledge of the beneficial owners, financial institution are primed to prevent money maundering and terrorism financing cases.

One of the case studies on beneficial ownership is the Panama Paper Leak which shows that known beneficial ownership opacity also make it easier for any entity to launder money, evade tax, and commit other financial crimes (Esoimeme, 2016). This underscores how critical it is to have transparency and make beneficial ownership information easily accessible in order to stop financial crimes and honour regulatory requirements.

From a law enforcement, regulatory and financial industry standpoint, beneficial ownership information is essential to track money flows to apprehend perpetrators of financial crimes (Crama et al., 2021). Verified information about beneficial owners enhances the scrutiny under which financial crimes can be detected and prosecuted. Transparency in beneficial ownership also signals to regulators, customers and the public that financial institutions are accountable and operate with high standards (Mugarura, 2017).

Accurate identification of beneficial owners is necessary to mitigating financial crimes such as money laundering, terrorism financing and other crimes. Ownership transparency ensures financial institutions are operating according to the law and act honourably (Mugarura, 2017). Reliable CDD processes with the support of accurate identification of beneficial owners helps the institution's compliance with laws, regulation and reducing criminal activities (Chitimira & Munedzi, 2023).

The procedure of identifying beneficial owners is often necessary in instances of intricate or veiled ownership structures, often requiring tedious investigations to ascertain the real beneficial owners

(Chitimira & Munedzi, 2023). Properly identifying beneficial ownership mainly supports regulatory compliance. It promotes the safety of the financial system, making it more secure and transparent.

In addition, by identifying beneficial owners, effective regulation is enhanced, and confidence in the integrity and stability of the financial system is considerably bolstered against financial crimes (Mugarura, 2017). Likewise, confidence in financial institutions is enhanced and their resilience and reputation insured since being able to identify beneficial owners creates the impression that the institutions employ prudence and observe rules (Mugarura, 2017). The true importance of beneficial ownership identification in the financial service innovations lies in FinTech's penchant for the secure and efficient digitalisation of business systems and financial transactions, as this reshapes the rhetoric of risk and compliance for useful purposes in the digital and interconnected financial world (Gaviyau & Sibindi, 2023).

Current State of Machine Learning in the UK Financial Market:

As per the survey conducted by Bank of England & FCA (2024), Machine learning (ML) is deployed extensively in the UK financial services sector, as it was reported that around two-thirds (72%) of UK financial firms (who participated in the survey) are either utilising ML in their systems or are in the process of developing a framework for their operations. This number leaves around 28 percent of the businesses that still do not utilise ML and BIG data capabilities. However, ML technologies' deployment is increasingly being used in processes related to AML, CFT and KYC, helping institutions improve efficiency, refine decision-making, and remain compliant with applicable regulatory requirements.

Despite widespread proliferation, ML penetration is not yet complete in the financial sector. This poses a risk but also an opportunity, since criminal and criminal-facilitating activities are themselves becoming ever more sophisticated, broadening and deepening the penetration of ML technologies in financial sector use cases will become ever more urgent. Introducing ML to the vast majority of institutions and use cases will help to realise the full potential of these technologies in enhancing beneficial ownership disclosure and verification, and in hardening all aspects of the financial system against illicit activities.

Current methods used for identifying beneficial owners and their limitations:

CDD in finance is one of the methods the defines how financial institutions can identify beneficial owners. According to Chitimira & Munedzi (2023), CDD means verifying the identity of who you are and the links that beneficial owners have in their business". In finance terms, CDD is important to determine beneficial owners since in most cases these beneficial owners are at high risk.

- Risk Based Approach (RBA): Risk based approach, for which regulators provide guidance, focus resources on those customers and transactions that are considered at the highest risk. By concentrating due diligence on the small, identifiable part of the population where financial crime

seems most likely, institutions can focus resources and increase the likelihood of identifying beneficial owners (Chitimira & Munedzi, 2023).

- Beneficial Ownership Registers: Such centralised data registries will record details of individuals with control or ownership over a legal entity. Under beneficial ownership registries, ownership is disclosed and registered across jurisdictions, thus improving the accuracy and efficiency of ascertaining beneficial ownership. Enhancing the ease of use of such registries can facilitate improved access (Chitimira & Munedzi, 2023).

Although it does have some limitations:

- Self-Reported Data: Customers also tend to report self-declared information during the onboarding stage, which can introduce inaccuracies or omissions in beneficial ownership data (Chitimira & Munedzi, 2023).
- Lack of Harmonization: The lack of systematised establishment of and access to information on beneficial ownership as a legal regulation across borders constrains operability in the international arena. Disparities in the requirements of different jurisdictions in disclosure pose a problem for international financial institutions attempting to adhere to the legal mandates (Chitimira & Munedzi, 2023).
- Cross-Border Operations: The differences in cross-border regulations can complicate due diligence processes, making it difficult for banks and other financial institutions to monitor complex subcontracting arrangements, particularly the problem of beneficial ownership (Jiao, 2023).
- Registry Accuracy: Beneficial ownership registries are indispensable tools for detection of money laundering and terrorism financing. Accuracy, timeliness and security of beneficial ownership information are essential to achieving the objectives of AML systems (Haenisch, 2024). Establishing mechanisms for keeping such registries up to date and safeguarding the confidential data underpinning them is paramount (Gilmour, 2020).

CDD, RBA and beneficial ownership registries can all be helpful tools, but none alone will successfully address the problem. To ultimately uncover beneficial ownership and curtail financial crime in all corners of the globe, it is crucial to strengthen the verification process, establish principles and harmonise regulatory frameworks, and standardise data management.

Challenges to effective identification of Beneficial Owners:

Complex ownership structures and cross-border entities, among other factors, present hurdles to identifying these beneficial owners. Furthermore, the employment of trusts, nominee shareholders and similar mechanisms make it challenging to pinpoint ultimate beneficial owners, which in turn makes it hard to follow the money and determine who benefitted from the movement of assets or the completion of a transaction (Mugarura, 2017). These issues are compounded by global financial networks, where these entities could be incorporated in a different legal regime where disclosure rules are less exacting (Feridun, 2023).

Inadequate harmonisation in the new beneficial ownership identification process across jurisdictions makes it difficult to combat financial crimes. Moreover, financial institutions are not able to comply with the requirements under the respective rules and regulations and accurately identify beneficial owners across jurisdictions when there is an inadequate approach to beneficial ownership (Feridun, 2023).

The *Danske Bank scandal,* about which it was reported in 2017 that 'at least $150 billion flowed through the Estonian branch of Danske Bank coming from Russia and other former Soviet states between 2007 and 2015 and was used for massive money laundering (Bjerregaard & Kirchmaier, 2019). The scandal revealed shortcomings in the bank's AML controls and the lack of regulatory oversight, which allowed criminals to move funds through the bank without adequately identifying the beneficial owners.

Another important case study is the *Panama Papers,* a data leak from the Panamanian law firm Mossack Fonseca, which managed offshore shell companies for numerous politically exposed persons, fraudsters, institutions, wealthy individuals and even organised crime groups behind economic crimes such as money laundering and tax evasion (Esoimeme, 2016). More specifically, the data leaks revealed conniving ways to obscure the core or the 'beneficial owners' of such complex ownership structures. It is no wonder that there is a growing focus on tracking down the ultimate owners of a corrupt network through meticulous investigation.

These cases illustrates the risks inherent in poor beneficial ownership identification and illustrate the importance of more rigorous AML supervision and due diligence checks. Through robust beneficial ownership identification and verification, financial institutions can better protect themselves against financial crimes and remain compliant with their KYC obligations.

1.1. Importance of Identification of Beneficial Ownership

Regulatory Compliance

Money laundering and terrorism financing pose serious threats to financial stability in the global economy and remain important questions for the control of crime across countries, financial institutions and corporations (Koker, 2006). However, many organisations tasked with detecting, preventing and punishing money launderers and terrorism financiers through Anti-Money Laundering (AML) and Combating the Financing of Terrorism (CFT) procedures continue to operate on electronic systems that are inadequate for the task at hand and incapable of identifying the real beneficial owners (Zavoli & King, 2021).

This ineffectiveness of AML regulations is due to the fact that the existing technologies are unable to help locate the real beneficiary/beneficial owners behind transactions, and big data technologies could be very important in helping to overcome these challenges, improve compliance and risk-management in financial markets by increasing the number of data points and identifying suspicious activities (Gaviyau & Sibindi, 2023).

This combination and development of improved AML performance through increased efficiencies includes asset seizures or forfeiture of the revenues from crimes and using seized funds as performance measures to assess performance of AML interventions, adopting international best practices in CDD and AML (Chitimira & Munedzi, 2023). Such reforms and development would enhance deterrence in compliance and risk management frameworks and the opportunity costs of using the proceeds from financial crimes.

Given the complexity and cross-jurisdictional nature of financial crime risks, a robust AML regulatory framework that involves beneficial ownership identification is urgently needed to reduce financial crime risks, such regulations and nomenclature regarding beneficial ownership are required (Feridun, 2023). Regulatory compliance relating to beneficial ownership is a prerequisite for all good AML and CFT regulations. Financial institutions are expected to report and identify beneficial owners as per regulations to mitigate risks of money laundering and terrorist financing (Koker, 2006). Therefore, financial institutions are required to embrace CDD and follow AML/CFT best practices across jurisdictions to be compliant (Chitimira & Munedzi, 2023).

For example, regulatory rules established at the European Union level directed all member states to provide national authorities with effective tools to identify beneficial owners to mitigate financial risks (Koster, 2020). Well-designed institutional settings allow banks to properly manage their risk portfolios, a key strategy being the obstruction of avenues of vulnerability to organised crime and terrorist activity (Chitimira & Munedzi, 2023). Regulatory approaches impose activity-based and entity-based checks on clients by the banks, especially those involved in activity or client relations linked to money laundering and terrorist financing (ML/TF).

Several regulatory bodies and frameworks are relevant to beneficial ownership identification:

- The Financial Action Task Force (FATF): It is an International policy making body that sets standards for combatting money laundering and terrorism financing. FATF stresses the need for more corporate transparency to tighten up financial crimes globally by making companies disclose the ultimate beneficial owners of legal entities (Chitimira & Munedzi, 2023).
- European Union: All EU member states must adopt and apply in full EU-initiated provisions transferred to national law by application of EU AML directives. AML directives mandate the identification of beneficial ownership and due diligence obligations with respect to customers (Mugarura, 2017).
- Financial Conduct Authority (FCA): The conduct regulator in the UK is responsible for over 58,000 regulated firms and financial markets, providing guidance on how an Ultimate Beneficial Owner should be identified as part of a firm's customer due diligence procedures to assess and monitor risks in line with AML and CFT legislation (Feridun, 2023).
- Basel Committee on Banking Supervision: A global standard setter for the prudential regulation of banks. It issues recommendations to improve banks' ability to identify beneficial owners of legal entities (Chitimira & Munedzi, 2023).

- The Organisation for Economic Co-operation and Development (OECD): provides guidance on best practices for combating financial crimes, including preventing 'anonymous companies'. OECD standards are mirrored in the UK regulatory frameworks that demand the identification of ultimate beneficial owners (Chitimira & Munedzi, 2023).

Together, these regulatory bodies and frameworks make up the UK regulatory environment for beneficial ownership transparency. Ensuring that every organisation complies with them will help assure that they are meeting AML/CFT legal and regulatory expectations as well as minimising financial crime risks.

Risk Management:

Accurate identification of beneficial owners helps mitigate financial risks. Correct beneficial ownership identification brings a reduction of business risk. On one hand, positive identification and monitoring of the beneficial owner can increase the efficiency of the entity carrying out business with the ultimate owner, the correct identification of the account holder in a bank or the actual owner of a company makes financial and business processes more effective.

As Koker (2006) points out, meeting the AML and CFT due diligence requirements of financial services providers discourages the involvement of innocent parties and associated money laundering and terrorism financing risk.

As also noted by Chitimira & Munedzi (2023), adherence to international best practices of customer due diligence and AML is also complied with, thus helping to ensure transparency, regulatory compliance and mitigating ML/TF risks.

Frimpong (2015) explains that accurate beneficial ownership identification contributes to organisational resilience regarding fraud and financial crimes, reducing corporate financial losses and reputational damage.

An accurate beneficial ownership register can improve risk management and reduce the downside of financial crimes such as money laundering or bribery. It helps to identify who is in control of a company and then see what steps should be taken in order to regulate and oversee that business.
Without accurate identification of beneficial ownership, financial institutions face significant risks, including:

- Reputation: Insufficient due diligence can result in a cumulatively degraded reputation for the host institution. As Koker (2006) comments, institutions that do not perform basic KYC and AML due diligence might create a dirty reputation in the eyes of their customers, regulators and the general public.
- Legal Sanctions: In addition to creating social instability and hampering economic growth, economic and financial crimes undermine the efforts of institutions and regulatory authorities to combat them. This could lead to legal proceedings, penalties and enforcements of regulatory orders based on the breach of AML and CFT regulations (Saddiq & Abu Bakar, 2019).

- Financial Losses: In addition to being exposed to the likelihood of financing terror, facilitating the flow of illicit money increases significantly the risk of huge financial losses that may result in the forfeiture of such assets, loss of business opportunities and negatively impact on shareholder value. Strengthening identification of beneficial owners will continue to guarantee the financial sustainability of these institutions (Chitimira & Munedzi, 2023).

But meeting regulatory requirements, performing credible due diligence and adopting best practices in AML/CFT risk management are necessary steps to avoiding these harms and protecting the integrity of the financial system.

Identifying beneficial owners with certainty remains the foundation for an honest financial ecosystem. if the regulators, prosecutors and courts know the true end beneficiaries of numerous financial transactions, then the financial system is more likely to grip money laundering, terrorism financing and other financial crime activities, with the result that the financial system retains the trust of its owners (Feridun, 2023).

Brexit has created ambiguity surrounding anti-money laundering rules, which are a precondition for regulatory compliance, and which the financial system needs for the safe and sound operation of the system in the light of the regulatory transformation (Mugarura, 2017).

1.2. Objectives and Research Questions

Primary Objective: This project will explore the use of big data and machine learning in relation to the identification of the beneficial owners of business accounts. It studies how these advanced analytical techniques could be improved in order to achieve higher accuracy, efficiency and consistency with regulatory requirements which can be taken as a best practice and have a reference point for emerging or less developed nations in terms of implementing stronger controls in their financial markets.

Secondary Objectives:

- Evaluate the effectiveness of current traditional methods.
- Identify the specific challenges faced in the identification process.
- Investigate big data technologies and machine learning algorithms that can address these challenges.
- Develop a practical framework for applying these technologies in the financial industry.

Primary Research Question: What do big data and machine-learning technologies hold for KYC, and the task of identifying beneficial owners?

Secondary Research Question:

- What are the limitations and challenges of traditional methods used for identifying beneficial owners?
- What big data technologies can be used to analyse and then integrate large sets of data in order to identify beneficial ownership?

- What machine learning algorithms can be designed to recognise the patterns and exceptions that signal beneficial ownership?
- What are the practical implications and benefits of using these technologies in real-world scenarios?
- Other than potential technological and logistical challenges, what are the downsides to adopting such advanced technologies?

1.3. Structure

This project will be structured as below:

Table 1: Structure of the Thesis

Chapter 1: Introduction	This section sets the research framework and lays out the problem space and identifying beneficial owners in the shadow financial world. The objectives of the research are then elaborated, and the remaining sections of the thesis are described.
Chapter 2: Literature Review	This chapter systematically reviews the relevant literature on beneficial ownership, covering the regulatory requirement, challenges in the conventional approach, and the cutting-edge technology (big data and machine learning) with potential solution.
Chapter 3: Methodology	This chapter will describe the applied research methodology, the methods for collecting and processing data, and the approaches to analysis. It will also cover the theoretical frameworks and models used in this study.
Chapter 4: Practical Showcase	In this chapter, some practical applications that have already emerged based on the application of big data and machine learning to the problem of beneficial ownership are showcased.
Chapter 5: Discussion	This chapter discusses the findings and the advantages and disadvantages of the methods proposed. it also shows the implications for supervisory compliance and risk management in the finance industry.
Chapter 6: Conclusion	The research results are summarised, along with an evaluation of their implications. The chapter concludes with recommendations for future research and practical guidance for the industry.

Source: Own representation

CHAPTER 2: LITERATURE REVIEW

With the help of literature review, this chapter describes the regulatory framework behind this effort, details the difficulties of legacy owner-due-diligence processes, and then dives into the possibilities of what big data and machine-learning are capable of.

2.1. Regulatory Requirements

This section provides an overview of the main regulatory regimes such as the European Union, the Financial Action Task Force, and the Basel Committee for enforcing transparency and accountability in relation to beneficial ownership to reduce money-laundering and terrorist-financing risks.

2.1.1. European Union

The EU is a major centre of financial governance, of AML and CFT in particular The EU's directives on banks and the wider financial system have a massive influence on making banks and other actors AML and CFT-compliant. This is particularly true when it comes to the economic impact of the UK post-Brexit.

Mugarura (2018) analyses how Brexit has impacted anti-money laundering regulation in the UK and how it is likely to impact money laundering regulation in the UK and EU, by looking at the law of EU Anti-Money Laundering (AML) Directives and their regulatory changes to the financial structures of the EU.

Chitimira & Munedzi (2023) puts stress on the legislative and regulatory framework provided by the EU in this area has played a significant role in shaping the UK's AML regulatory and enforcement framework, reiterating the fact that adherence to EU directives is essential for enhancing the AML regime in the UK.

Complying with EU regulations and directives helps financial institutions to improve risk management, increase transparency and contribute to a more secure and robust financial system for the EU. So far EU has published 6 AML directives, with each passing directive, EU thrives to make the controls and regulations around AML and CFT a lot more stringent:

Table 2: European Union AML Directives

Directive	Year	Regulation
1AML Directive	1991	Introduced measures to prevent the use of the financial system for money laundering. Focused on drug trafficking.
2AML Directive	2001	Extended the scope to cover a wider range of crimes and included requirements for customer due diligence and reporting suspicious activities.
3AML Directive	2005	Further expanded the scope to include terrorist financing. Introduced enhanced due diligence for Politically Exposed Persons (PEPs).
4AML Directive	2015	Increased transparency with stricter beneficial ownership requirements, enhanced customer due diligence, and new definitions for PEPs.

5AML Directive	2018	Expanded scope to include virtual currencies and prepaid cards, increased cooperation between Financial Intelligence Units (FIUs), and improved access to beneficial ownership information registers.
6AML Directive	2021	Harmonized definitions of money laundering offenses, expanded the list of predicate offenses, and introduced tougher penalties for AML violations.

Source: Own representation

Fifth Anti-Money Laundering Directive (5AMLD)

The 5AML Directive, (EU) 2018/843, issued by the EU marks a significant milestone on the regulations around Beneficial Ownerships. Understanding the depth of this concern, below are the snippets of the recommendations extracted directly from the official document (Directive (EU) 2018/843:

Key Recommendations:

- Increased Transparency for Corporate and Legal Entities: Member States must ensure accurate, verifiable, and up-to-date information on beneficial ownership. That information must be available to law enforcement, Financial Intelligence Units (FIUs) and to the general public.
- Beneficial Ownership Registers: Member States are required to establish central registers of beneficial ownership information. These registers shall be linked with each other in a common European Central Platform, and shall keep the aforementioned information for a period of 5 to 10 years following their dissolution.
- Trusts and Similar Legal Arrangements: Express trusts administered in Member States must maintain full details of settlor, trustees, protectors and beneficiaries. This information must be readily accessible to competent authorities, FIUs, and obliged entities.
- Access and Data Protection: Access might require registration (online), which may or may not involve a fee or fee schedule. Large-scale beneficial owners could be exempt from disclosure in exceptional circumstances (where non-disclosure might be detrimental to the beneficial owner).
- Mechanisms for Ensuring Accuracy: Mechanisms also have to be set up to ensure the integrity of central registers, while obliged entities must be forced to report discrepancies.
- Enhanced Due Diligence (EDD): EDD is required for high-risk transactions such as high-risk third countries (that are considered high-risk due to a higher risk of illegal activity) and implementing Enhanced Due Diligence measures, including a full customer and beneficial owner(s) check and increased monitoring.

Implementation and Compliance:

- Member States Responsibilities: These obligations to collect, retain and make beneficial ownership information available must be subject to a monitoring mechanism requiring each Member State to verify compliance. Appropriate and effective sanctions for non-compliance must be established.

- Deadlines for Implementation: They ordered member states to enact these measures by 10 January 2020, and they established new central registers of ownership for corporate entities and for trusts, to be incorporated by 10 March 2020. The 5AMLD, in transforming the EU's archaic system of corporate-entity identification into something capable of curtailing, as they put it, the use of legal 'avoidance and loopholes' that have enabled white-collar criminals from money-launderers to terror financiers, and many others to function, proudly claims that its intention is to eradicate the very loopholes in the law by which many are enmeshed.

The object is to fill the gaps in the EU's common legal framework, to enhance transparency, and stop the exploitation of shell companies and legal persons to launder illicit money and fund terrorism.

2.1.2. Financial Action Task Force

The global standard setter for Anti Money Laundering (AML) and Countering the Financing of Terrorism (CFT) regulations is the inter-governmental Financial Action Task Force (FATF) which, through its definition, is the natural person who ultimately owns or controls a customer or on whose behalf a transaction is conducted. The focus on beneficial ownership helps curtail the menace of Money Laundering and Financing of Terrorism (Chen, 2023).

All members need to ensure to implement FATF recommendations and make suitable amendments to maintain the integrity of their financial systems (Chitimira & Munedzi, 2022).

Cleaning up the ownership structures in the corporate sector is essential to make it difficult for criminals to benefit from and exploit vulnerabilities in the global financial system. Transparency of beneficial ownership and intergovernmental cooperation are imperative to fight money laundering and terrorist financing currently, and in the future (Haenisch, 2024).

Below are the recommendations extracted from the FATF Guidelines shared on Beneficial Ownerships (FATF, 2023):

Key Recommendations:

- Increased Transparency and Access to Information: FATF rules that information about beneficial ownership must be 'accurate, verifiable and up-to-date'. This information should be available to the general public. It should include the name of the beneficial owner, the country and the nationality.
- Member states would be required to cross-reference their central registers of beneficial ownership information. This information would need to be sharable and accessible across borders.
- Specific Requirements for Trusts and Similar Legal Arrangements: Records relating to the establishment and operation of all of a trust's management are to be maintained by and kept available to the competent authorities and FIUs in the member state in which it is established and to the obliged entities for purposes of CDD.

- Access and Data Protection: Beneficial ownership information should be made available subject to online registration and the payment of fees in the form of administrative charges. Public access should only be denied in instances where disclosure would represent a serious risk to the beneficial owner.
- Mechanisms for Ensuring Accuracy: When they find discrepancies between information in beneficial ownership registers and other readily available data, they should report these to the relevant authority, and correct them promptly.
- Enhanced Due Diligence (EDD): For high-risk cases (e.g., a political extremist, a place of business in a high-risk third country, a country recommended by an NGO for heightened attention), Enhanced Due Diligence procedures that include obtaining more information about the customer and beneficial owners, determining the intended nature of the business relationship, and monitoring to a greater extent.
- Implementation and Compliance: Compliance with the beneficial ownership information requirements is to be verified by member states, which are also charged with imposing 'effective, proportionate and dissuasive' sanctions in cases of non-compliance. The deadline for implementation of the measures was set for January 2020, together with timelines for the establishment of central registers for corporate entities and trusts.

Recommendations for International Cooperation

- Access Facilitation: Foreign competent authorities should be given access to beneficial ownership information in company registers.
- Shareholder and beneficial owner information exchange: An effective method of information exchange for shareholders and beneficial owners between member states should be established to facilitate international investigations.
- Avoidance of Restrictive Conditions: Information exchanges should not be accompanied by unduly restrictive conditions such as tax secrecy or bank secrecy provisions.

Included in these recommendations are new steps to increase the reliability, transparency and accessibility of beneficial ownership information for anti-money laundering, counterterrorism and other foreign policy objectives, particularly abroad.

2.1.3. Financial Conduct Authority:

The Financial Conduct Authority (FCA) in the UK has an important role which involves setting rules for business conduct in order to prevent financial crimes, such as money laundering, fraud and market abuse by implementing rules and regulations related to checking customers' identities as well as anti-money laundering (CDD and AML respectively). Ensuring that rules are met set by the FCA is paramount for financial institutions, which can avoid unnecessary risks and hence ensures the integrity of the UK economy (Haenisch, 2024).

The FCA puts emphasis on regulatory frameworks and regulation enforcement mechanisms, prioritising the fight against financial crime (Chitimira & Munedzi, 2023). The FCA promotes

transparency and accountability in financial deals which help eradicating the corrupt practices in the field of finance (Feridun, 2023).

The Financial Crime Guide (FCG) provides practical help and information for all sizes and sectors of firms that the FCA regulates in responding to the risk that they might be used to facilitate financial crime. Its contents draw heavily on their thematic reviews of firms, with additional material relating to other aspects of the FCA's financial crime remit (FCA, 2018):

Financial Crime Systems and Controls:

- Governance: Senior management should be clearly accountable for managing financial crime risks, including beneficial ownership. The firm's structure should facilitate co-ordination and information-sharing across the business.
- Risk Assessment: Ascertain what financial crime risks related to beneficial owners a firm is facing. This should include carrying out risk assessments, which should be done on a periodic basis.
- Policies and Procedures: Current policies and procedures that are appropriate to the business should be in place and accessible, effective, and known by the relevant staff, including policies for the identification and verification of beneficial owners.
- Staff Management: Appropriately trained staff, including those with the requisite knowledge of local trends in beneficial ownership, should be recruited and appropriately rewarded.
- Quality of Oversight: Regular, robust independent internal audits and compliance monitoring should seek to ensure that policies and procedures are being implemented effectively, including those relating to beneficial ownership.

Money Laundering and Terrorist Financing

- Perform Customer Due Diligence (CDD): Customer due diligence requires a firm to know its customer and beneficial owner, understand the nature and purpose of the customer's relationship with the firm, and identify and verify the customer. Firms must monitor customer transactions, and keep records.
- Enhanced Due Diligence (EDD): Firm must adopt EDD for all higher-risk cases: this involves seeking and attaining a more thorough understanding of the customer and associated risks and, in particular, the source of funds.
- On-going Monitoring: On an ongoing basis, monitor business relationships in a risk-sensitive way, ensuring that CDD information is kept up to date and that EDD is applied, where necessary.

Key Practices and Examples

- Good Practice: Carrying out detailed risk assessments, demonstrating commitment to internal audit and compliance functions and proper CDD/EDD functions.
- Poor Practice: It refers to inadequate risk assessments, insufficient internal audits, weak CDD and EDD processes and insufficient staff training on beneficial ownership risks.

Key Recommendations for Firms:

- Keep Information Correct and Current: Beneficial ownership information must be correct, verifiable, and up-to-date.
- Put in Place Robust CDD and EDD Measures: Conduct risk-based and enhanced due diligence as appropriate to establish beneficial ownership, when risk factors warrant such steps.
- Frequent Credible Risk Assessments: Risk assessments are needed often and should be updated as the business or regulatory environment changes.
- Encourage Governance and Oversight: Ensure senior management takes responsibility for the oversight of financial crime risks. Additionally, the board should receive good quality assurance and risk assessments from compliance and internal audit departments.
- Training: All staff who need to supply beneficial ownership and financial crime information should be trained to understand the rationale for doing so.

Such policies not only encourage transparency, but also meet regulatory requirements and lessen the risk of money laundering by beneficial owners.

2.1.4. Basel Committee on Banking Supervision

Basel Committee on Banking Supervision is an international committee of banking supervisory authorities comprising members from central banking governors of the 10 largest economies in the world (referred to as the Group of 10 or G10) that was created in 1974 to strengthen the resilience of the international banking system to promote financial stability. It aims to ensure sound banking systems across the globe (He, 2021).

Its rules, especially the Basel Accords, form the backbone of the way the world runs its banking, with implications for capital adequacy, risk management and prudential reporting on banking and finance the world over. Basel III introduced higher capital requirements, better liquidity requirements and improved risk management standards (He, 2021).

The Basel Committee's good practices on beneficial ownership are part of its wider approach on anti-money laundering and combating the financing of terrorism (AML/CFT). For banks, Basel's direction means that they must develop 'policies, procedures and controls' allowing for identification and verification of beneficial owners (Basel Framework 2020):

- Identify and Verify Beneficial Owners: Financial institutions must know the identity of beneficial owners (the people with control of a legal entity) and conduct robust due diligence to prevent undue influence and opacity in the beneficial ownership structure.
- Refusing Relationships with Persons of Suspicious Character: Banks should not open or continue business relationships with a client where the beneficial owner or person exercising effective control is suspected to be involved in narcotics trafficking, terrorist financing or money-laundering activities.

Implementation and the Three Lines of Defence Model:

The Basel Committee urges uptake of the 'three lines of defence' framework for managing financial crime risk:

- First Line of Defence: Operational management not only accounts for internal controls but also coordinates risk management activities relating to beneficial ownership on a day-to-day basis. First Line of Defence: Background checks are conducted and accounts are monitored for signs of suspicious activity. Second Line of Defence: Managers analyse reports and determine whether additional preventive measures are needed.
- Second Line of Defence: the risk management and compliance functions oversee and facilitate the proper, effective implementation of risk management processes, including continual CDD and, where appropriate, EDD.
- Third Line of Defence: Internal audit provides an independent verification of the effectiveness of the first and second lines of defence to provide assurance that the bank's AML/CFT policies and procedures have been effectively implemented and are being followed.

It guarantees that all banks are not only carrying out their AML/CFT obligations, but also that their risk systems are operating properly. Furthermore, with requisite safeguards, supervisors must work together with counterparts in other jurisdictions, and may need to share information or otherwise collaborate when identifying and reining in cross-border financial crimes.

These references confirm the relevance of supranational bodies such as the Basel Committee in providing international standards for financial regulation. The adoption of these standards plays a pivotal role for regulatory agencies to avert reputational and legal risks for them, as already happened in the US with the FDIC against the non-compliant banks (Wu & Salomon, 2017). The suggestions and direction that recommend the basis for the proper supervision of the financial sector itself become fundamental in the prevention of crimes such as money laundering (Valvi, 2023).

2.1.5. Organisation for Economic Co-operation and Development (OECD)

The Organisation for Economic Co-operation and Development (OECD) is an international organisation that works with member states to promote better economic growth, integrates environmental and sustainable development under a sustainable development committee and meets the challenge of financing the UN sustainable development goals to share common standards on aspects like money laundering and terrorism financing (Canhoto, 2021; Saddiq & Abu Bakar, 2019).

It advises member states on strengthening the international rule of law by improving international co-operation in the prevention of financial crimes. In advising on the national implementation of international standards, it proposes broad recommendations and legal guidelines for the world financial system to become more transparent, pure and accessible (Chitimira & Munedzi, 2023). These new OECD standards will promote transparency, combat money laundering and tax evasion by involving appropriate authorities in preventive actions. The OECD has a global mission to support countries and financial institutions to identify ultimate beneficial owners (Saddiq & Abu Bakar, 2019).

The recommendations aimed to support both national governments and private businesses to outline a set of legal and operational standards that could strengthen the resilience of financial systems against money laundering and the financing of terrorism, and enhance its transparency and security of the global financial system. These enable national governments and jurisdictions the world over to protect their fiscal integrity, increase predictability and stability, and reduce the risk in the international financial system (Matsuoka, 2020).

It is now the OECD that is the pace-setter for fostering responsible behaviour in business and tackling financial crime in the private sector, through the development of recommendations for countries and the private sector, and the provision of guidance on the design and operation of anti-money laundering measures intended to ensure that the global financial industry is sound, transparent, resilient and trusted.

2.2. UK's Approach

The implementation of a CDD regime due to anti-money laundering requirements in the UK is part of the general entrenchment and strengthening of the anti-money laundering regime in the UK (Chitimira & Munedzi, 2022). These measures are part of a progressive regulatory framework aimed at combating money laundering in financial institutions and markets (Chitimira & Munedzi, 2022). Additionally, The Brexit decision has led to the need for further development within money laundering regulations of the UK (Mugarura, 2018). Therefore, these efforts are contributing to its transparency and accountability by attempting to eliminate such crimes, which are traced to dealings in financial markets and institutions.

With respect to Brexit and its impact on the AML regulations in the UK, Brexit means to leave the European Union, but it also means to leave the core international standard setting directives and best practice guidelines on AML (the FATF recommendations have been adopted in the UK). The current neoliberal thinking on money laundering contrasts with a social democratic approach to money laundering (Mugarura, 2018).

Customer due diligence measures are essential in the UK, particularly in cases involving suspected money laundering activities or unreliable customer information Chitimira & Munedzi (2022). Even with simplified customer due diligence measures, financial institutions are still required to monitor all their customers' activities (Chitimira & Munedzi, 2022). The Financial Action Task Force (FATF) stresses the importance of beneficial ownership transparency through customer due diligence processes (Gilmour, 2020).

- Customer Identification: Regulated financial institutions are required to conduct regular know-your-customer (KYC) reviews on corporate clients to prevent money laundering and the financing of terrorism Haenisch (2024).
- Beneficial Ownership Identification: FATF promotes beneficial ownership transparency by urging institutions to verify customers' identification and financial transactions through the Customer Due Diligence (CDD) process (Gilmour, 2020). The Money Laundering Regulations 2017 mandate

comprehensive customer due diligence measures to be conducted on high-risk customers in alignment with FATF recommendations.

2.3. Challenges in Traditional Methods

Current regimes for determining beneficial owners face significant obstacles presented by complex ownership structures and manual processes, which hinder the ability to identify ultimate ownership and contribute to market inefficiencies and vulnerabilities in both deterring and combating financial crimes.

2.3.1. Complex Ownership Structures

The challenges of using traditional approaches to identify beneficial ownership in the UK are manifold, and could make anti-money laundering efforts less effective. some of the main challenges are:

- Concealed or complicated ownership structures: conventional due diligence procedures might prove inadequate where a middleman or dummy company is used to create a layer of obfuscation and to shield ultimate beneficial ownership. This is especially perceptible where the layers may span over various jurisdictions or where shell companies and trusts are used offshore. Unveiling the ultimate or beneficial owner in some of these complicated ownership structures may require the use of additional due diligence steps (Gilmour, 2020).
- High-Resource Labour Intensiveness: The identification of beneficial ownership through traditional means is high-resource labour-intensive. The manual review of the ownership information and the verification of beneficial ownerships require considerable human resource and financial input, hence the operational inefficiency (Feridun, 2023).
- Identifying beneficial owners could be done through more traditional approaches, but these could have data privacy and security issues. Gathering certain types of information and then sharing data across institutions could have data privacy and data security issues (Mhlanga, 2024).
- Reluctance to Share Information: Firms might be unwilling to share sensitive ownership information, either due to concerns for the security of their data, privacy, or competitive advantage. This could pose an obstacle to the identification of beneficial owners. Regulators must also be aware of this problem in the context of due diligence that requires firms to identify their beneficial owners (Feridun, 2023).
- High false positive rates: Conventional BOID identification tools might render high false positive rates, i.e., potential transactions or entities that are not suspicious but are flagged as such by the due diligence process. In the worst case, this can lead to inefficiencies in the due diligence process and interfering with the identification and prosecution of high-risk cases (Kurniabudi et al., 2019).
- The lack of a standardisation of processes and guidelines regarding beneficial ownership identification opens a door against standardised compliance. This is a challenge that many

financial institutions face as companies have to deal with different standards in identification of beneficial owners depending on the jurisdiction and the entity (Chitimira & Munedzi, 2022).

Case Study: Panama Paper Leak 2016 (Esoimeme, 2016)

Panama Papers case study relates to huge data leakage from Panamanian law firm Mossack Fonseca documenting massive money laundering, tax avoidance and corruption via offshore entities on an unseen scale before. Evidence from the Panama Papers showed that the world's most wealthy individuals and many other influential public figures were running their financial activities through offshore company structures.

But it required the massive leak that spurred the Panama Papers story in 2016 to bring the new forms of offshore intermediation and complex ownership structures to the attention of many members of the public. A total of 11.5 million documents from a Panamanian legal practice, Mossack Fonseca, reporting the business activities of more than 300,000 offshore clients, including politicians and presidents, celebrities, professional sports stars and drug lords became accessible to the general public. The size of the data leak is unmatched to date.

Among the public outcries, investigations, forced resignations, lawsuits and litigations triggered by the Panama Papers, which eventually led to convictions, suspected tax evaders and criminal activities, no country in the world had fewer problems than Panama itself.

Many of the problems and challenges involved in identifying beneficial ownership and, more generally, tackling money laundering become clearer through examination of the Panama Papers case study. These include:

- Complicated ownership structures and offshore entities: The use of complex chains of ownership and offshore entities to bury ultimate beneficial owners. In many cases it was almost impossible to tell who the human beings behind the complex structures were. The leaks made the case for reform more persuasive because they showed where enhanced due diligence is needed.
- Tax Evasion and Money Laundering: The documents revealed how the financial system allowed offshore accounts and shell companies to structure transactions and conceal evidence about tax evasion, money laundering offences and other financial crimes.
- Transparency and Accountability: The Panama Papers scandal showed how the international financial architecture contributed to lawlessness since it did little to prevent potential offenders from using the system to engage in criminal behaviour and poor governance. Therefore it demanded action to amend the rules so that global finance is reviewed and regulated with an appropriate level of scrutiny.
- Regulatory Reform: In the wake of the Panama Papers leak, a number of countries enacted regulatory reforms to improve beneficial ownership transparency, increase anti-money laundering supervision. The case became a catalyst for regulatory reforms that reduced tax evasion and money laundering in the process.

In short, the Panama Papers case study moved the global financial system one step closer towards identifying the chink in its armour and taking effective actions to plug that chink in an attempt to adopt transparent, accountable and compliant approach to the high-sounding supporting regulatory frameworks. The Panama Papers blow received back the spotlight on the issue of beneficial ownership disclosure, the effectiveness of AML controls and due diligence, in the global fight against money laundering and the enhancement of integrity of the global financial system.

These challenges cannot be tackled in any simple way without turning to more nuanced and technologically informed methods of beneficial ownership disclosure. Greater data analytic use and the more frequent application of machine learning techniques by the private and public financial sector could better identify beneficial owners in a more accurate and efficient manner than is currently possible for these institutions. In doing so, financial institutions will be better equipped to achieve a more robust anti-money laundering regime, and create greater financial transparency for their clients.

2.3.2. Manual Processes and Their Limitations

The challenges surrounding traditional methods of identifying beneficial ownership in the UK are substantial and can hinder the effectiveness of anti-money laundering. Here are the main issues with traditional methods:

- High Maintenance Investigations: Ineffective methods of beneficial ownership identification can involve manual processes, making for a high-maintenance procedure that requires expensive and time-consuming human support to conduct thorough investigations to find beneficial owners (Apene et al., 2024).
- Fragmentation: Beneficial ownership data might be fragmented across different databases and channels, which could impede efforts to consolidate and analyse this data in a single place, leading to inaccurate detection of beneficial owners of a particular entity. This could also result in errors and inconsistencies in the system (Apene et al., 2024).
- Scalability challenges: Traditional methods have a capacity limitation to scale to large volumes of data and complex ownership structures. With an increase of quantity and complexity of data, traditional methods could be overwhelmed and identifies would be delayed (Apene et al., 2024).
- Time Consumption: Manual approaches to beneficial-owner identification can be extremely time-consuming, specifically with complex ownership structures or cross-border entities. The time that is needed to properly investigate and verify beneficial owners can delay the compliance process and cost the business more time and resources (Apene et al., 2024).
- Data Privacy Concerns: Gaining access to and sharing information about ultimate beneficial ownership may increase the risk of data privacy and security breaches, potentially involving personal or confidential information (Apene et al., 2024).
- High Error Rates: Manual processes are susceptible to human error. Moreover, in a system that calls for the identification of beneficial owners, high error rates will create significant false positives

or false negatives. Either way, anti-money laundering (AML) efforts and the broader compliance endeavours are compromised (Apene et al., 2024).

- Limited Data Access: As traditional methods of detecting UBOs have largely been developed before the widespread availability of beneficial ownership data, they are likely to be less effective at using it. Limited data access means that traditional methodologies lack the necessary inputs to verify the BOD accurately, and thus run a greater risk of missing information (Apene et al., 2024).

The latter mitigates the former, as do financial-industry computational and data-based approaches to more dynamic beneficial ownership identification. Robust real-time data analytics, artificial intelligence and automation will empower financial institutions to more quickly and accurately identify beneficial owners, reducing the exposure of the global financial system, and individual firms, to the threats of corruption, money-laundering and other legal and illegal financial activities.

2.4. Technological Advances

This section reviews the role that big data and machine learning are playing in helping to identify beneficial owners. The use of big data and machine learning can automate detection of financial crime through applying many automated processes and advanced analytics to millions or even billions of data points.

2.4.1. Overview of Big Data and Machine Learning

Big data (a term used to describe the ever increasing data rendered into a structured and unstructured form in which firms operate each day in their day-to-day operations) is indeed a big driver of innovation in the both the private and public sector environments (Baru et al., 2013). In the financial technology space, big data has been the catalyst for new financial products and services, new-age risk management practices, and new-age operations capabilities for promoting new-age banking products (Mhlanga, 2024). In short, big data has been the game-changer for re-inventing the traditional banking products and services using modern-day technologies and capabilities in the financial sector. This has further expanded the horizon of financial inclusion and fished more and more people and businesses from poverty and financial exclusion.

Big data is needed to support machine learning and knowledge inference through a comprehensive pipeline leading to data pre-processing, mapping, fusion of high-dimensional data sets, extraction of new knowledge, and visualisation of new knowledge in terms that the user can easily understand (Holzinger, 2017). This shows that big data is needed to support advanced analytics to gain insights from big data. Furthermore, other cutting-edge research about machine learning and deep learning, both examples of data science, further accentuate the need for big data in research and innovation in artificial intelligence (Chahal & Gulia, 2019).

The regulatory implication of big data is even more significant. Financial crime driven countries, financial institutions and companies are expected to protect themselves from money laundering and

financing terrorism by complying with anti-money laundering (AML) and combating the financing of terrorism (CFT) regulations. These regulations heavily rely on analysing big data to mitigate illicit financial transactions (Koker, 2006).

It is known that big data, coupled with advanced technologies such as cloud computing, is able to help organisations overcome challenges related to data protection regulations, such as the General Data Protection Regulation (GDPR) (Labadie & Legner, 2023a). This shows how big data, when combined with advanced technologies, can enable organisations navigate regulatory frameworks and meet the requirements for strict data protection standards. Moreover, the application of NLP to big data, carry out thematic analysis in the financial sector to gain insights on how organisations are meeting financial goals through efficient use of data potential of big data in financial sectors to achieve financial inclusion and other important goals (Sharma et al., 2023).

Thus, it is essential for anomaly detection to define anomaly as a deviation or fluctuation beyond an understood behaviour or normal patterns. Consequently, big data can be applied in network anomalies mitigation to enhance cyber security systems (Kurniabudi et al., 2019). Using big data analyses of traffic anomalies can be an efficient way to enhance cyber security by protecting networks systems from traffic anomalies. Lastly, it is worth noting that utilising big data in project management, specifically in GDPR project implementation, points out to the significance of big data in transformation of business models and highlights the strategic importance of data in modern business ecosystems (Todorović et al., 2018).

In addition, the variety of industries that now use big data reflects its role in modern business. Financial crime, fraud detection and money laundering are all examples of important security-related issues that have been enhanced by the use of big data with artificial intelligence solutions involving machine learning algorithms (Apene et al., 2024). Given the increasing range of sophisticated criminal threats to the security, it is vital that analyses, such as this one, show how big data can help protect societies by improving law enforcement activities.

At the same time, the examination of customer due diligence and anti-money laundering (AML) systems in the UK demonstrate how big data can support regulators' efforts to prevent financial crimes (Chitimira & Munedzi, 2023). Through leveraging data analytics and monitoring methods, regulatory bodies can more effectively oversee and take actions against harmful activities, thereby strengthening the overall resilience of financial system to criminal activities. Lastly, an overview of beneficial ownership identification systems from the AML perspective, paints a picture of AML being complex, too, but also emphasises the importance of adopting data analytics to tackle new challenges that arise from the changing global financial landscape (Chen, 2023).

Big data must be the core for evidence-based decision-making in the 21st century since, in all sectors, companies are using big data to make better decisions, to increase operational effectiveness and reduce risk.

Machine learning, a subset of artificial intelligence, is an approach to technological problem solving wherein insights are discovered from (and predictions are made based on) patterns and relationships in data sets (Chahal & Gulia, 2019). Machine learning applications are developed by crafting algorithms that teach computers to learn from data and make decisions or predictions without being explicitly programmed (Holzinger, 2017). Machine learning techniques are used to uncover underlying relationships in data, automate decisions, and increase the accuracy of predictive outcomes.

A key characteristic of machine learning is its ability to analyse large sets of data with the objective to extract patterns that are invisible to humans (Chahal & Gulia, 2019). For example, using algorithms or statistical models, a machine learning algorithm is capable of analysing large data sets and derive decision-relevant knowledge to assist people in data-driven decision (Holzinger, 2017). Because of its ability to analyse large datasets, machine learning is often applied in solving problems in finance, healthcare, marketing, and cybersecurity, etc.

Moreover, anomaly detection (especially in the domain of cybersecurity) is an important application of machine learning (Najafimehr et al., 2023). Machine learning based detection approaches take network traffic patterns and respond to network anomalies to improve organisations' security posture and anomaly detection (Najafimehr et al., 2023). Through machine learning-based detection approaches, organisations can identify abnormal traffic and prevent threats.

Machine learning has recently assisted detecting financial crimes by analysing transactional data and identifying fraudulent patterns as indicators of money laundering and financial crimes. In the context of financial crime risks, machine learning has advanced the use of smart compliance such as fraud detection and anti-money laundering (Apene et al., 2024). This showcase the potential of machine learning that can be utilized in enhancing regulatory compliance and risk management practices within the financial sector.

Furthermore, machine learning along with big data analytics can accelerate the use cases in anomaly detection and predictive modelling (M & S, 2024). Organisations will be able to discover hidden patterns in large amounts of data with the help of artificial intelligence and machine learning techniques (M & S, 2024). Furthermore, it eases the anomalies detection with higher accuracy and thus speed up decision making process (M & S, 2024). This synthetic use of big data analytics with machine learning highlights how such technologies could be employed to drive innovation and improve efficiency in all sectors..

Machine learning is, moreover, integral to natural language processing (NLP), which is the ability to understand, interpret and write human language (Sharma et al., 2023). Using machine learning models to train computers to accomplish various NLP tasks (such as sentiment analysis, language translation or text summarisation), businesses can use unstructured text data to enhance customer support and automated communications, including those involving RPA-equipped robots, as well as to create more effective information retrieval systems (Sharma et al., 2023).

To conclude, machine learning is at the cutting edge of technological innovation. Using sophisticated algorithms and tools, machine learning will enable people to distil insights, make predictions, and automate decisions in virtually every domain of human expertise.

Relevance to Beneficial Ownership Identification

Jiao (2023) notes that practical analysis of the big data techniques demonstrated obvious advantages in the prevention and control of money laundering, including more accurate and timely analysis and effectiveness, a smoother process of regulatory compliance, and enhanced international cooperation among banks and regulatory bodies across the borders.. Machine learning algorithms can process massive amounts of data to look for patterns and abnormalities that can help identify suspicious beneficial ownership activities (Apene et al., 2024). big data can be used to monitor regulatory obligations under laws such as KYC and AML (Mhlanga, 2024).

Using machine learning and big data analytics, organisations will be able to reveal beneficial owners much more efficiently and accurately, through powerful algorithms that can analyse large data sets to derive insights and patterns on beneficial ownership. Machine learning and big data analytics can play an important role in enhancing due diligence practices and other regulatory compliance activities for organisations. This is because these methodologies can aid in identifying beneficial ownership linkages at a faster and more accurate pace than what was previously possible (Apene et al., 2024).

Machine learning and big data analytics are helping to enhance anti-money laundering verification activities and regulatory (and law enforcement) oversight on many fronts, making beneficial ownership details more transparent and traceable across national boundaries. As a result, the chances of detecting and preventing money laundering are getting more accurate and timely; compliance efforts are facilitated by improving specificity, and cross-border enforcement can be expedited through joint efforts between financial institutions and their regulators, especially on governance, surveillance and reporting (Jiao, 2023). Similarly, the technical and contextual affordances of machine learning algorithms are foundational to advancing effective efforts to curb money laundering and terrorism financing (Canhoto, 2021).

Similarly, machine learning in beneficial ownership identification can assist in bringing automation to the analyses, as the tasks can often be repetitive and mundane, pressing right buttons could save a lot of time and effort. Further, machine learning implementation can help organisations in ensuring efficient compliance procedures by automating the detection of whether specific forms were filed and if properly reported (Apene et al., 2024). Big data analytics can be used to manage risk, comply with regulatory requirements such as Know Your Customer and Due Diligence (Aderemi et al., 2024), set up control mechanisms to detect and prevent fraud, and employ algorithmic trading to speed up and enhance financial services. International best practice (such as customer due diligence measures set out by regulatory bodies like the BCBS and FATF) is an important tool in the global response to money laundering (Chitimira & Munedzi, 2022).

2.4.2. Previous Research and Applications in Financial Services

With its integration with Big Data Analytics, the financial services sector has experienced transformational changes, which it applies in managing financial assets, and other operational processes (Aderemi et al., 2024). The strength of Big Data lies in its ability to allow institutions to analyse complex data sets analyses, draw insight from past trends, and make predictions about future outcomes that have pushed innovation that enhances experiences for customers and optimise operational efficiency (Mhlanga, 2024). In financial services, big data Analytics allows for risk management and fraud Detection. By analysing the huge datasets generated over time by the firms, it allows institutions to identify irregularities in transaction patterns, pre-empt fraudulent transactions, and reduce financial risks (Abrol, 2023). The major objective of utilising Big Data Analytics is to result in the production of new and innovative financial products, and improve customer experience, which enhances organizational revenue, sales, and profitability.

Furthermore, big data analytics provides opportunities and challenges for decision making in the financial organisations (Aderemi et al., 2024). It involves in enhancing the insight of the market information for the strategic decision making but there are some challenges include in the data quality, privacy and constitution has to cover before use big data analytics (Aderemi et al., 2024). Resolve these challenges provide the big data analytics for the empowerment organization allows to move more success in their business.

Big Data analytics is having an immediate and substantial impact on portfolio management in the investments sector (Abrol, 2023). Big Data allows for the most informed investment decisions to be made with the use of available information (Abrol, 2023). Big Data allows for the management of portfolios and the definition of stocks, the financing, and trading models (Abrol, 2023). This helps us to understand the vitality of Big Data in making informed investment decisions and in creating a vital investment strategy for financial institutions.

Moreover, big data technology has redefined the financial services market competition by providing tools to reduce fraud, risk aversion as well as new business models (Aderemi et al., 2024). Big Data analytics for financial services can reveal previously unseen opportunities for companies, lower costs and improve operational efficiency, especially for payments services (Aderemi et al., 2024). This transformation shows that big data technology has the capacity to make banking and financial services easier, safer and more comfortable for consumers.

Big data is also a key tool for identifying, investigating and preventing fraud, due to the incorporation of several machine learning algorithms in the process (Canhoto, 2021). In fact, beyond transaction monitoring, the use of big data analytic systems and algorithms is critical to enable banks, for instance, to comply better with rules that combating money laundering and terrorism financing, among other financial crimes (Canhoto, 2021). The prevalent use of big data analytics reaffirms the concept that it is paramount for us to continue exploring ways to utilise this technology to further

boost security within the financial system, in order to discourage engagement in money laundering activities and other forms of financial crimes.

To summarize, it can be deduced that with Big data analytics, the field of financial services has been revolutionized by the reshaping of transactional and customer decisions, and behaviour patterns and trends (Aderemi et al., 2024). Through the utilization of Big Data, financial organizations can increase operational efficiency and risk mitigation, innovate, advance customer relationships and open new avenues for growth (Abrol, 2023). Although issues of data quality and privacy pose challenges for organizations interested in the utilization of Big data, the existing benefits and opportunities outweigh any obstacles, with financial organizations being poised for an era set within a rich data-available environment.

Applications of Machine Learning

In the banking sector, machine learning is already transforming payment transactions and improving services such as credit scoring, algorithmic trading, fraud detection and reduction (Holzinger, 2017). ML algorithms could analyse the amount and patterns of transactions to identify suspicious activities (Chahal & Gulia, 2019). This proves how machine learning (ML) has a major effect on improving risk management, reducing financial crimes and overall improving the quality of decisions regarding financial services in the society.

ML technologies are used by banks and financial institutions for credit scoring, the assessment of the creditworthiness of individuals and organisations to evaluate their ability to repay debts. ML helps financial institutions make more accurate and efficient credit decisions. In their analysis of credit risk in credit scoring, (Holzinger, 2017) highlights how ML 'can utilise massive datasets to implicitly capture the effects of risk factors that one might miss in manual credit score design'. As such, ML can speed up credit assessments, while improving the effectiveness of credit-lending decisions and reducing credit risk.

Moreover, ML methods have also been used in algorithmic trading, whereby decision-making regarding trading processes, such as identifying market data patterns and making optimal trades at point in time and price of trades are automated via ML methods (Holzinger, 2017). Financial institutions can develop or refine more sophisticated trading algorithms based on the ML analysis of historical data, market trends and real-time data, thereby facilitating the innovation and efficiency of financial markets.

To detect fraud in the financial sector, ML algorithms are crucial for explaining frauds in order to prevent frauds (Holzinger, 2017). The ML algorithm looks into the anomalies and frauds to the company so that any possible fraudulent transaction can easily be detected. Such a system allows the company to stay safe from frauds and other financial-related crimes. This explains how ML can be beneficial in improving the security of the financial system.

ML can be directly integrated into financial systems, giving institutions a deeper understanding of customer habits and transaction patterns so they can offer personalised services (Chahal & Gulia,

2019). With ML algorithms that access their customer data, financial institutions can figure out their preferences, predict future behaviours, and customise services to fit individual needs. This customised approach would boost customer satisfaction, build loyalty and stim financial services industry.

Credit scoring, algorithmic trading and fraud detection would not without ML. There are no reasons to believe that ML cannot be equally transformative in revamping other aspects of financial services. By leveraging ML algorithms, banks and other financial institutions will be able to more accurately predict the risks associated with specific clients, tweak decision-making processes to meet their clients' needs, and adopt new ways to interact with their clients in an effort to provide them with customised services.

Case Studies: Danske Bank Scandal

The Danske Bank money laundering case starkly shows major weaknesses in the current system of combating financial crime and that greater use of cutting-edge technology such as big data analytics and machine learning could advance practices of risk management and compliance with regulatory requirements (Maulidiyah, 2023). Had the Danske Bank had machine learning functionality, it could earlier have identified transaction patterns and therefore mitigated the damage and ensured its ability to prevent financial crimes (Pontes et al., 2021).

Danske Bank’s example highlights the role of cutting-edge technologies in strengthening anti- money laundering and the enhancement of regulatory compliance of financial services. Through using big data analytics, firms can extract insights about transactional and behavioural patterns of customers and become more knowledgeable about potential risks. By utilising artificial intelligence in large data aggregation, data analysis can detect irregularities and suspicious activities among the cases of money laundering (Gilmour, 2020). Furthermore, the machine learning algorithms play a significant role in the detection of frauds, benefiting the customers and the security of the transactions through the screening of the suspicious customers and the illegal activities (Mhlanga, 2024).

Furthermore, financial institutions can use big data analytics and machine learning to enhance their risk management and compliance functions, allowing them to better manage due diligence, flag transactions in real time and monitor for money laundering activities (Apene et al., 2024). This paradigm shift can be proactive in preventing financial crime and can ensure the stability of the financial system, help financial institutions maintain regulatory structures, and promote consultation and co-operation between institutions and the state itself.

Clearly, the Danske Bank case emphasises the need for Beneficial Ownership identification in order to stop money laundering and other illicit transactions. Machine learning and big data analytics can enhance the management of beneficial ownership structures in terms of identifying beneficial owners, suspicious transactions and anti-money laundering requirements as a whole (M & S, 2024).

Big data analytics and machine learning could automate due diligence processes, simplify regulatory reporting and enhance compliance (Feridun, 2023). The financial scandal at Danske Bank has raised awareness of the potential of big data analytics and machine learning to transform the financial sector's approach to financial crime prevention, fraud detection, regulatory reporting and compliance. Leveraging the power of data science, policymakers and private organisations can enhance their detection and prevention capabilities of financial wrongdoing, and contribute to a more transparent and secure financial sector.

Future Directions

The application of big data and machine learning in financial services will continue to evolve, with further potential to enhance regulatory compliance, risk management and operational efficiency in financial services. New research areas are underway that investigate the application of novel technologies (such as blockchain and alternative analytics) to further transform financial services and support beneficial ownership identification (M & S, 2024).

This movement will lead to a significant revolution in how financial services implement their regulatory compliance practices by leveraging big data and machine learning. It is now possible for a financial services organisation to monitor and enhance governance to regulate and reduce risk exposures to compliance issues such as fraud and regulatory changes. Moreover, these organisations can lead the change and embrace data analytics tools to control and mitigate regulations and new compliance remediation programmes published by the regulators.

Moreover, blockchain applications in financial services can help increase the transparency and security of beneficial ownership identification verification procedures. Possessing immutable and decentralised properties, the technology can be used by organisations to develop beneficial ownership verification systems that can enable faster and more secure sharing of information while increasing stakeholder trust. The same principle can be extended to other types of formal procedures where the reliability of information ranking is crucial. For example, it has been proposed by experts to use blockchain by enabling the real-time ranking of terms and conditions on airline ticket pricing. Blockchain applied to advanced analytics can enhance this accuracy and improve the automation of decision processes. Further, supporting the argument for machine-decision making, blockchain can also help mitigate the probability of fraud.

The intersection of big data analytics, machine learning and cloud computing is expected to spur growth in financial services by driving an increase in operational efficiency and improving innovation. Cloud-based data management can accelerate data processing, increase scalability and facilitate the sharing of data, thus driving improved operational efficiency, increased speed in decision-making and achieving a competitive advantage. Machine learning algorithms working on the cloud infrastructure can help organisations gain insights and make better decisions from vast amounts of data at a substantial speed, allocate resources more effectively to spur business growth, and generate insights into customer behaviour.

If embraced correctly, big data and machine learning can be revolutionary in the way regulatory compliance and risk management are applied, and thereby improve operational efficiency in financial services.

CHAPTER 3: Methodology

This chapter describes the research methods, which provides a literature review of existing practices, emphasises challenges of traditional research methods, identifies the opportunities of big data technologies and machine learning to identify the beneficial owners, presents a case study and the tools used to collect and analyse data.

3.1. Examination of Traditional Methods

Figuring out beneficial ownership is a challenge and traditionally involves a review of relevant documents, compliance with applicable rules and regulations, and adherence to industry best practices. The paper contains a comprehensive document review process, including materials from the regulator Financial Conduct Authority (FCA) in the UK, the Financial Action Task Force (FATF) and industry reports (Basel Framework, 2020). Current preventive rules, exposed in the papers, are insufficient, given the current dynamics of financial network structures.

Transparency and Compliance: Central here is the principle of transparency. Transparency has been, and remains, one of the key benchmarks by which financial regulators. For example, the Financial Action Task Force (FATF) assess whether adequate defences against criminal behaviour are in place in the operations of their member states' regulated banks, securities brokers and other financial institutions. Central to this framework are requirements that banks and other financial intermediaries undertake a rigorous and thorough Customer Due Diligence (CDD) of their customers. But it is also required to be able to assess the efficacy of CDD in practice in order to understand whether these approaches are effective or not (Gilmour, 2020).

- Technological Integration: It seems that a combination of old-style tools like identification and verification of beneficial ownership and new technologies can provide an increased reach and accuracy, by bringing the two approaches together. Apene et al. (2024) proposed a knowledge framework, which envisions a harmonious future. A combination of old approaches with new technological approaches could provide a more comprehensive solution.
- Anomaly Detection: On the one hand, techniques for anomaly detection for beneficial ownership disclosure need to aim for high detection rates and, on the other hand for lowest possible false alarms. While standard outlier-detection methods can work well in the case of false alarms and high detection rates, they might give unsatisfactory results when targeting strong anomalies, requiring a significant leap forward in anomaly-detection research (Kurniabudi et al., 2019).
- Machine Learning and Community Detection: One area that has shown fresh promise for finance is the use of machine learning, which has evolved to include supervised and unsupervised machine learning methods, whose implementation can improve the identification and verification of beneficial ownership (Warin & Stojkov, 2021). The Louvain algorithm for community detection provides fast and accurate methods for finding complex ownership networks More generally, the research literature provides further information on ownership transparency (Zhang et al., 2019).

- Impact of Regulatory Changes: Regulatory changes, for example also required by Brexit, drive adaptations in identification approaches based on the traditional information of beneficial ownership, and it is important to understand the role they play in ensuring good standards also in the future (Mugarura, 2017).

New technologies could significantly enhance ownership disclosure by overcoming existing methods' shortcomings and better enabling administrators to do their job.

3.2. Identification of Challenges

Some of the main obstacles in effective beneficial ownership identification were identified such as the lack of data, sparse and fragmented records, and the inherent difficulty of identifying the true owner of assets and institutions situated across countries. The section outlines the problem that these issues pose for regulatory compliance impediment and necessitate more sophisticated tools to overcome these challenges.

3.2.1. Case Study Analysis

Using case studies where these traditional approaches have failed to identify beneficial owners, specifically looking at the Panama Papers and the Danske Bank scandal, to shed light on how opaque corporate structures have been used intensively by a wide array of actors including traders, money launderers and others to circumvent regulatory and compliance obligations (Gilmour, 2020). The Panama Papers, for instance, the leak of data from the law firm Mossack Fonseca, showcased a world in which the purpose of so many companies was precisely to obscure actual ownership (Gilmour, 2020). The Danske Bank case demonstrated inadequate screening of high-risk customers and lack of monitoring to ensure compliance with anti-money laundering rules (Gilmour, 2020).

But hidden ownership structures, such as those characterised by shell companies and trusts, are likely to be created intentionally to keep beneficial ownership secret (Gilmour, 2020). Such structures require more sophisticated detection techniques to uncover the true ownership relations.

In addition, financial flows continue to be largely opaque (as evidenced in the Panama Papers and Danske Bank scandals), benefiting from lack of transparency in beneficiary owner identification requirements and in the legislation that provides them (Gilmour, 2020). The same gaps in regulatory compliance standards (for example, due diligence and ongoing monitoring) that have enabled widespread illicit use of opaque companies also make it difficult to deploy these familiar tactics for identifying beneficial owners.

It's not surprising that these methods have failed, largely due to the fact that there are often complex chains of ownership, an absence of transparency, and lax regulatory compliance. These deficiencies demonstrate the importance of increased transparency, more appropriate regulatory alignment and enhanced tools for identifying beneficial ownership of financial institutions.

3.2.2. Data Collection Issues

Records on beneficial ownership are hard to come by because data is fragmented, those countries with public registers discrepancies between how they maintain records and beneficial ownership can

be concealed through the use of offshore jurisdictions (Gilmour, 2020). Fragmentation results because of the localised nature of finance-related data, making it challenging to create an integrated register of ownership identities that could complement or enhance identification protocols. Moreover, there are two other reasons why data is not available, the information is held by different jurisdictions and institutions and/or the data is held by institutions that are unable or unwilling to compel to provide it (Feridun, 2023).

Public registries, the best alternative given the difficulty of obtaining direct information about beneficial owners, are further hampered by inconsistent reporting and disclosure requirements at the country level, and by varying standards of transparency. Registries also too often have weak verification mechanisms, resulting in substantial gaps in beneficial ownership information (Gilmour, 2020). These inconsistencies generally pose major problems for the effectiveness of approaches built on public records.

To make things more difficult, many offshore entities are owned behind multiple layers of secrecy, with many jurisdictions prioritising the privacy of those who register companies. This makes traditional models of identification more challenging (Gilmour, 2020). These challenges explain why attempts to collect self-identification data from suspects have generally failed.

These data collection challenges are directly linked to difficulties in assembly of beneficial owners. The incomplete and inaccurate facts about beneficial owners that financial institutions can access hamper their ability to meet the requisite due diligence, monitor transactions, and meet related obligations. To the extent that information about beneficial owners is incomplete and inaccurate, the financial system is undermined (Gilmour, 2020).

To overcome these obstacles, efforts should be made to improve data-collection protocols, to improve transparency in reporting data, and to strengthen beneficial ownership regulations. To standardise data, increase auditing and verification processes, and create cross-jurisdictional information-sharing deals are necessary next steps towards overcoming the barriers to collection of high-quality, accurate and complete information about beneficial ownership.

The hurdles posed by beneficial ownership data-collection, i.e., the data fragmentation, the inhomogeneity among public registries, and the use of shell entities, need to be overcome to allow identification protocols to perform well as a gatekeeper of a bank account. Better data handling protocols and regulatory reform are key to responding to the data-related problems identified.

3.3. Exploration of Big Data Technologies

This section will review many of the big data technologies in use today, including Apache Hadoop and Apache Spark, which allow for analysis of large amounts of data in order to identify beneficial owners. It will look at how these tools can be deployed to aggregate data across different data sources and help to improve compliance.

3.3.1. Data Sources

It is now possible to identify beneficial ownership patterns in the financial sector using big data technologies in strategic ways. Drawing from a number of diverse data sources that include public registries, payment, social media and regulatory databases, and financial transaction records stored in banks and payment service providers' systems, it is possible to integrate and aggregate rich data layers that provide the information needed to build detailed maps of the possible beneficial ownership structures (Sharma et al., 2023). These integrated views improve the identification and analysis of ownership patterns committed to conceal the true beneficial ownership of companies and individuals.

- Data Aggregation and Compliance: This is the aggregation of data from a large number of disparate sources, which when coordinated and assembled, creates a powerful information asset. However, such aggregated data must be collated in observance of data protection and privacy laws, such as the UK's General Data Protection Regulation (GDPR) (Labadie & Legner, 2023a). Across all stages of beneficial ownership identification, the right to privacy of the data subject and the right to confidentiality of their data should be protected in strict compliance with these types of regulations. Compliance with these motives is not only a legal duty but is also at the core of data best practices.
- Public Registries and Financial Data: There is an opportunity to enrich general public registries on company ownership with other publicly available data, and especially to cross-reference ownership information with third-party data. This way, use cases involving checks and balances, such as map-assisted registry-based research performed by journalists and investigative NGOs, can be improved in terms of the accuracy and richness of the collected information on camouflaged ownership structures and larger beneficial ownership network segmentation (M & S, 2024). Financial due diligence institutions can refine ownership records collected from general public registries by cross-checking their information with additional data and reduce errors caused by faulty data entry or typographical mistakes. Furthermore, there are vast opportunities in exploiting financial transaction records with evolutionary analysis to generate more abuse alerts and leads on ownership traces in and between entities (Bianchi et al., 2022).
- Social Media and Non-Traditional Data Sources: Social-media could serve as an additional source of information about beneficial owners, including relationships among owners that wouldn't be detectable with standard data sources (Pontes et al., 2021). This novel data source can add another layer of useful information, particularly when the relevant entities go to great lengths to obscure their ownership.
- Regulatory Databases and Compliance Information: Another source are regulatory databases that contain data about compliance, enforcement actions and other regulatory filings regarding beneficial ownership that allow analysts to assess the regulatory environments in which companies conduct business, as well as the jurisdictions from which they originate ('the jurisdiction of incorporation'). Data about this variable can be mined through a programme, potentially allowing for the discovery of patterns or anomalies in the nature of ownership (e.g.,

complex structures) that might indicate the presence of high or reputational risk or other problematic setups (Holzinger, 2017).

- Advanced Analytical Techniques: Big data technologies, including machine-learning and natural language-processing (NLP) algorithms, can further improve beneficial-owner(s) identification. These algorithms can identify patterns and anomalies within large data sets and identify complex ownership structures even if parties take elaborate steps to hide such information. For example, NLP can process data from social media and other sources that are difficult to access and can provide useful data that helps in beneficial-owner(s) analysis. For example, M & S (2024) demonstrate that machine learning and natural language-processing tools can improve the institutional capacity to identify beneficial-owners of debt-ridden firms.

Combining these data sources and processing methods into a single system, in a way that ensures compliance with data-protection laws, increases the speed and effectiveness with which they can identify beneficial owners, strengthening the beneficial ownership regime and making a positive contribution to the financial sector. Big data technologies that can process and analyse huge, complex datasets, in a way that complies with data-protection legislation, are essential for the kind of transparency and accountability that is necessary within modern financial systems.

3.3.2. Processing Tools

Platforms using Apache Hadoop and Apache Spark are the state-of-the-art in big data technologies when it comes to large-scale management and processing of beneficial ownership data. Apache Hadoop and Apache Spark are powerful tools to manage and analyse large amounts of unstructured and structured data of all kinds at scale. They provide the backbone that enables large-scale data analysis in financial contexts (M & S, 2024).

- Apache Hadoop: Apache Hadoop is used for distributed storage and processing of big data sets. It works by using clusters of commodity hardware for reliable, scalable, and distributed processing across all nodes in a cluster. As the need for data growth occurs, the cluster is scaled seamlessly so the Hadoop system can effectively handle the high-volume data-processing tasks such as in identifying beneficial owners (M & S, 2024). Hadoop's distributed, fault-tolerant architecture, for example, includes data integrity checks during computation, data replication with failover across nodes, data shuffling for optimal computation, etc. All of these are designed to mitigate data loss during hardware failures. Its ability to keep big datasets in a high state of integrity even during system breakdowns is fundamental to the capabilities described by (Salloum et al., 2019).
- Apache Spark: Apache Spark is better suited to real-time processing due to its capacity to process data in-memory which in turn allows for faster decision-making. Unlike Hadoop processing data in batches, Spark works on data processing in real-time which can help detect suspicious ownerships within scheduled time (M & S, 2024). Streaming and batch data processing is important for real-time processing, which is useful for identifying patterns in transactions that suggest money laundering (Holzinger, 2017). Financial institutions can use such technologies to

identify suspicious patterns of activity more effectively and respond more quickly, which in turn can lead to more effective regulatory responses to financial crime.

The fault-tolerant nature of their architectures also helps to ensure the integrity of the data, in case of hardware failures and crashes, as well as other technical errors, the system still works consistently (M & S, 2024). If a machine goes down, the system does not skip or inadvertently delete entries of data that was in the process of being captured and analysed. This 24/7 reliability is crucial to the purposeful nature of beneficial ownership analysis in that it will significantly enhance the data sources being analysed and lead to a more accurate and complete result.

One of the key benefits of big data technologies such as Hadoop and Spark for beneficial ownership identification is that they can process data very quickly, leading to almost instant detection of possible anomalies and suspicious patterns and ownership structures. This enables various automatic activities under the AML regime, for example Spark's real-time processing can significantly enhance the efficiency of AML activities (M & S, 2024). This speed will obviously be advantageous in financial crime detection, as it could lead to nipping the activity in the bud before possible further damage.

Both Apache Hadoop and Apache Spark can support the beneficial-ownership identification process in the context of customer due diligence. These tools provide a useful infrastructure for the successful use of data analytics in KYC and CDD activities. The ability of Hadoop and Spark to store large volumes of data, to tolerate errors, and to execute complex programs at high speed play a crucial role in ensuring near-real-time analysis and in facilitating better and faster detection and prevention of financial crime in private financial institutions.

3.4. Machine Learning Integration

In this section, machine learning techniques are outlined, like supervised or unsupervised learning, can be applied to the beneficial-owner disclosure process, and in particular how methods that can identify characteristics of beneficial ownership and disclosure-evasion networks, such as anomaly detection, natural language processing and graph theory, can be used to better detect beneficial owners and their networks.

3.4.1. Supervised Learning Techniques

Supervised learning techniques are a powerful machine learning approach that allows computers to adapt their behaviour through interactions with human agents over time, in order to improve the accuracy of their predictions based on labelled data without human intervention (Rahmaty, 2023). Supervised learning techniques allow computers to analyse large amounts of data to help predict outcomes. In finance, it was long believed that supervised learning could be employed for better predictive performance in areas such as risk assessment, fraud detection and beneficial ownership segmentation (Warin & Stojkov, 2021).

The machine learning model is trained on a training set of feature vectors, pairs of input data and the corresponding output values (label). It learns the mapping that transforms inputs into outputs. If that learning has been successful, the trained model is able to predict the outputs for future, unseen

feature vectors comprising input data. This applies to any classification or regression task (Holzinger, 2017).

Supervised learning plays a prominent role in the financial sector. For instance, it can be used for credit scoring (the machine learns how to predict an event like the default from a borrower), for fraud detection (the machine learns how to detect a potentially fraudulent transaction based on its patterns in the data from the past), and for beneficial ownership detection of the most opaque structures (the machine learns patterns and can then be used to filter out questionable ownership chains) (Warin & Stojkov, 2021).

Several supervised learning algorithms are particularly relevant in the context of financial data analysis:

- Support Vector Machines (SVM): A robust classifier that works perfectly with high-dimensional data, and one that is used in finance to segment customers based on their propensity to buy (prediction of outcome) or based on the metrics such as attrition, profitability, EBITDA etc, by pinpointing small section of data that can predict an entire large sample (Zhao, 2014).
- Random Forest: An ensemble learning method, also referred to as forest, that combines a large number of weak decision trees together to make a stronger (higher predictive performance) classifier. Random Forest is extensively used for credit risk modelling, fraud detection and the analyses of large datasets with hundreds of features (Zhao, 2014).
- eXtreme Gradient Boosting (XGBoost): XGBoost is an efficient and scalable implementation of gradient boosting that operates effectively in a distributed computing environment central to streaming production code. It's also a preferred option for classification and regression tasks in large-scale datasets as it reduces single-model overfitting issues and offers extremely high accuracy, making it best suited for complex simulations in financial models (Zhao, 2014).
- Deep Neural Networks (DNN): DNNs are a type of model with large numbers of parameters and neurons that can process hierarchical and complex patterns in data. In finance, examples of tasks that utilise DNNs include algorithmic trading and disruptive prediction models that rely on non-linear relationships in data to predict market movements (Chahal & Gulia, 2019).
- Anomaly Detection: These include recent anomaly detection techniques that can supplement supervised learning to further develop capabilities by combining machine learning and artificial intelligence to identify normal patterns and abnormal patterns in network traffic and detect fraud and other financial crimes (M & S, 2024).

One of the major benefits of supervised learning methods relative to unsupervised ones is that they scale well with larger volumes of data, especially big data. For example, distributed methods of data analytics can be used to detect anomalies in network traffic. Supervised learning methods can analyse petabytes of data in real-time which makes them suitable for automatically screening large amounts of trade and market data for complex patterns and anomalies (Vela et al., 2017).

3.4.2. Unsupervised Learning Techniques

Unsupervised learning is also an important function for identifying beneficial owners, especially in the absence of prior information about the entities of interest. Machine learning procedures can autonomously discover patterns and structures, such as linkages, in data (Wickramasinghe et al., 2021).

Clustering Algorithms:

- K-means Clustering: A frequently used clustering capability, splits data into clusters that share similarities. This sort of approach can group entities with similar characteristics, which could be very enlightening for purpose of grouping together companies sharing identical or similar structures, seeking to identify mapping hidden connections among beneficial owners (Sathya & Abraham, 2013).
- DBSCAN: DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is also a versatile, noise-tolerant method that detects clusters of arbitrary shapes and sizes, even if they are not necessarily connected. DBSCAN can be particularly useful for identifying suspicious ownership chains when they deviate from what's expected. Such anomaly detection can help identifying networks with irregular and non-linear ownership structures. DBSCAN is useful for finding noise in a dataset and identifying clusters of objects in a dense area. It's often used in domains like financial crime where irregular patterns within a network could signal some type of crime (M & S, 2024).
- Hierarchical Clustering: Hierarchical Clustering arranges observations in a nested hierarchy of clusters within clusters, enabling a more nuanced examination of underlying ownership relationships. It is especially useful at exposing the inherently hierarchical nature of ownership structures that may be obscured in standard analyses. For example, it can be used to identify entities that may be part of a larger ownership structure and highlight relationships that may indicate the influence of a single beneficial owner (Wickramasinghe et al., 2021).
- Anomaly Detection Techniques: Another kind of unsupervised learning, anomaly detection might be applied alongside clustering algorithms in order to enhance the identification of firms with abnormal ultimate beneficial ownership structures (Jiao, 2023). Unsupervised learning algorithms can mathematically and nonlinearly solve problems of classification (Sathya & Abraham, 2013).
- Isolation Forest: It is very effective because Isolation Forests recursively partition the data until only a single instance remains, splitting the anomalies while keeping the normal entities together. They are effective for anomaly detection, for instance, when lists of beneficial ownership entities are queried. All three of the algorithms described are powerful machine-learning techniques for detecting entities (or groups of entities) that are discrepant from the norm. This can point to risky owners or more complex ownership structures that law enforcement may want to review again (Alharbi et al., 2021).

- Autoencoders: One type of neural network, called the autoencoders, are trained to reconstruct input data from a lower-dimensional representation. This representation process focuses the model on the parts of the input data that are typical of normal patterns, while also highlighting slight deviations from the norm. Autoencoders that have been trained on ownership data are able to identify atypical owners in ways that indicate artificial camouflaging of beneficial ownership (M & S, 2024). This method appears especially useful in detecting sophisticated camouflage schemes.

Unsupervised methods such as clustering and detecting outliers are invaluable in discovering hidden relationships and aberrations in beneficial ownership data that could otherwise go unnoticed, improving monitoring of complex ownership structures and ensuring the overall integrity of the financial system for tackling financial crime.

3.4.3. Natural Language Processing (NLP)

NLP is a very handy tool for unstructured text data analysis, which is what is needed for tracking beneficial ownership. Machine processing based, for instance, on Named Entity Recognition (NER) and sentiment analysis, enables organisations to identify data and map a sense of sentiment to it wherever it occurs, in news articles, in financial reports, in blogs and social media posts.

Named Entity Recognition (NER) is a basic NLP approach for automatically extracting and categorising entities such as persons, organisation and places from unstructured text (Sharma et al., 2023). NER would be helpful for companies whose internal teams are trying to derive automatically-constructed pictures from large text datasets to analyse understood ownership structures. For example, if NER approach can identify the main person or organisation of interest, this can help fine-tune the relationships identified between these entities in the picture.

Sentiment analysis is the automated assessment of the emotional tone, positive, negative or neutral, communicated in text towards specific entities. By applying sentiment analysis to information such as news, financial statements and social media, relying entities can have some insight into how public the beneficial owner is. This gives a first hint as to how risky their beneficial owner may be (Sharma et al., 2023). For instance, negative sentiment can be an indicator of reputational risks or potentially criminal entities worthy of further scrutiny.

Tools such as Stanford NER or Illinois NER automate the tagging of entities in unstructured data, making it easier to extract beneficial ownership information (Sulaiman et al., 2017). Techniques such as Stanford NER and Illinois NER can be helpful in automated coding of entities from unstructured data (Sulaiman et al., 2017). Sentiment analysis and NLP can also be useful to discern ownership structures (Sharma et al., 2023). These ways of deploying techniques can aid in identifying red flags in patterns of ownership, along with the vision for improving the quality of beneficial ownerships (Gilmour, 2020).

3.4.4. Graph Theory and Network Analysis

Graph theory and network analysis have offered powerful new tools to model and represent the kinds of complex ownership networks depicted by a figure such as that supplied above. For example, it has become possible to structure and visualise overwhelming web-like portrayals of ownership structures, revealing otherwise concealed linkages between owners and interests (Bianchi et al., 2022). It does so via NetworkX, a Python library that makes it possible to easy draw the graphs and assign each entity a node and relationships as a directed edge. Analysts can then look at those networks and calculate different statistics, centrality measures and so forth when trying to unravel patterns and relationships in the network (Bianchi et al., 2022). This allows us to visualise complex networks as this diagram shows and therefore allows us to discern what otherwise is almost impossible to grasp at first sight or even getting in a spreadsheet.

In the case of network analysis, beneficial ownership structures can be mapped to reveal hidden links and structural anomalies, such as pyramidal or serial structures. Important techniques include measuring different centrality indexes such as degree centrality, eigenvector centrality, betweenness centrality and closeness centrality, to identify the most influential nodes of the network which may be control nodes or beneficial owners (Bianchi et al., 2022).

- Degree centrality: the count of direct links that a node holds. In ownership networks, positions of sustained node with high degree centrality (large number of connections coming in and going out) tend to point out the positions of the central nodes in the direct network and the controlling owners (Bianchi et al., 2022).
- Eigenvector Centrality: Gauges the quality of connections of a node instead of the network, based on the number of links – nodes with high eigenvector centrality are linked to other nodes with high eigenvector centrality, and are, thus, considered to be more influential in the network (Bianchi et al., 2022).
- Betweenness Centrality: It measures how often a node lies along the shortest path between two other nodes. Nodes with high betweenness centrality are controlling the flow of information, almost becoming natural gatekeepers in the ownership structure. Hence, nodes with high betweenness would represent entities that are not linked to too many others, but are often used to connect different parts of the network together (Bianchi et al., 2022).
- Closeness Centrality: Provides information on how close a node is to all other nodes in the network via the shortest path. Nodes with high closeness centrality can rapidly reach the rest of the network and have better access to information compared with other nodes. An actor in an ownership network may have a high closeness centrality if it is capable of quickly reaching other nodes in the entire structure or rapidly relay information throughout the network (Bianchi et al., 2022).

These centrality measures indicate important nodes within ownership networks that likely constitute a beneficial owner, or proxy. Network analysis further reveals clusters or communities within the network, and together they create a coherent representation for the ownership structure (Bianchi et

al., 2022). This allows the identification of critical value nodes that are highly connected, while nodes that are less connected are less critical.

Graph theory and network analysis, supported by tools like NetworkX, are invaluable for uncovering and visualizing complex ownership structures. By applying centrality measures and network analysis techniques, regulatory bodies and organizations can better identify who truly controls networks and detect potential beneficial owners (Crama et al., 2021). These methods enhance the understanding of ownership networks, revealing patterns, outliers, and key players, which is essential for effective oversight and risk management.

CHAPTER 4: PRACTICAL SHOWCASE

This section offers a practical demonstration of how big data and machine learning can potentially be used to determine beneficial owners, by using case studies and tools for anomaly detection and network analysis to unravel complex ownership chains.

4.1. Network Analysis: A Practical Showcase

The network analysis of the Panama Papers dataset illustrates the power of network analysis where a financial network can be crisply visualised and analysed to unravel hidden and untold stories. This section details the formation and the functioning of a Network Analysis Dashboard which was built to highlight and to get insights on relationships among entities within the Panama Papers dataset. The Network Analysis Dashboard was built using Python, Dash, and NetworkX libraries. It allowed the visualisation of dense interconnection between entities in the Panama Papers network which could be companies, intermediaries and officers.

4.1.1. Data Collection and Preprocessing

To produce the Network Analysis dashboard shown below, all the sub-datasets of the Panama Papers available online at ICIJ (2017) are loaded and cleaned first, in order to create aggregates of nodes and edges representing respectively owners, addresses, companies, intermediaries, officers and other types of entities. The datasets were loaded in Pandas data frame, and then aggregated, joined and cleaned to resolve large inconsistencies and duplicates.

Figure 1: Data Preparation

```
# Load and combine the datasets
def load_data():
    nodes_address_df = pd.read_csv('/path_to_data/
panama_papers.nodes.address.csv', low_memory=False)
    nodes_entity_df = pd.read_csv('/path_to_data/
panama_papers.nodes.entity.csv', low_memory=False)
    nodes_intermediary_df = pd.read_csv('/path_to_data/
panama_papers.nodes.intermediary.csv', low_memory=False)
    nodes_officer_df = pd.read_csv('/path_to_data/
panama_papers.nodes.officer.csv', low_memory=False)
    nodes_other_df = pd.read_csv('/path_to_data/
panama_papers.nodes.other.csv', low_memory=False)
    edges_df_full = pd.read_csv('/path_to_data/
panama_papers.edges.csv', low_memory=False)

    nodes_combined_df = pd.concat([nodes_address_df,
nodes_entity_df, nodes_intermediary_df, nodes_officer_df,
nodes_other_df])
    nodes_combined_df.fillna("N/A", inplace=True)
    nodes_combined_df.replace("", "N/A", inplace=True)

    return nodes_combined_df, edges_df_full
```

Source: Own representation

4.1.2. Dashboard Layout and Entity Selection

The dashboard layout allows users to quickly explore the underlying network of entities. From the options menu, users can select an entity. After selecting an entity, detailed information is displayed about the selected entity, including the property address and jurisdiction, the type of entity, whether it is in good standing, and its primary officer.

When an entity from the dataset is selected, its details are shown on the right sidebar and connected entities. The connected entities are separated into tables including the information about the connected entities, the related jurisdictions, their type of company and further information.

Here is a code snippet demonstrating how the connections are fetched and displayed:

Figure 2: Dashboard layout and entity selection

```
@app.callback(
    Output('connections-section', 'children'),
    Input('entity-dropdown', 'value')
)
def update_connections(selected_entity):
    entity_node_id = nodes_data[nodes_data['n.name'] ==
selected_entity]['n.node_id'].values[0]
    connections = edges_data[(edges_data['node_1'] ==
entity_node_id) | (edges_data['node_2'] == entity_node_id)]

    connection_rows = []
    for _, row in connections.iterrows():
        connected_node_id = row['node_1'] if row['node_2'] ==
entity_node_id else row['node_2']
        connected_entity_data = nodes_data[nodes_data['n.node_id']
== connected_node_id]

        connection_rows.append({
            "Connected Entity":
connected_entity_data['n.name'].values[0],
            "Address": connected_entity_data['n.address'].values[0],
            "Jurisdiction":
connected_entity_data['n.jurisdiction_description'].values[0],
            "Company Type":
connected_entity_data['n.company_type'].values[0],
            "Status": connected_entity_data['n.status'].values[0],
            "Additional Info":
connected_entity_data['n.note'].values[0]
        })

    connection_df = pd.DataFrame(connection_rows).drop_duplicates()
```

Source: Own representation

4.1.3. Network Visualization

The main feature in the dashboard will include the network graph, visualising the connections between the chosen object and other ones in the dataset. This network will be generated through the use of NetworkX and Plotly, counting nodes (each entity) and edges (the relationship between them) in the plot and, at the same time, applying centrality measures of betweenness, closeness and eigenvector centrality to characterise the importance of each chosen node within the network.

Here is how the network might be visualized:

Figure 3: Network visualisation

```
@app.callback(
    Output('network-graph', 'figure'),
    Input('entity-dropdown', 'value')
)
def update_network_graph(selected_entity):
    G = nx.Graph()

    entity_node_id = nodes_data[nodes_data['n.name'] ==
selected_entity]['n.node_id'].values[0]
    connections = edges_data[(edges_data['node_1'] ==
entity_node_id) | (edges_data['node_2'] == entity_node_id)]

    for _, row in connections.iterrows():
        node1_id = row['node_1']
        node2_id = row['node_2']
        node1_name = nodes_data[nodes_data['n.node_id'] == node1_id]
['n.name'].values[0]
        node2_name = nodes_data[nodes_data['n.node_id'] == node2_id]
['n.name'].values[0]

        G.add_edge(node1_name, node2_name,
weight=row.get('transaction_amount', 1))

    pos = nx.spring_layout(G, k=0.3, iterations=50)

    edge_trace = []
    for edge in G.edges(data=True):
        x0, y0 = pos[edge[0]]
        x1, y1 = pos[edge[1]]
        edge_trace.append(go.Scatter(
            x=[x0, x1, None],
            y=[y0, y1, None],
            line=dict(width=edge[2]['weight'] * 2, color='#888'),
            hoverinfo='none',
            mode='lines'))

    node_trace = go.Scatter(
        x=[],
        y=[],
        text=[],
        mode='markers+text',
        hoverinfo='text',
        marker=dict(
            showscale=True,
            colorscale='YlGnBu',
            size=[20 * betweenness[node] for node in G.nodes()],
            color=[eigenvector[node] for node in G.nodes()],
            colorbar=dict(
                thickness=15,
                title='Eigenvector Centrality',
                xanchor='left',
                titleside='right'
            ),
        ),
        textfont=dict(size=14)
    )
```

Source: Own representation

The Network Analysis Dashboard serves as a practical tool for visualizing and analysing complex financial networks, as demonstrated using the Panama Papers dataset. By integrating and displaying detailed entity information, connections, and centrality measures in an intuitive manner, the dashboard provides valuable insights into potential illicit activities and the intricate relationships within financial networks. This practical showcase highlights the power of network analysis in combating financial crimes and can be applied to other datasets within the financial sector to uncover hidden connections and mitigate risks. Refer *appendix A.1 & B.1* for datasets and working script used.

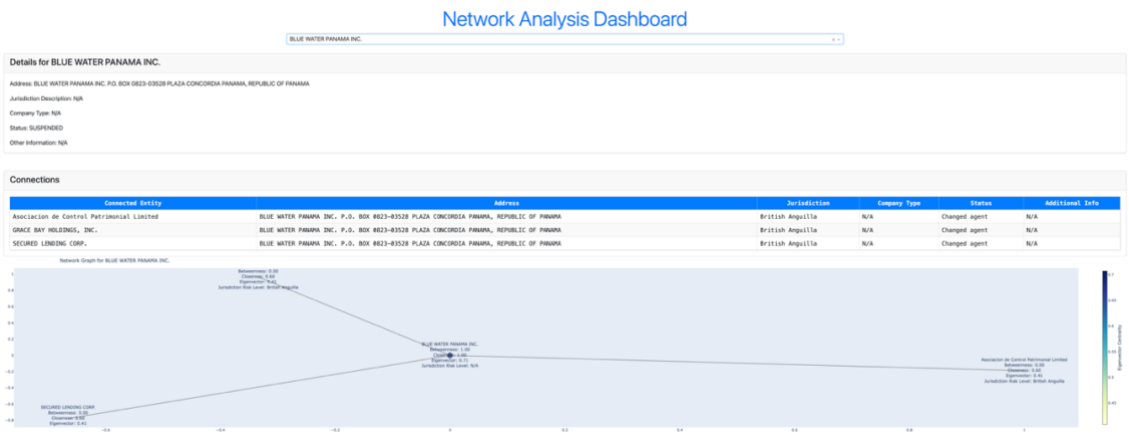
4.1.4. Case Study: Panama Papers Analysis

The practical exhibit uses the Panama Papers data to study how to identify key entities, and their relationships. The network analysis identifies central actors of the network, as in the figure above,

depicting companies or intermediaries that represent key roles in the offshore financial incorporated network.

Example Discovery: ‘BLUE WATER PANAMA INC.’ could be one of the key entitles identified.

Figure 4: Community Detection Algorithm



Source: Own representation from python

4.2. Anomaly Detection

This section talks about the application of anomaly detection techniques to detect fraudulent financial transactions, demonstrating various algorithms such as Isolation Forests and Autoencoders that can monitor and detect anomalies before these may result in a financial crime.

4.2.1. Concept and Tools

Anomaly detection is crucial in identifying irregularities within financial datasets and could prove to be a key factor in the prevention of money laundering such as what occurred at *Danske Bank*. Given the complexity of transactions, conventional monitoring of suspicious transaction scenarios is unlikely to be effective. Machine learning algorithms such as Isolation Forest, Autoencoders, and DBSCAN are utilized to effectively detect anomalies or outliers that significantly deviate from the norm in financial networks (Alsuwailem et al., 2022; Mhlanga, 2024). These techniques are essential for identifying suspicious activities and ownership structures that necessitate further investigation, thereby enhancing risk management and fraud detection capabilities in the financial sector (Aderemi et al., 2024; Jiao, 2023). Furthermore, the integration of big data analytics and advanced machine learning methods can strengthen compliance processes, facilitate cross-border collaboration, and improve the transparency of international financial flows (Crama et al., 2021; Jiao, 2023).

4.2.2. Data Preparation and Overview

This practical application integrates Isolation Forests, Autoencoders, and DBSCAN (Density-Based Spatial Clustering of Applications with Noise) methods, and visualizes the results using Dash, a Python framework for building analytical web applications:

- Transaction Amount: The total amount of each transaction.
- Account Balance: The balance of the account after the transaction.

– Transaction Frequency: The frequency of transactions made by the account.

The dataset is pre-processed for the anomaly detection and their structure can be described as follows:

Figure 5: Dataset pre-processing for Anomaly Detection

```
import pandas as pd
import numpy as np
import plotly.express as px
from sklearn.ensemble import IsolationForest
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Dense
from sklearn.cluster import DBSCAN

# Load the synthetic financial dataset
data = pd.read_csv('synthetic_financial_data.csv')

# Visualizing the dataset
fig = px.scatter_matrix(data, dimensions=["Transaction Amount",
"Account Balance", "Transaction Frequency"],
                        color="IsoForest_Anomaly",
                        title="Overview of Synthetic Financial
Dataset")
fig.show()
```

Source: Own representation

4.2.3. Implementing Anomaly Detection Algorithms:

Several algorithms are applied to the data to detect anomalies:

Isolation Forest, a method that randomly selects features and split values to find anomalies, can be used for anomaly detection. Anomalies in large datasets, such as those occurring in financial institutions, are isolated and detected since they are distinct from normal observations (Chang et al., 2007). These ideas can be combined with machine learning methods to detect fraudulent financial transactions, where unusual observations lead to financial criminal activities. Combating financial crimes, therefore, is just another battlefront to prevent money laundering (Ashfaq et al., 2022)

Figure 6: Isolation Forest

```
# Apply Isolation Forest
iso_forest = IsolationForest(contamination=0.05)
data['IsoForest_Anomaly'] =
iso_forest.fit_predict(data[['Transaction Amount', 'Account
Balance', 'Transaction Frequency']])
```

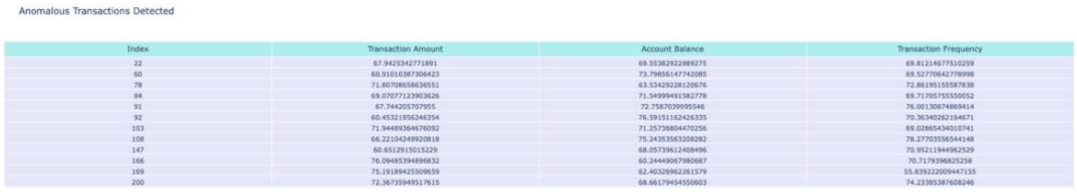
Source: Own representation

Visualization: Below is the dashboard visualization from the Isolation Forest method.

Figure 7: Suspicious Transaction Data through Isolation

Source: Own representation

Figure 8: Anomalous Transactions Detected through Isolation

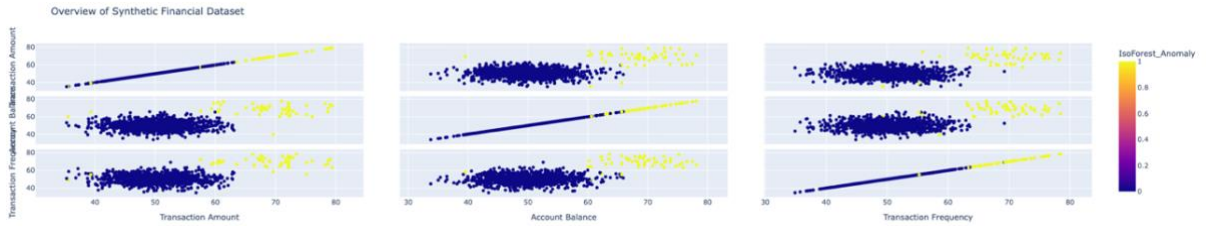


Anomalous Transactions Detected

Index	Transaction Amount	Account Balance	Transaction Frequency
22	67.9425342771891	69.55382922889275	69.81214677510259
60	60.91010387306423	73.79856147742085	69.52770642778998
78	71.80708658636551	63.53429228120676	72.86195155587838
84	69.07077123903626	71.5499949158277	69.71705755550052
91	67.744205707955	72.7587039995546	76.00130874869414
92	60.45321956246354	76.59151162426335	70.3634026216467
103	71.94489364676092	71.25736804470256	69.02865434010741
108	66.22104249920818	75.34353563208282	78.2770355644148
147	60.651291501522	68.05739612408496	70.95211944962529
166	76.09485394896832	60.24449067980687	70.717939682525
169	75.19189425509659	62.40326962261579	55.83922200944715
200	72.36735949517615	68.66179454550603	74.2338538760824

Source: Own representation

Figure 9: Overview of Financial Dataset (synthetic) through Isolation



Source: Own representation

Autoencoder, a type of neural network, is utilized to learn a condensed representation of data through encoding and then reconstruct the data from this encoding through decoding. Anomalies are identified based on the reconstruction error, where anomalies typically exhibit higher reconstruction errors (M & S, 2024).

Figure 10: Autoencoder

```
# Create an Autoencoder
autoencoder = Sequential([
    Dense(64, input_dim=3, activation='relu'),
    Dense(32, activation='relu'),
    Dense(64, activation='relu'),
    Dense(3, activation='sigmoid')
])

autoencoder.compile(optimizer='adam', loss='mse')

# Train the Autoencoder
autoencoder.fit(data[['Transaction Amount', 'Account Balance',
'Transaction Frequency']],
                data[['Transaction Amount', 'Account Balance',
'Transaction Frequency']],
                epochs=50, batch_size=16, shuffle=True)

# Calculate reconstruction error
reconstructions = autoencoder.predict(data[['Transaction Amount',
'Account Balance', 'Transaction Frequency']])
loss = np.mean(np.abs(reconstructions - data[['Transaction Amount',
'Account Balance', 'Transaction Frequency']]), axis=1)
data['Autoencoder_Anomaly'] = (loss > np.percentile(loss,
95)).astype(int)
```

Source: Own representation

Visualization: The Autoencoder results are visualized in the following dashboard:

Figure 11: Suspicious Transaction Data through Autoencoder

Source: Own representation

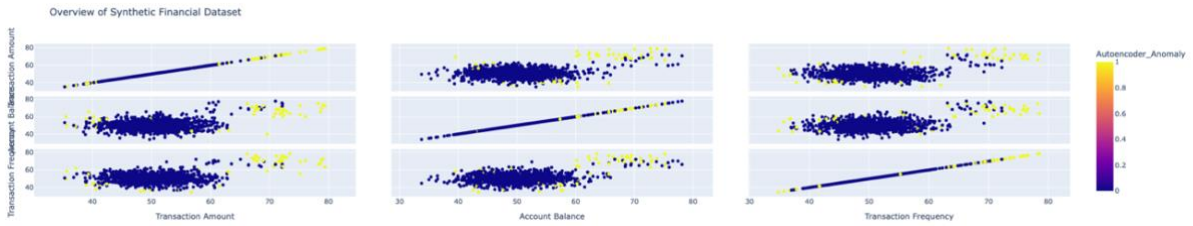
Figure 12: Anomalous Transactions Detected through Autoencoder

Anomalous Transactions Detected

Index	Transaction Amount	Account Balance	Transaction Frequency
14	48.14692833956947	53.85849355324172	35.75728689684962
22	67.9425342771891	69.55382922889275	69.81214677510259
78	71.80708658636551	63.53429228120676	72.86195155587838
80	37.18833170602345	49.04486118699916	62.063077108064576
91	67.744205707955	72.75870399955546	76.00130874869414
103	71.94489364676092	71.25736804470256	69.02865434010741
108	66.22104249920818	75.24353563208282	78.27703556544148
166	76.09485394896832	60.24449067980687	70.7179396825258
169	75.19189425509659	62.40326962261579	55.839222009447155
200	72.36735949517615	68.66179454550603	74.23385387608246
205	69.63118853218162	39.57943530250706	58.62348324880554
222	57.93008408072676	43.810922505865754	60.665166873281336

Source: Own representation

Figure 13: Overview of Financial Dataset (synthetic) through Autoencoder



Source: Own representation

The density-based algorithm, DBSCAN (Density-Based Spatial Clustering), is an algorithm that performs well at clustering high-density areas and also identifying outliers or anomalies in low-density regions (Fuhnwi et al., 2023). Its usefulness and applicability in many different fields suggest it for detection of anomalies, for detection of obstacles in cleaning out a warehouse, and also for prediction of financial frauds (Akinola & Ayano, 2017; Bushra & Yi, 2021)

Figure 14: DBSCAN

```
# Apply DBSCAN
dbscan = DBSCAN(eps=0.5, min_samples=5)
data['DBSCAN_Anomaly'] = dbscan.fit_predict(data[['Transaction
Amount', 'Account Balance', 'Transaction Frequency']])
```

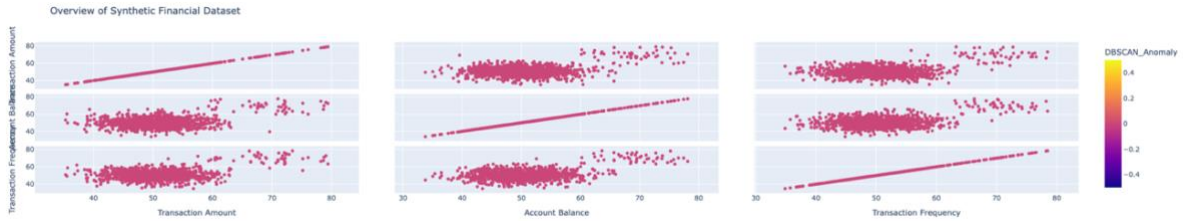
Source: Own representation

Visualization: The results of DBSCAN anomaly detection are shown in the dashboard below.

Figure 15: Suspicious Transaction Data through DBSCAN

Source: Own representation

Figure 16: Overview of Financial Dataset (synthetic) through DBSCAN



Source: Own representation

4.2.4. Interpretation of Results:

The application of these anomaly detection techniques has successfully identified a set of transactions that deviate significantly from the norm, suggesting potential irregularities or fraudulent activities. The table below lists these anomalous transactions, which are flagged by at least one of the detection methods.

Figure 17: Identifying Flagged Transactions

```
# List of Anomalous Transactions
anomalies = data[(data['IsoForest_Anomaly'] == 1) |
(data['Autoencoder_Anomaly'] == 1) | (data['DBSCAN_Anomaly'] == -1)]
anomalies_display = anomalies[['Transaction Amount', 'Account
Balance', 'Transaction Frequency']]
anomalies_display.head()
```

Source: Own representation

This practical showcase demonstrates the effectiveness of using Isolation Forests, Autoencoders, and DBSCAN for anomaly detection in financial datasets. Each method provides unique insights into potential anomalies, and when combined, they offer a comprehensive toolset for detecting fraudulent or irregular financial transactions. The results from these methods, visualized through Dash, provide an actionable framework that financial institutions can leverage to enhance their fraud detection systems and improve regulatory compliance. Refer *appendix A.2 & B.2* for dataset and script used.

4.3. Data Integration

This section argues that better identification of beneficial owners is possible only through the integration of data from different sources (e.g., financial transactions and social media), and points

to the role of big data technologies, such as Apache Spark, to process and visualise integrated datasets.

4.3.1. Concept and Tools

Modern financial systems require the monitoring of multiple risk indicators as they unfold in real time. This permits organisations to make choices that minimize financial crashes and enhance the integrity of institutions. The Risk Dashboard in this project can be designed to offer data visualisation, data analysis, and risk assessment to guard against deadly financial system collapses. The Risk Dashboard in this project uses Dash, a Python-based framework for analytical web applications. It provides an interactive and flexible visualisation interface for data analysis. Its modular architecture makes Dash an ideal tool for building modern analytical and interactive web applications like this one. Meanwhile, Apache Spark, a widely used qualitative research programming language, operates on multiple nodes to efficiently analyse data on parallel clusters.

4.3.2. Data Collection and Integration

Hence, it entails the processing and mashup of different datasets, each contributing individual insights on the companies. The most important datasets leveraged and incorporated into the dashboard are registration number, incorporation date, registered address, nature of business and operating status.

It is the information on companies from whom the data has been collected. Data on company are essential for risk assessment of companies.

- Persons with Significant Control (PSC): PSC holdings are an important gauge of the governance structure of companies, and a sensible risk gauge associated with ultimate ownership.
- Financial Transactions: Information on transactions, in the context of these companies any financial activity conducted by the company, including date of transaction, amount involved, and input parties. This feed can be used to accept financial funds and as a way to keep tabs on money flow for any irregularities or suspicious patterns in the data
- Risk Scores: Risk scores data assigns risk scores to an individual transaction type, company, or an entire company's activities based on a broad set of financial, governance and extrinsic metrics. Risk scores provide a metric of risk that can be quantitative in nature and helps evaluate the relative risk in prioritising risk mitigation activities.
- Social Media Sentiment: Feed on sentiment towards companies via social media This data feed provides sentiment towards a company, based on the content of social media posts. By monitoring sentiments expressed in this way useful indicator of reputational risk can be gauged.

4.3.3. Data Integration Visualization

Below information describes the dataset integration concept and the scientific project it applies to:

- Companies House: UK register of companies registering office, location when registering, business type, operating status.

- Financial Transactions: Synthetic data generated to cover a believable range of financial transactions including amounts, dates, and risk scores.
- Social Media Data: It expresses the sentiments of the public and social media interactions from the companies in a synthetic format.
- Regulatory Data: UK Beneficial Ownership (People with Significant Control) register on companies.

Data Aggregation: This integrates Big Data frameworks for streaming, processing and analysis of real-time public domain social media data (based on Apache Kafka) and large volumes of historical data (from Companies House and the Beneficial Ownership register) tabled using Apache Hadoop, and using Apache Spark for real-time processing and analysis.

Figure 18: Aggregating Data from Apache Hadoop to Spark

```
from pyspark.sql import SparkSession

# Initialize Spark session
spark =
SparkSession.builder.appName("RiskDashboard").getOrCreate()

# Load datasets from HDFS using Spark
companies_df =
spark.read.csv("hdfs://localhost:9000/user/neeleshkhantwal
/Datasets/synthetic_companies_house.csv", header=True,
inferSchema=True).toPandas()
psc_df =
spark.read.csv("hdfs://localhost:9000/user/neeleshkhantwal
/Datasets/synthetic_psc_register.csv", header=True,
inferSchema=True).toPandas()
transactions_df =
spark.read.csv("hdfs://localhost:9000/user/neeleshkhantwal
/Datasets/synthetic_financial_transactions.csv", header=True,
inferSchema=True).toPandas()
risk_scores_df =
spark.read.csv("hdfs://localhost:9000/user/neeleshkhantwal
/Datasets/synthetic_risk_scores.csv", header=True,
inferSchema=True).toPandas()
social_media_df =
spark.read.csv("hdfs://localhost:9000/user/neeleshkhantwal
/Datasets/synthetic_social_media_data.csv", header=True,
inferSchema=True).toPandas()
```

Source: Own representation

As such, they are ingested, cleansed, processed, transformed, and standardised through the integration process prior to visualisation inside a dashboard.

4.3.4. Cleansing and Preprocessing

Data cleansing and preprocessing are very important steps in preparing the integrated data set for analysis. These include data cleaning by removing/correcting the incorrect/incomplete data from the integrated data set (Salloum et al., 2019). In addition, it includes the data preprocessing such as formatting and standardising the data which is very important step before streaming the data set into machine learning model and hole the necessary dashboard visuals about the model analysis (M & S, 2024).

Data Cleaning with Spark: The cleansed integrated dataset is obtained with an architecture that uses Apache Spark's distributed processing engine. Spark provides massive parallelism over clusters of computers, thus it's the ideal component for scaling to the enormous data volumes dealt with in the

project. The cleansing process involves identification and removal of duplicates, management of missing values, as well as removal of data element inconsistencies.

Figure 19: Data Cleaning with Spark

```
# Example of cleaning data
companies_df = companies_df.dropna()  # Drop rows with missing
values
transactions_df['transaction_amount'] =
transactions_df['transaction_amount'].fillna(0)  # Fill missing
transaction amounts with 0
```

Source: Own representation

Data Preprocessing: all the numerical features are normalised, map categorical features to a number, and eliminate any features that are deemed irrelevant for a wider and meaningful comparison. All this preprocessing makes the data ready for machine learning. One can use Spark's machine-learning library for this, called MLlib.

Figure 20: Data Pre-processing with Spark

```
from pyspark.ml.feature import StringIndexer, VectorAssembler

# Example of preprocessing
indexer = StringIndexer(inputCol="company_status",
outputCol="status_index")
companies_df = indexer.fit(companies_df).transform(companies_df)

assembler = VectorAssembler(inputCols=["status_index",
"registration_number"], outputCol="features")
companies_df = assembler.transform(companies_df)
```

Source: Own representation

Integration Techniques: Using these data, a triggered flow of data with real-time streaming between social media data and regulatory databases can be developed. Apache Kafka can be used to ensure real-time streaming integration and up-to-date record triggering. With enhancing the timeliness of the datasets and Apache Spark can be utilised for batch processing of large volumes of data from public registries and financial datasets.

Figure 21: Data Integration with Spark

```
# Example of real-time data integration using Kafka (pseudocode)
from kafka import KafkaProducer

producer = KafkaProducer(bootstrap_servers='localhost:9092')
for post in social_media_posts:
    producer.send('social_media_topic', post)
```

Source: Own representation

4.3.5. Analysis and Insights

After integrating and preprocessing the data, then last step is to perform analytics using Big Data technologies and machine learning algorithms. Apache Spark's MLlib library can be used to perform the analytics over the integrated corpus at scale.

Risk Scoring at Scale: The risk scored can be calculated using Spark's services enabling machine learning over distributed large datasets.

Figure 22: Risk Scoring

```
from pyspark.ml.classification import LogisticRegression

# Example of a simple logistic regression model for risk
scoring
lr = LogisticRegression(featuresCol='features',
labelCol='risk_label')
model = lr.fit(companies_df)
```

Source: Own representation

Pattern Recognition and Clustering: Clustering algorithms, such as K-Means, find groups of similar transactions or entities that ought to be put together so that by clustering is it easier to spot outliers or anomalies.

Figure 23: Pattern Recognition and Clustering

```
from pyspark.ml.clustering import KMeans

# Example of K-Means clustering
kmeans = KMeans().setK(2).setSeed(1)
model = kmeans.fit(companies_df)
```

Source: Own representation

Dashboard for real time risk monitoring: Risk scores of critical entities derived during the analysis are visualised using a real time dashboard, leveraging Apache Spark for continuous risk monitoring of entities of interest.

Interactive Network Graphs: Visualise integrated data in interactive network graphs and show relationships and ownership structure to explore massive amounts of data.

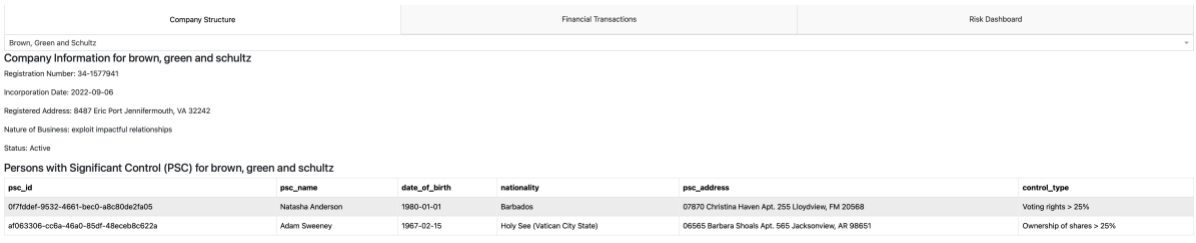
Figure 24: Risk Monitoring

```
import dash
from dash import dcc, html

app = dash.Dash(__name__)

app.layout = html.Div([
    dcc.Graph(
        id='risk-scores',
        figure={
            'data': [{'x': risk_scores_df['transaction_date'],
'y': risk_scores_df['risk_score'], 'type': 'line'}],
            'layout': {'title': 'Risk Scores Over Time'}
        }
    )
])

if __name__ == '__main__':
    app.run_server(debug=True)
```

Source: Own representation

Visualization: The results can be generated as below:

Figure 25: Company Information on Risk Dashboard



Source: Own representation

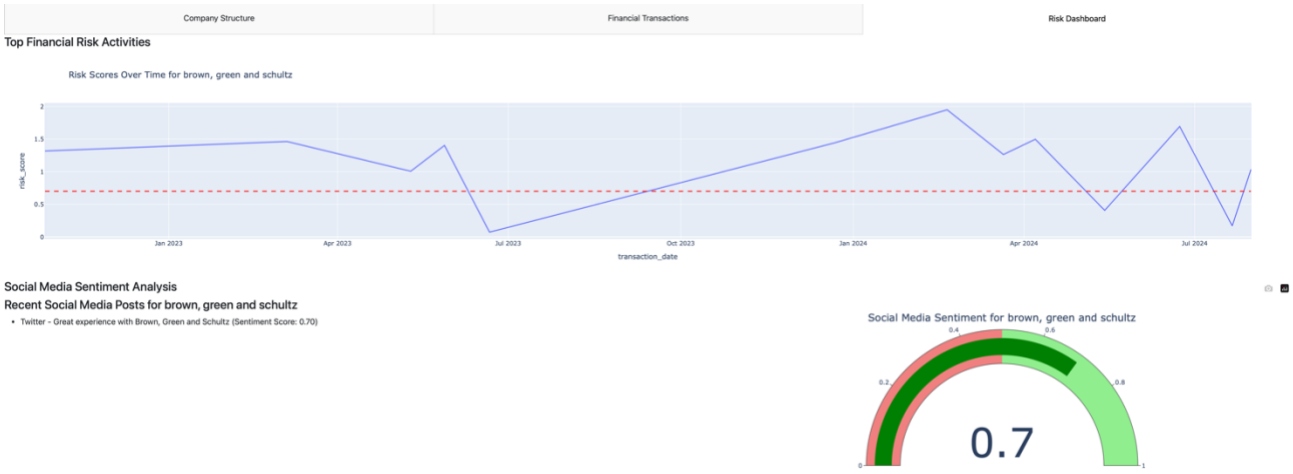
Figure 26: Transaction Information on Risk Dashboard

Company Structure | Financial Transactions | Risk Dashboard

transaction_id	company_id	transaction_date	transaction_amount	sender_account	recipient_account	transaction_type	financial_risk
fee2378b-cb17-4657-abbd-234511d4d703	cff1f7e9-79b1-4b17-940e-d461c49d86ee	2022-12-22	602062.96	GB45XYXF80617946016960	GB53MRKI94743361621497	Debit	0.26732839704616607
d21dddb0-0ad5-4eac-9f13-891db76769e9	cff1f7e9-79b1-4b17-940e-d461c49d86ee	2022-10-31	411099.43	GB27XPET14295339026848	GB08BTKD65639462687827	Credit	0.47952566096106075
dbe4ec76-78d5-4190-88a9-c30da76d6c9e	cff1f7e9-79b1-4b17-940e-d461c49d86ee	2022-11-05	135328.45	GB89CMQE30239647663966	GB70ZOLM53196073302175	Credit	0.42448837666894457
e11677f0-1cbe-4a8e-a7b9-5ce87db375e9	cff1f7e9-79b1-4b17-940e-d461c49d86ee	2023-01-20	64039.87	GB49IGLH03254827323935	GB11AVRC70513186764830	Credit	0.2407890924973726

Source: Own representation

Figure 27: Financial and Social Media Sentiment Analysis



Source: Own representation

The data integration process described in this section illustrates how Big Data technologies such as Apache Hadoop, Spark and Kafka can be utilised to aggregate and process multiple data sources. Refer *appendix A.3 and B.3* for dataset and script used.

CHAPTER 5: DISCUSSION

The discussion centres on the results from the demonstration project, and the implications of the use of big data and machine learning for financial institutions. This includes the various aspects of applying big data and machine learning to risk management, regulatory compliance and decision making.

5.1. Overview of Findings

The main objective of this research was to investigate the application of big data technologies in AML more specifically network analysis, anomaly detection, and data integration, all of which is key to combating increasingly sophisticated financial crimes such as money laundering, fraud, and regulatory breaches. By utilising large data sets, advanced and sophisticated algorithms, and real-time data-processing capabilities, the ability to detect, analyse and mitigate risk more efficiently has become a reality.

Chapter 4 provided hands-on demonstration illustrates how these technologies work to combat financial crimes, using real and synthetic data as multiple data sources merged into a holistic risk-management approach. Outcomes showed that selected machine learning and big data technologies, if harnessed correctly, can indeed break the mould of how financial crime is detected. The approaches will allow financial institutions to make better decisions and mitigate their risks.

Network Analysis (Panama Papers)

Panama Papers dataset (ICIJ, 2017) was used in network analysis. In particular, five patterns that were indicative of probable money laundering operations emerged. Specifically, the relationships among the entities (people, companies and offshore accounts) were mapped to identify cases in which the entities were connected to, Network graph depicting cluster-like entities that were tightly connected to each other, mostly involving jurisdictions with weak sanctions, which was indicative of money laundering.

One significant finding was some suspicious clusters of irregular proportions in terms connections, reflecting more complicated ownership constructs with complex chains of intermediaries and, subsequently, more sophisticated attempts to conceal the underlying nature of financial transactions. Simply put, network analysis works, and it should therefore be a core component of any effective financial crime-detection strategy.

Anomaly Detection (Synthetic Data for Danske Bank Case)

Anomaly-detection demonstration utilised simulated data from the Danske Bank case. It contains hundreds of simulated accounts and in the thousands simulated transactions per day, with transaction features that intentionally emulate some of the patterns in real non-simulated datasets.

These patterns could be detected using an anomaly detection algorithm with machine learning. The unusual characteristics included large transactions, countries with high levels of risk in relation to financial transactions, and another pattern that indicated layering the act of concealing in a series of

financial transactions as a step in money laundering. The results demonstrated the value in the early detection of anomalies using anomaly detection algorithms, reducing the risk that small criminal transactions by non-trusted clients grow into major banks' money laundering scandals like the one that happened with Danske Bank.

Data Integration and Risk Dashboard

The bringing of these data from different sources together into a single risk dashboard was an important achievement in terms of realising the operational potential of real-time monitoring and decision-making in banks. This risk dashboard went one step further by aggregating data from company registries, financial transactions and social media sentiment analysis to real-time alerts to help manage risk surrounding entities or transactions.

Making complex data analytics understandable and eye-catching for financial analysts and decision-makers with the risk dashboard was one of the most significant advantages of the technology. Real-time processing not only allowed for display of most significant risk areas, but also for immediate processing of new data and updating of risk assessments as data arrived. This is particularly important in fast-moving financial systems where decisions based on outdated information can pose serious financial or reputational risks.

The success of this data integration also emphasised the importance of using big data technologies at scale. The complexity involved in collating and aggregating data from disparate sources would have been extremely difficult to manage without scalable big data technologies, such as Hadoop and Spark. The study showed that financial firms could implement similar dashboards to help them streamline their risk management processes, reduce operational costs, and aid in the fulfilment of regulatory compliance obligations.

5.2. Implications for Financial Institutions

Network Analysis Interpretation

The network analysis of the Panama Papers dataset found key information about concealed relationships and groups of those who were potentially responsible for criminal activities. Network analysis turned out to be the perfect tool, because it is good at finding complex pathways and groups that are often deliberately hidden (Esoimeme, 2016). The pathways often went through entities in countries known to be less regulated by international oversight, and which facilitated money laundering and other types of crime (Dumitrescu et al., 2022).

These patterns of clusters in the Panama Papers network are quite apt demonstrations of the value of network analysis as a tool in fighting financial crime. When institutions can see it, they can do something about it. For example, clusters with particularly high levels of transactions and densely connected entities across a network could be signs of a coordinated effort to move money in large volumes through this sort of firm, obscuring both their origin and destination, typical for money laundering.

The findings of these investigations emphasise the potential of network analysis as a pre-emptive risk assessment tool in financial matters.

Banks and other financial institutions can apply the same kind of methods to their internal datasets such as transaction logs, customer relationships, ownership structures to pick up on hidden relationships, such as a payment routed through a series of opaque firms, that could be early warning signs of something amiss (Saddiq & Abu Bakar, 2019). Network investigation could be incorporated into the risk management toolkit of institutions in advance of their becoming the epicentre of the next headline-generating scandal (Canhoto, 2021). As crime gets smarter, these methods could be essential to police the increasingly sophisticated networks of shell companies, trusts and intermediaries that criminals use to hide proceeds and launder money (Vemuri et al., 2023).

Additionally, network analysis could be used to model potential future risk by examining how structural change in one part of the network might affect the whole. For instance, if an entity in a financial network was deemed a suspicious activity, network analysis could be used to predict which other entities might also be affected in real-time, so that institutions could take preventive action (Esoimeme, 2016). This predictive power would help us create better steading walls to handle the risk we'll inevitably come up against.

Anomaly Detection Insights

The results of the anomaly detection on the synthetic dataset modelled around the Danske Bank case provide crucial guidance on how to detect precursors of financial crime in transactions of financial institutions. It showed that machine learning algorithms are particularly well-suited for detecting non-normal transactions, which often appears before commencing financial crimes like fraud or money laundering.

One of the main takeaways from this analysis is the significance of early diagnosis when it comes to preventing financial crime hitting massive proportions. In the case of Danske Bank, billions of euros have been transferred over several years by non-residents via the bank's Estonian branch, before the money laundering scheme was discovered (Bjerregaard & Kirchmaier, 2019). By having anomaly detection in advanced stage, the suspicious activity could have been picked up before a lot of damage would have been done (Vashistha & Tiwari, 2024).

To make the synthetic dataset as realistic as possible, the training included patterns of transactions that appear in money laundering, such as unusually large transfers by a client, frequent transfers by a client to a high-risk jurisdiction, and rapid movement of funds between a client's accounts. The anomaly detection algorithms correctly identified these patterns. The principles underlying these techniques are the same principles that can be applied in the real world. Financial institutions could deploy such techniques to improve the identification and investigation of suspicious transactions at an early stage, reducing their vulnerability to fines, losses and reputational damage.

From an implementation perspective, the financial institutions that employ anomaly detection algorithms could monitor transactions data in real time, in order to catch the suspicious activities as they are happening, rather than later. If transactions behaviour is delivered in real time for a particular customer and an algorithm identifies a spike of transactions that deviate from its typical behaviour the system could trigger a red alert. In today's ultra-fast financial environment, this kind of timely monitoring is crucial since there's a high cost in terms of conducting an investigation after the event.

Furthermore, it could be calibrated to the needs of the unique businesses within each institution (Vashistha & Tiwari, 2024). A retail bank might use it to detect fraudulent transactions by individual customers, while an investment bank may turn to it for detecting anomalous movements in large-scale markets. Their flexibility allows for various settings of context, making these algorithms a promising tool in combatting financial crime.

Risk Dashboard Utility

The risk dashboard developed for this study demonstrates how different data sources can be brought together and combined effectively to provide real-time risk management capabilities. The advantage of having such a dashboard is that it can combine information from multiple sources, such as company registries, financial records and social media sentiment into one cohesive view that allows financial institutions to manage risk on an ongoing basis, using the most up-to-date data available.

Real-time warnings of areas of risk to an institution eliminate the need for a constant filtering of risks. Automatically, the process of integrating and analysing large quantities of data and providing alerts about high-risk areas becomes more efficient because fewer analysts are employing their human judgments in that process. Such data-driven efficiency liberates employees to focus on higher-value work (Mhlanga, 2024).

Beside operational efficiency, risk dashboard can be used to maintaining regulatory compliance. The financial institutions usually face multiple regulations for preventing money laundering (AML) and other financial crimes. By utilising several laws and guidelines, governments are seeking to eliminate financial crimes. Know your customer (KYC), which often requires information such as gathering detailed information on all interactions with clients so that financial institutions can detect and prevent illegal activities, is one of the regulations for financial institutions in preventing crimes like money laundering (Chitimira & Munedzi, 2022). Dashboard can be used to monitor the activities of users and comply with relevant regulations when it finds possible violations such as a transaction exceeds threshold or it involves high-risk regions.

Additionally, the real-time functionality of the risk dashboard allows for timely responses to emerging threats. In a rapidly changing financial marketplace, the institution's capacity to make decisions based on 'live' information front and centre can be the difference between avoiding and sustaining major loss. The use of real-time data sources enables the institution to work with the latest

information available. With proactive intervention, institutions stand a better chance of preventing a potential risk from becoming a serious threat (Jiao, 2023).

5.3. Challenges and Limitations

Technical Challenges

- Data Quality Issues: clean-up of data supplied in inconsistent formats, missing values, and records redundancy was large in scope and time-intensive. Additionally, the data analysis was performed using synthetic data, which created a whole new level of challenges since it had to be matched with real-world characteristics of money and financial transactions.
- Algorithm Complexity: The machine learning and network analysis algorithms implemented in the project require technical competence, since it demanded adjusting the algorithm parameters and achieving the best performance, which is time-consuming, and therefore requires in-depth knowledge of both the algorithms and domain (Dumitrescu et al., 2022).
- Computational Power: The need to process large amounts of data was an obvious and demanding requirement for this project. Being able to address processing growth in this area was therefore a crucial criterion for success. The ability to do so with low latency was particularly important with real-time applications, such as risk dashboards (Jiao, 2023).

Data Privacy and Security Concerns

- Handling Sensitive Data: The fact that synthetic data was used to lower the risk of data breaches doesn't mean privacy concerns can be ignored. A balance was made in the attempt to keep the project compliant with the law, and thus protected by privacy regulations like GDPR, with creating synthetic data that mirrored real-world situations accurately.
- Accuracy of Synthetic Data: The most obvious limitation to using synthetic data was the danger that certain aspects of 'real' transactions could be under-represented by synthetic datasets. If certain behaviours were less likely to be modelled by synthetic data, then the conclusions drawn would not be generalisable. Furthermore, representing actual behaviours with sufficient realism, as well as incorporating any identified anomalies, would require advanced techniques for creating synthetic data that was representative, privacy-respecting and yet realistic.
- Data Security: Secure data protection was implemented from the time of collection to destruction. If done right, access controls and encryption provide a high level of security for data over its complete lifecycle (Labadie & Legner, 2023b). However, if the data is crossing borders, then this poses additional and complicated compliance requirements. To comply with EU regulations during the collection, analysis and preservation of data requires special care to prevent a breach.

Adoption Barriers

- High Initial Costs: Infrastructure, software and staff packages represents a significant initial investment. In particular, smaller financial institutions may find it hard to afford the cost of investment in big data technologies. Maintaining this infrastructure over time and scaling it, for

example doubling its capacity to hold more data as data volumes grow, becomes an additional fiscal burden (Jiao, 2023).

- Resistance to Change: Owing to old legacy systems and procedures, many organisations simply do not see the necessity to change. In other words, high levels of organisational resistance can make it hard to embrace the Big Data technologies. Successful change management strategies had to be used, including the creation of a culture of innovation (Gilmour, 2020).
- Skill Shortage: One of the major impediments to implementation would be a lack of skilled staff to build and maintain the technology. Financial services is one sector that has been faced with a chronic lack of professionals with the competencies required to take full advantage of advanced machine-learning and network-analysis tools. One response might be to invest in updating the workforce with the necessary skills, or to enter into a relationship with an external expert (Dumitrescu et al., 2022; Maulidiyah, 2023).
- Phased Implementation: Since, big data solutions may have substantial risks and costs, organisations should implement them in phases, by starting with a pilot project. This way, the value of these technologies can be demonstrated early on, and investment costs can also be incurred gradually over time. Organisation can create a culture that encourages long-term learning and experimentation to ensure long-term competitiveness (Rostami & Mondani, 2015). Firms should not only hedge against uncertainty of future developments by investing in their capabilities for continued innovation, they must also ease growing cultural resistance to change.

5.4. Real-time Data Processing and Decision-making

This section underscores the significance of capturing decisions with real-time data processing in financial institutions, where the authors explain how with real-time analysis it will be much easier to catch any errors and irregularities in transactional data, especially through effective monitoring of their customers' transactions and therefore staying in line with regulations and standards.

5.4.1. Importance of Real-time Data Processing

Financial institutions rely on real-time processing to modernise their operation and to comply with AML regulations. Machine learning and big data analytics allow banks to monitor transactions as they happen, enabling them to detect and react to suspect activities in real-time. This is especially important for anomaly detections where abnormal transaction patterns can signal financial crimes (Aderemi et al., 2024; Vashistha & Tiwari, 2024).

Real-time capable institutions can flag unusual transactions and investigate these incidents in time to stop a minor anomaly from evolving into a serious financial crime. Aside from helping the institution itself, real-time processing also helps client interests. It builds a sense of trust in the bank and promotes compliance to keep clients' money safe (Canhoto, 2021; Jiao, 2023). With financial crimes becoming more nuanced, it is evident that integrating advanced analytics in real-time processing is central to the integrity of finances (Canhoto, 2021).

5.4.2. Practical Applications

Real-time data processing brings substantial real-world benefits to Customer Due Diligence (CDD) and Know Your Customer (KYC) processes that are part of risk management and regulatory compliance in financial institutions, previously heavily reliant on time-consuming, document-heavy processes that deployed significant human resources to verify identity, assess risk and ensure compliance with anti-money laundering (AML) directives (Mhlanga, 2024).

For example, being able to access real-time information about the customer from multiple sources (e.g., credit reports, social media profiles, or transaction histories) allows institutions to greatly reduce the time and effort needed to verify customer identities and risk factors, and can dramatically accelerate the adoption of new customers, without manually assembling and pooling data from various sources.

Besides, real-time analysis is better suited to a company's ongoing monitoring of customers a fundamental part of KYC compliance processes. Since most financial crimes originally occur in the context of ordinary business activities, an institution can take a proactive step such as conducting enhanced due diligence or flagging certain activities for regulators before the risk becomes a financial crime (Jiao, 2023).

Real-time processing of data also helps to maintain accurate and up-to-date records in the databases of the financial institution (Mhlanga, 2024). For example, the bank knows when a customer has modified or liquidated a position in the bank records as soon as it happens, enabling the institution to maintain an accurate and up-to-date reflection of the customer's status and risk profile. If regulatory requirements mandate that this information be reflected in the records of the financial institution, real-time processing can facilitate regulatory compliance.

5.4.3. Future Implications

The potential new uses for real-time data processing are not limited to CDD and KYC, and many opportunities exist for enhancing the industry's financial crime prevention and risk-management efforts. Financial criminals are improving their capabilities and exploiting new technology to cover their tracks and evade old detection techniques.

Going forward, this may involve additional incorporation of artificial intelligence and machine learning algorithms into existing real-time anomaly detection technologies. This would allow these systems to gain greater sophistication and catch anomalies, either that would not normally be caught by human analysts, or to catch them earlier. This, in turn, would reduce the incidences of false positives and increase the number of real threats being identified (Canhoto, 2021; Jiao, 2023).

An increase in real-time processing capacity could allow for more forward-looking risk management mechanisms. Financial institutions could begin to avert certain threats before they occur, even at the micro-level, if they had better networks of data (Gilmour, 2020).

Real-time processing is not a technological revolution, but a cultural revolution, marking the paradigm shift in risk management, customer service, regulatory compliance and other areas that the finance industry must pay attention to in order to respond to new developments in the financial environment and face new challenges swiftly, effectively and proactively.

5.5. Implications for Practice

This part examines how financial institutions can effectively utilise big data and application of machine learning to enhance operational efficiency and risk management in an effort to reduce the costs associated with meeting compliance standards and deliver a superior customer experience, in real-time financial environments.

5.5.1. Enhanced Risk Management

The incorporation of big data technologies into the finance industry could transform the way corporations handle traditional risk management by turning it to a proactive system, where risk is constantly analysed and monitored. In finance, risk management has traditionally used static models that rely on historical data and are manually compiled to detect and mitigate risks, this approach is not much responsive to emerging threats (Mhlanga, 2024). However, by harnessing big data, institutions would be empowered to analyse extensive quantities of data from numerous information sources to spot risks.

Predictive analytics uses both historical and real-time data to identify trends and patterns that can signal emerging risks. For instance, machine learning models can be trained to identify slight variations in transaction patterns, such as many withdrawals from different accounts in rapid succession, which may be indicative of attempted fraud. In this case, early intervention by the institution may prevent the risk from growing into a much bigger problem (Jiao, 2023).

Further, risks can be reassessed as changes take place in real time, meaning that, for example, a bank can monitor its transactions for suspicious signs of money laundering and can freeze accounts or report activity to the authorities as and when it takes place, ultimately reducing the bank's risk and enhancing compliance with its requirements (Dumitrescu et al., 2022).

By bringing together information from a variety of sources, from customer transaction files to social media feeds to market feeds, risk management is bolstered by providing risk managers with a view of an entity's risk profile. Indeed, in today's financial system, where risks are more complicated than before, this holistic approach to risk is an increasingly important part of the equation (Johansson & Andersson, 2023).

5.5.2. Operational Efficiency

There are also expectations of operational efficiencies and cost savings through the use of big data technologies. For example, many automatable clerical tasks that tended to be time-consuming and

expensive when operated manually can now be performed much faster and more economically using big data tools and machine-learning algorithms (Jiao, 2023).

For instance, customer due diligence (CDD) and know-your-customer (KYC) identification procedures could leverage big-data technologies to automatically check customer information against risk-factor data from a variety of external sources at a high rate of speed, reducing the need for long, costly manual vetting, speeding up the onboarding process and freeing up resources for better-value activities (Gaviyau & Sibindi, 2023).

Within the institutions, big data can help to identify operational inefficiencies such as collecting and crunching large data sets in near actual time, enabling the institution to track how external flows into and out of its transactions (flowing stocks and bonds, for example) and its interface with customers (such as the time it takes for cash to clear) work in real time, and thus identify bottlenecks and changes in the marketplace that can either be rectified or taken into account. It also enhances the agility of the institution in enabling it to respond in a flexible way by anticipating and adapting to those changes in the way it conducts itself.

5.5.3. Improved Customer Experience

Besides enhancing risk management and operational efficiencies, the use of big data technologies would also help to provide better customer experience. For instance, banks will be able to better utilise data-driven insights to comprehend customers' needs, preferences and patterns of consumption. On such basis, banks can tailor more personalised products and services for individuals (Mhlanga, 2024; Sharma et al., 2023).

For instance, analysing transaction data or market prices can enable institutions to pinpoint patterns in customer spending and saving behaviour, allowing them to offer more targeted and appropriate financial products. Another benefit of real-time data processing is that it can facilitate a more rapid responses to customer questions, concerns and complaints, improving the user experience (Aderemi et al., 2024).

Combined with big data technology, better customer experience can be a significant competitive advantage for a financial data technology, institutions can boost risk management, improve operational efficiency, and strengthen customer relationships, leading to enlarged bank size and value.

Overall, big data technologies represent vast potential benefits for banks and other financial institutions to develop practices that are more robust, efficient and customer-centric.

CHAPTER 6: CONCLUSION

Network Analysis, Anomaly Detection, Data Integrated Risk databases and other big data technologies mark a new chapter for financial institutions, allowing them to better track and manage risk and non-compliance, and build a strategic perspective on their operations. In this thesis we've looked at a number of methods as well as specific cases from dealing with the leaked Panama Papers to the systematic investigation of Danske Bank, and have introduced a real-world, practical risk dashboard that highlights the potential of these new technologies for transforming the financial world.

The Panama Papers case, exposing an extraordinary global network of offshore entities involved in tax evasion and money laundering, is a good example of a big data case. Throughout the study, how big data technologies can help investigators to identify illegitimate relationships and networks by 'looking beyond' information by visualising and exploring the connections between actors over large and intricate datasets through network analysis is explored.

Network analysis helped show the underlying structure of offshore networks, identifying common links or ultimate beneficial owners of companies. This allowed investigators to more easily trace the trail of money. Without network analysis, it would have been much harder to find both critical and merely interesting information among the vast amount of documents that had been leaked. The Panama Papers are a perfect example of how crucial network analysis will be in the future for financial institutions and regulators to expose financial crimes and prevent money laundering.

Moreover, the Panama Papers case illustrates some of the challenges facing big data investigations on how to manage huge amounts of unstructured data that one can easily find. As noted by the tools used, investigators had to deal with massive amounts of data with processing, analysing, and making sense of it involved huge amounts of digital ink. Sophisticated data-processing techniques built on top of network analysis and other big data technologies was needed to face this challenge and turn the data into intelligence.

For instance, the much-publicised Danske Bank scandal, which involved the laundering of almost €200 billion ($227 billion) in payments through its Estonian branch in recent years, underscores the need for robust anomaly detection systems to be developed and integrated within banks and financial institutions. Anomaly detection involves the detection of patterns in data that deviate from expected behaviour, and can potentially identify suspect flows that point to fraud, money laundering or other financial crime.

The anomaly approach, as applied to the Danske Bank case, entails using transaction data to detect deviations from the norm that are easy indicate of money laundering. Essentially, any kind of statistical anomaly was searched for uncommonly large transactions, regular transfers between accounts with no conceivable commercial purpose, and transactions drastically different from what a customer usually does were all flagged as anomalous.

Applying machine learning algorithms to anomaly detection enables financial institutions' systems to detect suspicious activities with increasing precision in real time. These powerful tools are capable, for example, of being trained on an institution's historical database to detect patterns inherent in criminal financial activities, based on this training, the tool can automatically detect further anomalies in new activities, and alert the system that follows suspicious activities. In the Danske Bank case, the application of such technologies could have prevented either a blowup or major containment.

Anomaly detection, however, comes with its own set of problems. The efficiency of such a system is dependent on the quality of the data and the intricacy of the algorithms used to evaluate them. Some systems score well on one aspect yet poorly on others a false positive would tend to see a real but banned transaction as an anomaly, putting staff investigating it under a tremendous strain. A false negative, on the other hand, which would see something suspicious and allow it to move into the clear, carries a significant consequence for financial institutions' workload and, potentially, the safety of the interbank system. These systems must be constantly refined and updated to maximise their efficiency.

Another focus looked at a consistent disconnect at integrating heterogenous datasets into a risk dashboard. Large financial firms maintain repositories of data from numerous sources related to transactions and their participants, such as customer data, markets data, regulatory filings, etc. By enabling information sharing among relevant participants, firms can potentially glean meaningful insight into the institution's comings-and-goings and risk zone.

This study developed a user-friendly risk dashboard using big data technologies, such as Apache Hadoop, Spark and Kafka, for evaluating real-time risk exposures of a company. This novel data dashboard integrates multiple big-data sources, such as company information, financial transactions and social media sentiment data, to provide instant signals of potential risks simultaneously. This enables the sharing of timely information across various risk units within financial institutions and monitors multi-factor material risk indicators, such as credit risk, market volatility and reputational risk, in one unified platform.

The creation of the risk dashboard stressed the role of data cleaning and pre-processing in data integration, while it allowed us to help the development officers clean and modify their datasets using Spark's distributed processing capabilities. For instance, the overall shape and statistics of the fundraising data have helped them identify millions of dollars of lost donations.

The impact of real-time data-processing can also be seen in the risk dashboard, which uses streaming data from social media and other channels to help financial institutions react to evolving risk and market conditions. This capacity to react in real-time is especially important in a world of fast-moving finance where problems can quickly lead to large losses.

These findings from the Panama Papers, the Danske Bank saga, and the development of the risk dashboard illustrate the transformative potential of big data technologies in the financial sector,

where the regulatory and financial costs of regulatory failures can be significant. Beyond enabling financial institutions to detect and prevent financial crime, these technologies also promote operational efficiency, optimise risk management and decision-making.

However, the use of these technologies do not come easy, technical problems such as data quality, computational power and complexity of algorithms hinder the full exploitation of big data. The protection of personal data privacy, data security and regulatory limits pose a real challenge to avoid the legal and ethical pitfalls in the usage of big data.

In the future, ongoing innovations in the use of big data technologies in finance shall be expected. Data volumes will only increase over time, and money crimes will only become more sophisticated. Financial firms will have to invest in developing and implementing robust data analytics platforms that are capable of handling the complexities of ever more complicated modern finance.

Furthermore, the role of human expertise remains fundamental. Although big data technologies deliver strong analytical tools to help filter and interpret the data, the findings that will be achieved by those technologies will depend on the data gathered and the skills of the experts that take the pedals.

Finally, this study has shown how big data technologies such as network analysis, anomaly detection and data integration can help address many of the most burning issues in finance today. Using these technologies, financial institutions can better identify and prevent financial crimes, manages their risk and take effective decisions, and make more informed operations. What has been found in this study has significant implications for the future of finance, and can act as a guide for the field of big data in the near future when more and more big data technologies are developed.

REFERENCES

Abrol, S. (2023). *Corporate Governance Insight, Volume: 5, Number:1, June 2023, eISSN: 2582-0834*. *5*(1), 82–111. https://doi.org/10.58426/cgi.v5.i1.2023.82-111

Aderemi, S., Olutimehin, D. O., Nnaomah, U. I., Orieno, O. H., Edunjobi, T. E., & Babatunde, S. O. (2024). Big data analytics in the financial services industry: Trends, challenges, and future prospects: A review. *International Journal of Science and Technology Research Archive*, *6*(1), 147–166. https://doi.org/10.53771/ijstra.2024.6.1.0036

Akinola, S. O., & Ayano, O. (2017). A multi-algorithm data mining classification approach for bank fraudulent transactions. *African Journal of Mathematics and Computer Science Research, 10*, 5-5–13. OpenAIRE.

Alharbi, A., Faizan, M., Alosaimi, W., Alyami, H., Agrawal, A., Kumar, R., & Khan, R. A. (2021). Exploring the Topological Properties of the Tor Dark Web. *IEEE Access, Access, IEEE, 9*, 21746-21746–21758. IEEE Xplore Digital Library. https://doi.org/10.1109/ACCESS.2021.3055532

Alsuwailem, A. A. S., Salem, E., & Saudagar, A. K. J. (2022). *Performance of Different Machine Learning Algorithms in Detecting Financial Fraud*. *62*(4), 1631–1667. https://doi.org/10.1007/s10614-022-10314-x

Apene, O. Z., Blamah, N. V., & Aimufua, G. I. O. (2024). *Advancements in Crime Prevention and Detection: From Traditional Approaches to Artificial Intelligence Solutions*. *2*(2), 285–297. https://doi.org/10.59324/ejaset.2024.2(2).20

Ashfaq, T., Khalid, R., Yahaya, A. S., Aslam, S., Azar, A. T., Alsafari, S., & Hameed, I. A. (2022). A Machine Learning and Blockchain Based Efficient Fraud Detection Mechanism. *Sensors (14248220)*, *22*(19), 7162-7162–7181. Academic Search Ultimate. https://doi.org/10.3390/s22197162

Bank of England, & FCA. (2024, August 8). *Machine learning in UK financial services*. https://www.bankofengland.co.uk/report/2022/machine-learning-in-uk-financial-services

Baru, C., Bhandarkar, M., Nambiar, R., Poess, M., & Rabl, T. (2013). Benchmarking Big Data Systems and the BigData Top100 List. *Big Data ; Volume 1, Issue 1, Page 60-64 ; ISSN 2167-6461 2167-647X*. BASE. https://doi.org/10.1089/big.2013.1509

Basel Framework. (2020). *Sound management of risks related to money laundering and financing of terrorism: Revisions to supervisory cooperation*. https://www.bis.org/bcbs/publ/d505.htm

Bianchi, P. A., Causholli, M., Minutti-Meza, M., & Sulcaj, V. (2022). *Social Networks Analysis in Accounting and Finance**. *40*(1), 577–623. https://doi.org/10.1111/1911-3846.12826

Bjerregaard, E., & Kirchmaier, T. (2019). The Danske Bank Money Laundering Scandal: A Case Study. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3446636

Bushra, A. A., & Yi, G. (2021). Comparative Analysis Review of Pioneering DBSCAN and Successive Density-Based Clustering Algorithms. *IEEE Access, Access, IEEE*, *9*, 87918-87918–87935. IEEE Xplore Digital Library. https://doi.org/10.1109/ACCESS.2021.3089036

Canhoto, A. I. (2021). Leveraging machine learning in the global fight against money laundering and terrorism financing: An affordances perspective. *Journal of Business Research*, *131*, 441-441–452. ScienceDirect. https://doi.org/10.1016/j.jbusres.2020.10.012

Chahal, A., & Gulia, P. (2019). Machine Learning and Deep Learning. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, *8*(12). https://www.ijitee.org/portfolio-item/l35501081219/

Chang, R., Charlotte, U., Ghoniem, M., Kosara, R., Ribarsky, W., Jing Yang, Evan Suma, Daniel Kern, Agus Sudjianto, & The Pennsylvania State University CiteSeerX Archives. (2007). WireVis: Visualization of Categorical, Time-Varying Data from Financial Transactions. *Http://Kosara.Net/Papers/2007/Chang_VAST_2007.Pdf*. BASE. http://kosara.net/papers/2007/Chang_VAST_2007.pdf

Chen, M. (2023). Compare the Beneficial Ownership Identification System from the Perspective of Anti-Money Laundering Between China and the UK. *Highlights in Business, Economics and Management*, *21*, 197–203. https://doi.org/10.54097/hbem.v21i.14184

Chitimira, H., & Munedzi, S. (2022). *Overview International Best Practices on Customer Due Diligence and Related Anti-Money Laundering Measures*. *26*(7), 53–62. https://doi.org/10.1108/jmlc-07-2022-0102

Chitimira, H., & Munedzi, S. (2023). An evaluation of customer due diligence and related anti-money laundering measures in the United Kingdom. *Journal of Money Laundering Control*, *26*(7), 127–137. https://doi.org/10.1108/JMLC-01-2023-0004

Crama, Y., Hübner, G., Leruth, L., & Renneboog, L. (2021). *Identifying Ultimate Beneficial Owners: A Risk-Based Approach to Improving the Transparency of International Financial Flows*. https://doi.org/10.2139/ssrn.3937884

Directive (EU) 2018/843 of the European Parliament and of the Council of 30 May 2018 Amending Directive (EU) 2015/849 on the Prevention of the Use of the Financial System for the Purposes of Money Laundering or Terrorist Financing, and Amending Directives 2009/138/EC and 2013/36/EU (Text with EEA Relevance), CONSIL, EP, 156 OJ L (2018). http://data.europa.eu/eli/dir/2018/843/oj/eng

Dumitrescu, B., Baltoiu, A., & Budulan, S. (2022). Anomaly Detection in Graphs of Bank Transactions for Anti Money Laundering Applications. *IEEE Access*, *10*, 47699-47699–47714. Directory of Open Access Journals. https://doi.org/10.1109/ACCESS.2022.3170467

Esoimeme, E. E. (2016). Wealth Management, Tax Evasion and Money Laundering: The Panama Papers Case Study. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2790543

FATF. (2023, October 3). *Guidance on Beneficial Ownership of Legal Persons*. https://www.fatf-gafi.org/en/publications/Fatfrecommendations/Guidance-Beneficial-Ownership-Legal-Persons.html

FCA. (2018). *FCG 3—FCA Handbook.* https://www.handbook.fca.org.uk/handbook/FCG/3/?view=chapter

Feridun, M. (2023). *Cross-Jurisdictional Financial Crime Risks: What Can We Learn From the UK Regulatory Data? 31*(3), 608–617. https://doi.org/10.1108/jfc-03-2023-0044

Frimpong, K. (2015). Back to basics: Fighting fraud and austerity. *Journal of Financial Crime, 22*(2), 219–227. https://doi.org/10.1108/JFC-11-2013-0065

Fuhnwi, G. S., Agbaje, J. O., Oshinubi, K., & Peter, O. J. (2023). An Empirical Study on Anomaly Detection Using Density-based and Representative-based Clustering Algorithms. *Journal of Nigerian Society of Physical Sciences*, *5*(2), 1-1–13. Academic Search Ultimate. https://doi.org/10.46481/jnsps.2023.1364

Gaviyau, W., & Sibindi, A. B. (2023). Customer Due Diligence in the FinTech Era: A Bibliometric Analysis. *Risks*, *11*(1), 11. https://doi.org/10.3390/risks11010011

Gilmour, P. M. (2020). Lifting the veil on beneficial ownership: Challenges of implementing the UK's registers of beneficial owners. *Journal of Money Laundering Control*, *23*(4), 717-717–734. Emerald Insight. https://doi.org/10.1108/JMLC-02-2020-0014

Haenisch, H.-Joachim., Speaker Von. (2024). *Know your customer: Unravelling the challenges* (edshst.AJP8358). Henry Stewart Talks. https://hstalks.com/article/8358

He, X. (2021, January 1). *The Impacts of Basel III on the Global Banking Regulations and the Responses of Regulatory Systems* (edsbas.182B452E). Proceedings of the 6th International Conference on Financial Innovation and Economic Development (ICFIED 2021) ; Advances in Economics, Business and Management Research ; ISSN 2352-5428. BASE. https://doi.org/10.2991/aebmr.k.210319.019

Holzinger, A. (2017). Introduction to MAchine Learning & Knowledge Extraction (MAKE). *Machine Learning and Knowledge Extraction; Volume 1; Issue 1; Pages: 1-20*. BASE. https://doi.org/10.3390/make1010001

ICIJ. (2017, November 5). *ICIJ Offshore Leaks Database*. https://offshoreleaks.icij.org/pages/database

Jiao, M. (2023). Big Data Analytics for Anti-Money Laundering Compliance in the Banking Industry. *Highlights in Science, Engineering and Technology*, *49*, 302–309. https://doi.org/10.54097/hset.v49i.8522

Johansson, L., & Andersson, E. (2023). Exploring the Distinctions in Characteristics, Framework Designs, and Toolsets of Big Data Systems. *International Journal of Research Publication and Reviews*, *4*, 1520-1520–1524. OpenAIRE.

Koker, L. D. (2006). Money laundering control and suppression of financing of terrorism: Some thoughts on the impact of customer due diligence measures on financial exclusion. *Journal of Financial Crime*, *13*(1), 26–50. https://doi.org/10.1108/13590790610641206

Koster, H. (2020). *Towards Better Implementation of the European Union's Anti-Money Laundering and Countering the Financing of Terrorism Framework*. *23*(2), 379–386. https://doi.org/10.1108/jmlc-09-2019-0073

Kurniabudi, K., Purnama, B., Sharipuddin, S., Darmawijoyo, D., Stiawan, D., Samsuryadi, S., Heryanto, A., & Budiarto, R. (2019). *Network Anomaly Detection Research: A Survey*. *7*(1). https://doi.org/10.52549/ijeei.v7i1.773

Labadie, C., & Legner, C. (2023a). *Building Data Management Capabilities to Address Data Protection Regulations: Learnings From EU-GDPR*. *38*(1), 16–44. https://doi.org/10.1177/02683962221141456

Labadie, C., & Legner, C. (2023b). Building data management capabilities to address data protection regulations: Learnings from EU-GDPR. *Journal of Information Technology*, *38*(1), 16–44. https://doi.org/10.1177/02683962221141456

M, V. P. P., & S, S. (2024). Advancements in Anomaly Detection Techniques in Network Traffic: The Role of Artificial Intelligence and Machine Learning. *Journal of Scientific Research and Technology*, 38–48. https://doi.org/10.61808/jsrt114

Matsuoka, A. (2020). The establishment of the OECD Asia-Pacific academy for tax and financial crime investigation. *Journal of Financial Regulation and Compliance*, *28*(4), 541-541–554. Emerald Insight. https://doi.org/10.1108/JFRC-12-2019-0139

Maulidiyah, D. N. (2023). *Consensus on the Role of Culture in Restraining Financial Crime: A Systematic Literature Review*. *31*(4), 883–897. https://doi.org/10.1108/jfc-05-2023-0103

Mhlanga, D. (2024). The role of big data in financial technology toward financial inclusion. *Frontiers in Big Data*, *7*, 1184444. https://doi.org/10.3389/fdata.2024.1184444

Mugarura, N. (2017). Tax havens, offshore financial centres and the current sanctions regimes. *Journal of Financial Crime*, *24*(2), 200–222. https://doi.org/10.1108/JFC-01-2016-0008

Mugarura, N. (2018). *The Implications of Brexit for UK Anti-Money Laundering Regulations*. *21*(1), 5–21. https://doi.org/10.1108/jmlc-07-2016-0032

Najafimehr, M., Zarifzadeh, S., & Mostafavi, S. (2023). DDoS attacks and machine-learning-based detection methods: A survey and taxonomy. *Engineering Reports*, *5*(12), e12697. https://doi.org/10.1002/eng2.12697

Pontes, R., Lewis, N., McFarlane, P., & Craig, P. (2021). *Anti-Money Laundering in the United Kingdom: New Directions for a More Effective Regime*. *25*(2), 401–413. https://doi.org/10.1108/jmlc-04-2021-0041

Rahmaty, M. (2023). Machine learning with big data to solve real-world problems. *Journal of Data Analytics; Vol. 2 No. 1 (2023): Winter*. BASE. https://doi.org/10.59615/jda.2.1.9

Rostami, A., & Mondani, H. (2015). *The Complexity of Crime Network Data: A Case Study of Its Consequences for Crime Control and the Study of Networks*. *10*(3), e0119309. https://doi.org/10.1371/journal.pone.0119309

Saddiq, S. A., & Abu Bakar, A. S. (2019). *Impact of Economic and Financial Crimes on Economic Growth in Emerging and Developing Countries*. *26*(3), 910–920. https://doi.org/10.1108/jfc-10-2018-0112

Salloum, S., Huang, J. Z., & He, Y. (2019). Exploring and cleaning big data with random sample data blocks. *Journal of Big Data*, *6*(1), 1-1–28. Directory of Open Access Journals. https://doi.org/10.1186/s40537-019-0205-4

Sathya, R., & Abraham, A. (2013). *Comparison of Supervised and Unsupervised Learning Algorithms for Pattern Classification* (edsbas.A40F4828). BASE. https://doi.org/10.14569/IJARAI.2013.020206

Sharma, R. K., Bharathy, G., Karimi, F., Mishra, A. V., & Prasad, M. (2023). Thematic Analysis of Big Data in Financial Institutions Using NLP Techniques with a Cloud Computing Perspective: A Systematic Literature Review. *Information*, *14*(10), 577. https://doi.org/10.3390/info14100577

Sulaiman, S., Wahid, R. A., Sarkawi, S., & Omar, N. (2017). Using Stanford NER and Illinois NER to Detect Malay Named Entity Recognition. *International Journal of Computer Theory and Engineering*, *9*(2), 147–150. https://doi.org/10.7763/IJCTE.2017.V9.1128

Todorović, I., Komazec, S., Krivokapić, Đ., & Krivokapić, D. (2018). *Project Management in the Implementation of General Data Protection Regulation (GDPR)*. *8*(1), 55–64. https://doi.org/10.18485/epmj.2018.8.1.7

Valvi, E.-A. (2023). The role of legal professionals in the European and international legal and regulatory framework against money laundering. *Journal of Money Laundering Control*, *26*(7), 28-28–52. Emerald Insight. https://doi.org/10.1108/JMLC-12-2021-0139

Vashistha, A., & Tiwari, A. K. (2024). Building Resilience in Banking Against Fraud with Hyper Ensemble Machine Learning and Anomaly Detection Strategies. *SN Computer Science*, *5*(5), 556. https://doi.org/10.1007/s42979-024-02854-w

Vela, A. P., Ruiz, M., & Velasco, L. (2017). *Distributing Data Analytics for Efficient Multiple Traffic Anomalies Detection*. *107*, 1–12. https://doi.org/10.1016/j.comcom.2017.03.008

Vemuri, S., Jahnavi, P., Manasa, L., & Pallavi, D. R. (2023). Money Laundering: A Review. *REST Journal on Banking, Accounting and Business*, *2*(2), 19–24. https://doi.org/10.46632/jbab/2/2/2

Warin, T., & Stojkov, A. (2021). Machine Learning in Finance: A Metadata-Based Systematic Review of the Literature. *Journal of Risk and Financial Management, Vol 14, Iss 302, p 302 (2021*. BASE. https://doi.org/10.3390/jrfm14070302

Wickramasinghe, C. S., Amarasinghe, K., Marino, D. L., Rieger, C., & Manic, M. (2021). Explainable Unsupervised Machine Learning for Cyber-Physical Systems. *IEEE Access*,

Access, IEEE, 9, 131824-131824–131843. IEEE Xplore Digital Library. https://doi.org/10.1109/ACCESS.2021.3112397

Wu, Z., & Salomon, R. (2017). Deconstructing the liability of foreignness: Regulatory enforcement actions against foreign banks. *Journal of International Business Studies*, *48*(7), 837. JSTOR Journals.

Zavoli, I., & King, C. (2021). The Challenges of Implementing Anti-Money Laundering Regulation: An Empirical Analysis. *The Modern Law Review*, *84*(4), 740–771. https://doi.org/10.1111/1468-2230.12628

Zhang, Z., Jiang, J., Wu, W., Zhang, C., Yu, L., & Cui, B. (2019). *MLlib*: Fast Training of GLMs Using Spark MLlib* (edseee.8731565). 1778-1778–1789. IEEE Xplore Digital Library. https://doi.org/10.1109/ICDE.2019.00194

Zhao, X. (2014). *Based on Gravity Method of Logistics Distribution Center Location Strategy Research:* International Conference on Logistics Engineering, Management and Computer Science (LEMCS 2014), Shenyang City, China. https://doi.org/10.2991/lemcs-14.2014.134

APPENDICES

Appendix A	Data Sources and Preparation	File
A.1	Panama paper leak dataset	Panama_paperleak
A.2	Synthetic Data generation script for Anomaly Detection	Synthetic Data Generation
A.3	Synthetic Dataset used in Data Integration	Data_Integration_Dataset.zip

Appendix B	Code Snippets and Algorithms	File
B.1	Python code used for Network Analysis	Network Analysis
B.2	Python code used for Anomaly Detection	Anomaly Detection
B.3	Python code used for Data Integration	Data Integration

APPENDIX C: THESIS EXPOSÉ

IU UNIVERSITY OF APPLIED SCIENCES

Exposé (07/2024)

Name: Neelesh Khantwal

Supervisor: Dr. Tobias Broweleit

Matriculation: 32209508

Study Program: MBA (Big Data Management – 90 ECTs)

Working Title: Enhancing Traditional Methods of Identifying Beneficial Owners with Big Data and Machine Learning

1. INTRODUCTION

Identification of beneficial owners (BO) for the anti-money laundering (AML) and countering the financing of terrorism (CFT) is of utmost priority. It is not only important to protect the customers to be a victim of a financial fraud but also to abide the International and national regulations. As per Financial Action Task Force (FATF) estimates, approximately 2-5 per cent of global GDP, trillions of dollars, get laundered annually highlighting the urgency of these identification processes besides its strategic importance. However, traditional methodologies of identification failed to process the sheer volumes and complexity of modern data, especially as banks grow more entangled in global financial networks. Multinational business corporations have increased the complexity of ownership structures and utilise shell companies in other jurisdictions to obscure ownership and evade anti-laundering/terrorist financing guidelines.

This project aims to leverage big data and machine learning analytics to supplement human resource-intensive identification processes for the identification of beneficial owners in business banking, yielding significant performance benefits for the banks in terms of substantially reducing the time taken as well as improving the quality of identification processes.

Currently, identifying beneficial owners in business banking takes anywhere between a few days and weeks, while with the help of machine learning models we can reduce this significantly to a few hours. This will enable banks to achieve superior performance flexibility in KYC processes, compliance accuracy and operational efficiency (Chawla & Khattar, 2020). The proposed technology will not just facilitate banks to comply with AML regulations but present a scalable solution that adapts to changing nature of FIs and regulatory mandates.

2. OBJECTIVES

This thesis aims to:

- Evaluate the current challenges in identifying beneficial ownership in business accounts.

- Develop a machine learning model that utilizes big data to identify beneficial owners more accurately and efficiently.
- Compare the performance of the model with traditional identification methods.

3. LITERATURE

3.1. Review Regulatory Requirements

- Regulatory Challenges: Examination of global anti-money laundering (AML) requirements and the problems with current compliance practices (Golightly et al., 2022).
- Traditional vs. Modern Approaches: Analysis of existing methodologies for beneficial ownership identification and their limitations (Latifian, 2022).
- Technological Innovations: Discussion on the application of big data analytics and machine learning in financial services, focusing on their potential to transform traditional practices (Vaisman & Zimányi, 2022).

4. METHODOLOGY

4.1. Region and Regulation for Research:

The UK has robust regulations in place for anti-money laundering (AML) and countering the financing of terrorism (CFT), primarily governed by the Financial Conduct Authority (FCA). The UK also adheres to the European Union's directives on AML, despite Brexit, and has its own set of detailed rules under the Proceeds of Crime Act and the Money Laundering, Terrorist Financing and Transfer of Funds (Information on the Payer) Regulations 2017.

4.2. Key Aspects to Consider:

- Regulatory Bodies:
  - Financial Conduct Authority (FCA): The FCA plays a pivotal role in overseeing AML and CFT activities within the financial sector. Banks and financial institutions must comply with FCA guidelines to prevent financial crimes (Golightly et al., 2022).
  - Companies House: Companies House is crucial for the registration and disclosure of corporate information, including beneficial ownership details, which is essential under the UK's implementation of the Fifth Money Laundering Directive (5MLD) (Latifian, 2022).
- Regulatory Framework:
  - Fifth Money Laundering Directive (5MLD): Transposed into UK law, the 5MLD aims to increase transparency by mandating that companies obtain and hold accurate and current information on their beneficial ownership (Vaisman & Zimányi, 2022).
  - Sanctions and Anti-Money Laundering Act 2018: Sanctions and Anti-Money Laundering Act 2018: This act provides the UK with the autonomy to apply international sanctions and manage its AML measures post-Brexit, reflecting the country's ongoing commitment to international compliance standards (Chawla & Khattar, 2020).

4.3. Data Collection:

UK-Specific Data Sources:

Companies House: Use the data provided by Companies House for access to registered company details and their reported beneficial owners.

UK's Financial Conduct Authority (FCA) Reports and Databases: Leverage financial transaction reports and other relevant data made available by the FCA for compliance checks.

4.4. Data Processing and Model Development:

- Use of Apache Tools:
  - Apache Hadoop and Apache Spark will facilitate the storage and real-time processing of large datasets, essential for monitoring and detecting transactional anomalies within the scope of UK regulations (Navlani et al., 2021).
- Machine Learning with Scikit-learn:
  - Machine learning models from Scikit-learn will analyse data to identify potential non-compliance or discrepancies in reported beneficial ownership, applying techniques suitable for predictive accuracy and regulatory adherence (Pajankar & Joshi, 2022).

4.5. Regulatory Compliance and Operational Efficiency

Real-time Monitoring and Reporting: Develop systems using these technologies to enable real-time monitoring and automated reporting, ensuring that financial institutions can maintain continuous compliance with UK AML regulations.

4.6. Challenges and Solutions:

- Data Privacy and Security: It will be crucial to comply with the UK's Data Protection Act 2018 to ensure data privacy while processing large volumes of sensitive information (CompTIA, 2020).
- Systema Relevance: Continuously update the compliance framework to align with any legislative changes, ensuring that the system remains relevant and effective under new regulatory conditions.

5. EXPECTED RESULTS

The expected outcome is a robust model that:

- Achieves higher accuracy and efficiency in identifying beneficial owners than traditional methods.
- Provides actionable insights that can be used by financial institutions to enhance compliance with AML regulations.
- Demonstrates scalability and adaptability to different data environments and regulatory requirements.

6. PRACTICAL IMPLICATIONS

This research will provide:

- A proof of concept that big data and machine learning can significantly enhance regulatory compliance in the financial sector.
- A framework that can be adopted by banks worldwide to improve their risk management practices and compliance with AML directives.

7. TIMELINE AND RESOURCES

- Timeline:
  - Jul'24: Literature review and initial data collection.
  - Jul'24-Aug'24: Data preprocessing and model development.
  - Aug'24: Model evaluation and comparative analysis.
  - Aug'24-Sep'24: Documentation and presentation preparation.
- Resources Needed:
  - Access to Python and data analytics libraries (Johansson, 2018).
  - Subscription to data feeds and public registries for real-time data access.
  - Computational resources for data processing and model training.

8. CONCLUSION

This project aims to bridge the gap between regulatory requirements and current technological capabilities, offering banks a more effective tool to identify beneficial owners. The application of machine learning and big data is anticipated to not only enhance the efficiency of these processes but also to provide a scalable solution that can adapt to evolving regulatory landscapes.

REFERENCE

Alejandro Vaisman & Esteban Zimányi. (2022). Data Warehouse Systems: Design and Implementation. eBook Collection (EBSCOhost) (3336619). Retrieved from https://search.ebscohost.com/login.aspx?direct=true&db=nlebk&AN=3336619&site=eds-live

Avinash Navlani, Armando Fandango, & Ivan Idris. (2021). Python Data Analysis: Perform Data Collection, Data Processing, Wrangling, Visualization, and Model Building Using Python. eBook Collection (EBSCOhost) (2725992). Retrieved from https://search.ebscohost.com/login.aspx?direct=true&db=nlebk&AN=2725992&site=eds-live

Chawla, H., & Khattar, P. (2020). Data Lake Analytics Concepts. In Data Lake Analytics on Microsoft Azure: A Practitioner's Guide to Big Data Engineering (pp. 1–10). Springer Nature eBooks (edssjb.978.1.4842.6252.8.1). https://doi.org/10.1007/978-1-4842-6252-8_1

Latifian, A. (2022). How does cloud computing help businesses to manage big data issues. Kybernetes, 51(6), 1917-1917–1948. Emerald Insight (edsemr.10.1108.K.05.2021.0432). https://doi.org/10.1108/K-05-2021-0432

Lewis Golightly, Victor Chang, Qianwen Ariel Xu, Xianghua Gao, & Ben SC Liu. (2022). Adoption of cloud computing as innovation in the organization. International Journal of Engineering Business Management, 14. Directory of Open Access Journals (edsdoj.5d947de2360342e4a8e2d1b39021027b). https://doi.org/10.1177/18479790221093992

Pajankar, A., author, & Joshi, A., , author. (2022). Hands-on machine learning with Python: Implement neural network solutions with Scikit-learn and Pytorch /. VLeBooks (edsvle.AH39944401). Retrieved from https://www.vlebooks.com/vleweb/product/openreader?id=none&isbn=9781484279212

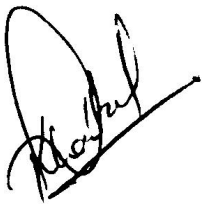
Quentin Docter & Cory Fuchs. (2020). Cloud networking and storage. In CompTIA Cloud Essentials+ Study Guide: Exam CLO-002 (pp. 35–76). Wiley. https://doi.org/10.1002/9781119642138.ch2

Robert Johansson. (2018). Numerical Python: Scientific Computing and Data Science Applications with Numpy, SciPy and Matplotlib. eBook Collection (EBSCOhost) (1990936). Retrieved from https://search.ebscohost.com/login.aspx?direct=true&db=nlebk&AN=1990936&site=eds-live

INTERNATIONAL UNIVERSITY OF APPLIED SCIENCES

Declaration of Authenticity

I hereby declare that I have completed this MBA thesis on my own and without any additional external assistance. I have made use of only those sources and aids specified and I have listed all the sources from which I have extracted text and content. This thesis or parts thereof have never been presented to another examination board. I agree to a plagiarism check of my thesis via a plagiarism detection service.



Berlin, 09/09/2024

Place, Date

Student signature