TALLINN UNIVERSITY OF TECHNOLOGY
School of Information Technologies

Bridget Chinye Kifordu
196387IVGM

Machine Learning in Tax administration: A case study of Nigeria

Master's thesis

Supervisor: Ben Sadok Yahia
Professor

Tallinn 2021

TALLINNA TEHNIKAÜLIKOOL
Infotehnoloogia teaduskond

Bridget Chinye Kifordu
196387IVGM

Masinaõpe maksuametis: Nigeeria juhtumiuuring

Magistritöö

Juhendaja: Ben Sadok Yahia
Professor

Tallinn 2021

Author’s declaration of originality

I hereby certify that I am the sole author of this thesis. All the used materials, references to the literature and the work of others have been referred to. This thesis has not been presented for examination anywhere else.

Author: Bridget Chinye Kifordu

04.05.2021

Abstract

Tax evasion is a global problem with huge impact on governmental revenues. Although digital taxation systems exist, there is room for further cost reduction in the area of selecting companies for further investigation. As tax administrations process huge amounts of corporate data, this data can be fed into data mining and Machine Learning methods to detect patterns among taxpayers to help detect companies which are more likely to commit tax fraud.

This research aims to investigate the usefulness of these ML methods in attaining tax objectives in companies with a special focus on Nigeria. The research problem highlighted the challenges being faced in the current system as a result of poor data management methods. The study proceeds to examine how the use of ML methods can help to solve the issues raised. To achieve the research objective, the case study methodology was adopted as appropriate. The means of data collection were interviews and systematic review.

The results obtained show that the stakeholders in the company tax administration process both want improvements in the tax system through greater use of data. The FIRS officials want this from the point of view of greater control of company activities and increased number of companies in the tax net which would translate to greater revenue, and the tax consultants, from the aspect of increased convenience afforded by the use of such data centric methods. However, the research discovered some limiting factors to the adoption of these models after reviewing the tax administrations of some selected countries. Conclusively, the research proposed a framework for the adoption of ML methods in the tax administration of companies in Nigeria.

This thesis is written in English Language and is 66 pages long, including 5 chapters, 4 figures and 2 tables.

Keywords: Machine Learning, Tax Evasion, Tax Administration, Data Mining, Nigeria

Annotatsioon

Masinaõppe mudelite rakendamine maksuametis: Nigeeria juhtumiuuring

Maksudest kõrvalehoidumine on ülemaailmne probleem, millel on suur mõju valitsuse tuludele. Kuigi digitaalsed maksustamissüsteemid on olemas, on ettevõtete edasise uurimise eesmärgil valimise valdkonnas ruumi kulude edasiseks vähendamiseks. Kuna maksuhaldurid töötlevad tohutul hulgal ettevõtte andmeid, saab need andmed sisestada andmekaevandamise ja masinaõppe meetoditesse, et avastada maksumaksjate seas mustreid, mis aitavad avastada ettevõtteid, kes tõenäoliselt panevad toime maksupettusi.

Selle uuringu eesmärk on uurida nende ML meetodite kasulikkust maksustamiseesmärkide saavutamisel äriühingutes, mis keskenduvad eelkõige Nigeeriale. Uurimisprobleem tõi esile probleemid, mis praeguses süsteemis seisavad silmitsi puudulike andmehaldusmeetodite tõttu. Uuringus uuritakse, kuidas ML meetodite kasutamine aitab lahendada tõstatatud küsimusi. Teadusuuringute eesmärgi saavutamiseks võeti vastavalt vajadusele vastu juhtumiuuringu metoodika. Andmete kogumise vahendid olid intervjuud ja süstemaatiline läbivaatamine.

Saadud tulemused näitavad, et ettevõtte maksuhaldusprotsessi sidusrühmad soovivad andmete suurema kasutamise kaudu parandada maksusüsteemi. FIRSi ametnikud soovivad seda ettevõtete tegevuse suurema kontrolli ja suurema arvu äriühingute maksuvõrgus, mis tooks kaasa suurema tulu, ja maksukonsultantide seisukohast, mis on seotud selliste andmekesksete meetodite kasutamise mugavuse suurendamisega. Siiski avastasid teadusuuringud mõned piiravad tegurid nende mudelite vastuvõtmisele pärast mõnede valitud riikide maksuametite ülevaatamist. Lõplikult pakkus uurimistöö välja raamistiku ML meetodite vastuvõtmiseks Nigeeria ettevõtete maksuhalduses.

See tees on kirjutatud inglise keeles ja on 66 lehekülge pikk, sealhulgas 5 peatükki, 4 numbrit ja 2 tabelit.

Märksõnad: masinaõpe, maksuheitlus, maksuamet, andmekaevandamine, Nigeeria

Acknowledgement

To the Almighty God, ultimate thanks for Your grace and strength needed to write this thesis successfully.

To my supervisor, Prof Ben Sadok, gratitude for your continuous support throughout the duration of the thesis.

To Silvia Lips, your assistance was indispensable.

To the entire eGov team for your support throughout the study period.

To all my parents and family, for your prayers and unending love.

To my course mates, I could not have asked for a better family.

To my friends, you guys are awesome.

List of abbreviations and terms

ANN	Artificial Neural Networks
CIT	Company Income Tax
CGT	Company Gains Tax
FAR	False Alarm Rate
FDR	Fraud Detection Rate
FIRS	Federal Inland Revenue Service
DBSCAN	Density-based spatial clustering of applications with noise
ML	Machine Learning
NGN	Nigerian Naira
NBS	National Bureau of Statistics
OECD	Organization for Economic Cooperation and Development
PIT	Petroleum Income Tax
PPT	Petroleum Profit Tax
SIRS	State Inland Revenue Service
SVM	Support Vector Mechanism
TCC	Tax Clearance Certificate
TIN	Tax Identification Number
VAT	Value Added Tax
WHT	Withholding Tax

Table of contents

List of figures

List of tables

1 Introduction

Taxation is a major source of revenue to governments in many countries around the world. Regardless of how reliant a country is on tax revenue to fund its public expenditure; every country seeks to maximise the benefits of taxation against the costs of collecting it. González & Velásquez (2013) also describes taxation as a major indicator of a government's effectiveness in the execution of its functions. A major setback to the attainment of optimal tax collection is the issue of tax evasion.

Tax fraud is a challenge that many countries grapple with (González & Velásquez, 2013 [in (Davia et al., 2000)]). Evidence establishes that tax evasion is extensive and even mainstream in almost all countries (Bird et al., 2003). A major reason for tax evasion is the fact that citizens and companies cannot justify their payments of tax to the governmental benefits received (González & Velásquez, 2013) and therefore, tax evasion is resorted to. Tax evasion basically occurs due to a gap in information between the tax officials and taxable entities which the latter takes advantage of (Martikainen, 2012).

To curb this issue, tax administrations conduct audits on taxable entities that are suspected of non-compliance. This is a painstaking process and with an increasing number of taxable entities, it is also not economically beneficial or even feasible to conduct audits manually. Consequently, different techniques from electronic tax systems, pre-filled forms, and more recently, data mining and analytic techniques have been and are being adopted with the aim of discovering these entities and principally maximising tax returns.

Bird et al., (2003) underscored that the challenge of tax evasion in developing countries is, "widespread and, indeed, systemic", and therefore has more serious implications than compared to developed countries. The Nigerian tax administration systems also faces these issues of tax evasion. Although an electronic tax system exists, a large proportion of tax returns are filled manually due to challenges encountered in the use of the e-tax filing system (Olaghere & Adewuyi, 2020). Manual tax filing system leaves opportunities for collusion between tax officials and business owners to under-declare their tax

liabilities. In addition, taxable entities have better opportunities to exploit all loopholes available to reduce their tax expenses.

Tax administrations have massive amounts of data that can be used to tackle tax evasion through data mining and Machine learning techniques (da Silva et al., 2016). This research will focus on the effective utilization of the data available to the tax administration in Nigeria to tackle the two forms of tax fraud highlighted above.

In this study, chapter 1 gives a general introduction to the research. Chapter 2 highlights the preceding research done in this subject area and includes highlighted studies on the case and proposed models. The third chapter gives the research questions and explains the case and method selection. Chapter 4 gives the findings from all data sources and the final chapter discusses these findings, summarises the research impact and gives a framework for implementation.

In this chapter, the author introduces the research problem in more detail in section 1.1 The research objectives are stated in section 1.2, and the research context is explained in section 1.3.

1.1 Research Problem

With the increase in incidences of tax fraud, tax administrations have also doubled down on measures to reduce such incidences. With scarce resources in monetary and labour terms, governments have adopted technological means and the use of data analytics to reduce the cost of tax audit and collection processes. Using the vast amounts of data available to tax administrations, Machine Learning (ML) techniques are increasingly employed to categorize taxpayers according to their propensity to default in tax obligations, and better drive efforts of tax officials in tax audit processes.

About 53% of OECD countries employ data mining techniques in their tax administration process (Centre for Public Impact, 2019). Successes worthy of mention has been achieved by some of these countries, particularly Australia and Norway, where different ML models have been employed in the process of eliminating tax fraud (Mendez, 2020). Investigation will be done as to the use of data analytics in tax systems where automated tax data capture exists and in systems where such does not.

In Nigeria, as manual systems are massively being utilized and with components of the tax filing process still involving physical involvement of the taxable entities, all loopholes for tax evasion and collusion still exist. Lack of proper data collection and management methods keeps the tax gap large, and this forms a major leakage in the revenue that is due to the government.

Therefore, there is a compelling need to maximise government taxation revenue while using resources available to the tax administrations- data. This is sought to be done with the aim of reducing the resources spent in tax audits and improving communication strategies with taxable entities. Thus, a more efficient system which uses data to predict non-compliant entities and communicate with these entities is desirable to curb the problems of tax leakages and increase government revenue.

1.2 Research Objective

The goal of this research is to highlight the current features of the Nigerian tax administration system, examine the loopholes for tax fraud that exist, and propose ways in which the application of ML models can be employed to reduce incidences of tax fraud and tax collusion. This research also aims to explore the role the tax consultant plays in tax evasion by examining the tripartite relationship between taxpayers, tax authority and tax consultant. Following this, the research objectives are highlighted below:

- Identify current features and loopholes of the Nigerian tax administration system.
- Propose insights on means of improvement of this system using ML techniques and data already available to the administrators.
- Examine elements which may inhibit or promote usability of these insights.
- Investigate the ethical implications of these propositions and its influence on stakeholders.
- Advance a framework to guide the execution on how this can be adopted taking the above factors into consideration.
- Explore the role of the tax consultant in facilitating or preventing tax evasion.

1.3 Background

Tax evasion is a criminal offence in which a taxable entity intentionally decreases its taxation incidence (David & Abreu, 2013). Although it is a challenge which can take many forms (Allingham & Sandmo, 1972), this research will focus on tax evasion initiatives advanced by corporate entities. Tax fraud detection is a prime concern of tax administrations globally of which they are to develop cost-efficient mechanisms of addressing it (de Roux et al., 2018). Tax administrations have previously resorted to experiences of auditors and rule-based systems in the fight against tax fraud. The problems associated with these include their inability to detect new fraud mechanisms and the high costs of building and maintaining rule-based systems (de Roux et al., 2018).

Taking an example from the Brazilian case, (da Silva et al., 2016), which considers the imbalance of tax officials who conduct audits to the increasing number of entities to audit, governments must strike a balance between with the costs of recouping leakages and the actual benefits derived from these auditing activities. This has resulted in the advancement of technological means of collecting tax and auditing submitted tax information. This transition commenced with the transition of filing tax returns from paper-based processes to electronic filing. This enabled the administrators to track the process better, identify erroneous entries faster and general improvements in tax revenues received.

Despite the benefits associated with electronic tax filing, the argument exists that it is a mere digitization of the manual processes and taxable entities understood the process well enough to carry on tax evasion practices even with e-filing methods. This is especially the case in the Nigerian state, where the culprits of tax fraud sometimes stem from both ends of the tax audit spectrum. The tax officials themselves may collude with taxpayers to under declare taxes. Tax evasion is heavily prevalent in the Nigerian state mainly due to poor data collection and management practices needed to tackle tax evasion (Momoh, 2018) and high levels of corruption pervasive in all levels of government.

The Nigerian tax system has its principal tax collection body as the Federal Inland Revenue Service (FIRS). The FIRS is responsible for the collection and remittance of all taxes due to the federal government. These taxes include: Value Added Tax (VAT), Company Income Tax (CIT), Petroleum Profit Tax (PPT), Capital Gains Tax (CGT),

Withholding Tax (WHT) and other Tax types (*FIRS*, 2021). Personal income taxes (PIT) are the prerogative of the State Internal Revenue Service (SIRS) of each state. This research will focus on the FIRS function in collecting CIT.

All corporate entities in Nigeria are identified for tax using the Tax Identification Number (TIN). The estimated total number of businesses in Nigeria, according to the National Bureau of Statistics (NBS) is 41.5m (Adesoji, 2019). Of this number, the total number of registered businesses in Nigeria shortly exceeds 3m (Iloani, 2019). The shortfall in the number of companies which remit taxes is apparent to begin with. The tax gap (the difference between what is due as tax remittance and what is remitted) is a further leakage to the total amount of revenue that is lost to tax evasion. The former FIRS chief, Tunde Fowler, stated that the Nigerian government loses as much as N15 billion annually to tax evasion. This was a statement made at the seminar titled, "Exchange of Information as a tool to combat Offshore Tax Evasion" ("Nigeria Loses $15 Billion Annually to Tax Evasion, Says FIRS," 2019).

There is, therefore, the need to employ ML techniques in the classification of taxpayer data and in taxpayer communication to ensure increased efficiency in tax administration and increased revenues. Although there are similarities in taxes collected and collection processes globally, tax administration is a country-specific process and insights from research must be finetuned to every country's specific case. The two major tax categories are direct and indirect tax. Investigation into CIT in Nigeria is necessary because CIT is hinged upon Tax Identification Number, with the option of electronic tax returns and therefore has some structure from which ML application can serve as an improvement.

A variety of ML models have been researched for use in tax administration. These models are broadly categorized into supervised, semi-supervised and unsupervised ML models. These models have been researched and adopted by different countries who either use one or a combination of approaches to meet their tax objectives. This research will explore these models with a focus on predictive and communicative models with a view to addressing challenges identified in the Nigerian tax administration.

Having this kind of system in place will reduce the tax gap, increase efficiencies in the system and increased learning of the ML models adopted will ensure continuous improvement of the results obtained. Overall, automated tax audit information capture is

sought, and this feature is advanced by the concept of ML specifically and Data Mining in general.

2 Literature Review

This chapter will highlight all relevant literature and related work done by other researchers that led to this thesis. It will also bring attention to related work that is significant to this research. The first section will provide a summary of previous work done in this area as well as give definitions to some of the terms as used in the current research context. It will also give a recommendation on models to be applied depending on tax objectives sought by any tax administration. The second section will give a theoretical framework on the research as work done by previous researchers are explored and insights on theoretical concepts and practices backing this research will be explained.

Moreover, the theoretical framework section will explore the tripartite relationship between the taxpayer, tax consultant and tax administrator. It will also highlight the successes derived from the application of ML in taxation as seen in two countries: Australia and Norway. This will be done with the aim to recommend a similar implementation in Nigeria and any other country with similar features in their tax administrations. Finally, a chapter summary is given.

2.1 Preceding Research

The fight to reduce tax gap by governments around the world has progressed through different methods. The reduction in tax gap is a goal that is aimed at to ensure that actual governmental revenue that accrues to the government is what is received. The data available to the tax administrators can be used in achieving this objective. Historical data used in ML models with taxpayer attitude pre classification can be used in predictive analysis to streamline the tax audit and administration process. ML can also be applied in detecting faulty/anomalous returns (Gedde & Sandvik, 2020) and even enhancing tax policy (Battiston et al., 2020). Consequently, the use of ML in tax can be used in correcting all the challenges and costs associated with tax audit patterns and selections which are not based on data.

2.1.1 Tax Evasion/Fraud

Tax evasion is the intentional alteration of tax-relevant details with a view to reducing the tax liability due to the government (khlif & Achek, 2015). One of the more established

theories of tax evasion, advanced almost 50 years ago by Allingham & Sandmo (1972), explains the taxpayer's decision of tax declaration as one of uncertainty. This uncertainty arises because the result of his (false) tax declaration can take either of two directions: It can either go undetected or be detected with the condition of a fine (Allingham & Sandmo, 1972). This position of uncertainty is what stems all forms of tax evasion. Tax evasion is primarily a violation of the law, an false declaration of income which makes the culprit liable to legal action from the authorities (Sandmo, 2005).

As this research is focused on corporate income taxes, the work of Chen & Chu (2005) on advancing a model on corporate tax evasion is also highlighted. Chen & Chu (2005) include other factors such as the alteration of the compensation scheme offered to employees of the company, distortion in manager's efforts and the efficiency loss of internal control. On the aspect of corporate income taxes, Feinstein (1991) also notes that the costs of tax evasion does not stop at decreased governmental revenue, but also extends to the inequity between tax evaders and honest filers. Bird et al. (2003) expounds upon this to include the following effects of tax evasion:

- It affects the taxes that compliant taxpayers face and the public services that citizens receive.
- It creates misallocations in resource use.
- It increases government expenditure in the quest to determine the magnitude of tax evasion, deter non-compliance and penalize tax-evaders.
- It alters income distribution in unpredictable ways.
- It incites feeling of unfair treatment and disrespect for the law. It weakens morale and leads to further tax evasion.
- It affects the accuracy of macroeconomic statistics.

Tax audit on the other hand is the inspection of the tax returns of a taxable entity to determine the non-alteration of income and expenses.

2.1.2 Electronic tax filing

E-taxation is the process of assessing, remitting and collecting tax through electronic means (Umenweke & Ifediora, 2016). Electronic in the definition of e-taxation encompasses the deployment of computer systems and networks in the payment of taxes (Richards & Ekhator, 2019). One major form of digitized taxation is the use of pre-filled

tax forms. This process, in which the tax administration, rather than the tax-payer is the originator of the tax returns document for the majority of individual taxpayer tax filing process (OECD - Forum on Tax Administration, 2006). The work of the taxpayer is to confirm that the information displayed is correct, making changes where necessary. This method simplifies the tax remittance process, and, in Estonia, the individual tax payer process typically takes 3 – 5 minutes (*Taxes in Estonia*, 2019).

The adoption of pre-populated returns is one giant leap in digitalization of taxation. Beginning in Denmark in 1988 and extending to Nordic regions (Estonia, Finland, Iceland, Norway, and Sweden) and other countries (such as Chile and Spain) (OECD - Forum on Tax Administration, 2006).

The digitization of the taxation process has a two-fold significant benefit to the government. On the one hand, it is considerably cheaper to administer, and on the other hand, it can help in decreasing incidences of tax avoidance and tax evasion (ICAEW, 2019). The latter benefit is felt most significantly when the data analytics are involved as additional tools are available to for tax data analysis and tax omission detection (ICAEW, 2019). The use of prefilled forms will require deep collaboration with third parties and might alter the tripartite relationship between taxpayers, the tax authorities and their tax advisors/consultants (ICAEW, 2019). A downside to it is that some taxpayers may not question the default information displayed and may end up paying more tax than what is due. Therefore, explanatory materials must be present to enable taxpayers "audit their returns" (ICAEW, 2019).

(Martikainen, 2012) notes that in developed economies, taxation amounts for taxable entities are derived (by calculation or validation) from tax returns and other data submitted periodically. The occurrence of certain taxable events like sale of assets, declaration of dividends, etc, can also affect the tax returns. A large proportion of what constitutes pre-filled forms is data that originates from third parties like banks, employers, unions. This data is also used for comparison and initial data check (Martikainen, 2012). This is the case for Finland where prefilled forms are used for income tax assessment and self-assessment is used for corporate tax.

2.1.3 E-Tax in Nigeria

The process to E-tax use in Nigeria began from a report from the visiting teams of the International Monetary Fund (IMF) Fiscal Affairs Division in 2004, 2005 and 2006 which gave recommendation for the implementation of an Integrated Tax Administration System (ITAS) (Richards & Ekhator, 2019). "The FIRS was given approval by the Federal Executive Council to procure, install and implement the ITAS which was aimed at re-engineering and automating the FIRS tax administration processes" (Richards & Ekhator, 2019). Although initial efforts at e-taxation began from the release of the NICIS software for self-assessment and implementation in 2010 (Umenweke & Ifediora, 2016), more traction in this regard was gained by the introduction of six new E-Tax filing services in 2017 by the FIRS. These include e-Registration, e-Stamp Duty, e-TaxPay, e-Receipt, e-Filing, and e-TCC (Tax Clearance Certificate) in 2017.

In Nigeria, tax is collected at the Federal, State and Local Government level. The Joint Tax Board instituted by the Federal Government brings all these tax bodies together for discussions (ICAEW, 2019). The FIRS, though, is the only body with automation in its taxation processes (Richards & Ekhator, 2019). Corruption is a major challenge in Nigeria, with a rank of 136th place out of 176 in the Corruption Perceptions Index conducted by Transparency International (ICAEW, 2019). The FIRS is not immune to this as "a research conducted by the UN office on crime and drug research shows that 27.3% of interactions with tax and customs officers include a request for a bribe" (ICAEW, 2019).

Despite the preference for the elimination of human interaction, the (OECD, 2004) report on tax compliance still emphasizes the need to have "local" knowledge in detecting what returns require further investigation. This will serve as a "support" to the digitization and data mining systems being sought as the systems do not contain all knowledge and the feedback provided by the tax officials provides an "invaluable aid in improving their quality" (OECD, 2004).

Digitalization initiatives, where online payments are made through a third-party payments' provider, efforts have begun for some types of tax (at least for VAT, CIT and some forms of WHT). There are, however, several challenges that inhibit its full implementation. To begin, taxpayers may be required to print evidence of online payment and proceed to the tax office for verification. This automatically eliminates all benefits

that may have been achieved from the online payment system (ICAEW, 2019; Olaghere & Adewuyi, 2020). Moreover, it provides the opportunity for further corrupt practices as tax officials may demand a bribe before processing their tax filing (ICAEW, 2019).

In addition, the issue of legislative cover for the digitalisation initiative is prevalent as the FIRS' power to mandate online tax filing or use electronic tax returns in court is unclear. Security challenges are also a major challenge, as a simple login password gives a taxpayer access to the filing system and this resulted in companies resisting the request by FIRS to install a plugin to extract data automatically from their accounting systems due to privacy and security concerns (ICAEW, 2019).

Moreover, low computer literacy levels, lack of widespread broadband internet connectivity, high cost of setting up a truly encompassing taxation system, epileptic power supply and, not surprisingly, cyber-criminal activity which aim at compromising the authenticity of the tax payment portals (Richards & Ekhator, 2019) are also challenges to successful implementation. Finally, there are significant downtimes faced by online taxpayers as they have difficulty accessing the online portals due to the inability of the servers to accommodate the volume of the online users (Olaghere & Adewuyi, 2020).

2.1.4 Machine Learning Application to Digitized tax system

Although e-taxation has profound benefits in terms of greater convenience to taxpayer and reduced costs to tax administration, it must not be equalled with the use of data analytics and Machine Learning. The use of ML techniques in achieving the over-arching goal of increasing tax revenue in tax administrations may be aided by the presence of an e-taxation system, and vice versa (Mendez, 2020). The presence of information systems is crucial in reducing the tax information gap from which tax evasion emanates (Martikainen, 2012). Also, the possibility that tax administrations can learn from past behaviour or taxable entities in determining future patterns is an advantage that data mining provides (Martikainen, 2012).

As behavioural patterns change, the models which are effective in deciphering fraudulent transactions from genuine ones in certain years may turn out to be less effective in subsequent years. Machine Learning is useful in this area when the models involved are integrated into the tax administration system. In this instance, the models are retrained by subsequent transactions as soon as they occur and this enables them to make more

accurate predictions in the future, and detect emergent fraudulent patterns (Nethone, 2019). The concepts of data mining and Machine Learning will be examined more closely in the following sections.

2.1.5 Machine Learning and Nearby Concepts Explained

Machine Learning, as defined by Arthur Samuel in 1959, is the discipline that studies the ability given to computers "to learn without being explicitly programmed (Ge et al., 2017). (Louridas & Ebert, 2016) explain further by stating that machine learning is a field where a computer learns to do a task on a new dataset after learning what decisions to make from a training dataset. It investigates the "study and construction of algorithms that can learn from and make predictions on data" (Ge et al., 2017). Although machine learning is not a new concept, what makes it useful in problems with greater complexity is the increase in computing power and its ability to the applied to a variety of domains is due to the surge in the volume of data being experienced globally (Louridas & Ebert, 2016).

Machine learning has been applied to spam filtering, to critical infrastructure repair, weather prediction and aviation turbulence, healthcare data access privacy breaches, advertising targeting and, interestingly, the linkage of "census datasets to track an individual's career trajectory" (Rudin & Wagstaff, 2013). Its use also spans recommendation engines (e.g. Netflix and Amazon), credit scoring and a host of other functions (Nethone, 2019).

It has been recognized as a powerful tool in the general area of fraud detection and numerous studies have been conducted in this regard. The idea behind its application in this area is that features exist which distinguish fraudulent transactions from genuine ones. A model can be trained with sufficient variables and data patterns to make these distinctions, based on a massive amount of interrelated information, "that sometimes may seem completely unrelated to a human being" (Nethone, 2019).

Worthy of note is that the use of ML models in this area will not replace the experts in the field, but will serve as a powerful tool in their amour to speed up fraud detection (Nethone, 2019). Its major advantage lies in the fact that the intensity of the "fraud like" patterns discovered can help automate what transactions are approved, and what are

handed over for further investigation (Martikainen, 2012; Nethone, 2019). The high accuracy levels make it a fantastic choice to detect fraudulent transactions.

Machine Learning as a discipline lies between statistics, computational science and other fields which deal with inference deduction and decision under uncertainty. (X.-D. Zhang, 2020) states that machine learning is a subset of Artificial Intelligence. The major approaches employed by machine learning includes Unsupervised, Semi-supervised and Supervised ML methods.

- **Supervised Learning Approach**

In this approach, the dataset involves input samples alongside corresponding output targets (Ge et al., 2017). The training set includes data and a task within that dataset. A labelled classification of the correct output is given and the supervised model is trained with this to identify the correct result in a new dataset that it has not encountered before (Louridas & Ebert, 2016). Ge et al. (2017) explain further that the data sample label may be discrete or continuous. When a discrete data label is used, supervised learning can be used for classification purposes. On the other hand, if the label of the data sample is a continuous value, regression models used for estimation and prediction can be created (Ge et al., 2017).

Consequently, two major categories of supervised ML algorithms are classification and regression algorithms. “Classification can employ logic regression, classification trees, support vector machines, random forests, Artificial Neural Networks (ANNs), or other algorithms while regression algorithms “predict the value of an entity’s attribute” (Louridas & Ebert, 2016). These categories with examples are detailed in Fig 1.

- **Unsupervised Learning Approach**

Here, the training set involved does not include a target output. The algorithm must discover the solutions independently. Typically, it involves identifying the underlying structure from a given set of patterns (Louridas & Ebert, 2016). Clustering algorithms and dimensionality reduction algorithms typically comprise unsupervised ML models as seen in Fig 1 below. In clustering algorithms, a dataset which covers various dimensions is taken as input. This dataset is then divided into clusters which have satisfied previously defined criteria (Louridas & Ebert, 2016). Dimensionality reduction algorithms on the

other hand project a dataset to fewer dimensions from an initial position covering various dimensions with the intention that these fewer dimensions can capture the essential aspects of the data. Ge et al. (2017) adds a further category of unsupervised learning called density estimation which involves "estimating the distribution of data within the input space" to which Self-Organizing maps are included.

(Gedde & Sandvik, 2020) in seeing the gap in the exploration of unsupervised methods in tax anomaly detection undertook a study in this regard. They propose what variables and models are suitable for "tax type data" including the situations best suited for the applications of these models.

- **Semi supervised ML approach**

This approach assumes for modelling a small proportion of labelled data and a large amount of unlabelled data. This approach is especially useful when it is too costly to use fully labelled training data. Both supervised and unsupervised models can be made semi-supervised (Ge et al., 2017).

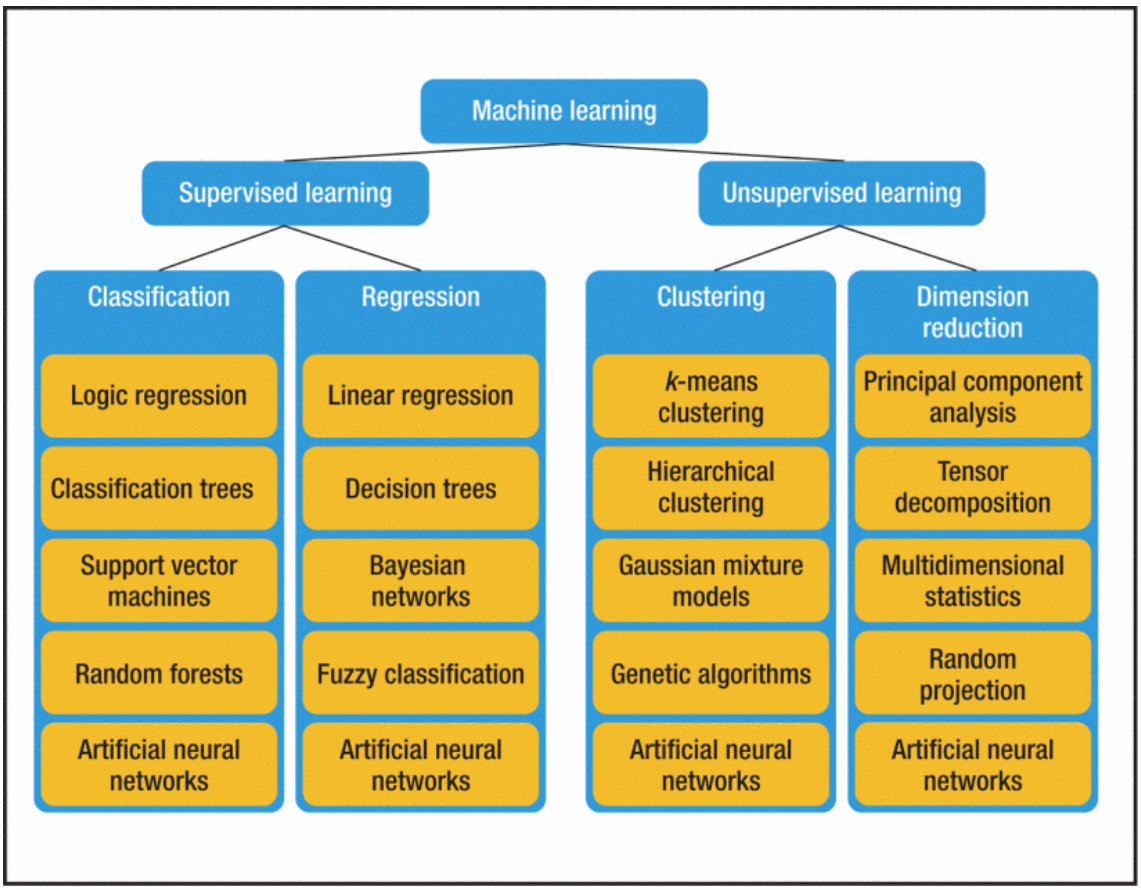

Figure 1. Machine Learning and its classifications. Source: Louridas & Ebert, 2016

Other concepts related to Machine Learning and whose definitions must be clarified at this point are given below:

- **Data Mining**

This is the extraction of previously unknown information from large volumes of data using mathematical or statistical techniques (Jun Lee & Siau, 2001). It involves analysing large datasets to detect "unsuspected relationships" with a novel summary which is both comprehensible and useful to the user (Hand et al., 2001). Machine Learning serves as a "computational engine" to data and analytics and serves to achieve "information extraction, data pattern recognition and predictions" (Ge et al., 2017).

(Hand et al., 2001) describe statistics as a necessary component of data mining, as "the aim of data mining is to provide inferences beyond the available database". (Hand et al., 2001) also detail the intricate relationship between these statistical methods and data mining: while statistical methods can be applied to any dataset, regardless of size, large datasets may require the application of sophisticated methods and statistical models to expose insights which may not be obvious if those techniques were not present.

In exploring materials for this research, materials with keyword "data mining" was also considered in doing the review as the models in these literatures were similar to the "Machine Learning" materials discovered. As some of the authors in the former category also described categories of data mining methods as supervised and unsupervised, the terms data mining and Machine Learning will be used interchangeably in some parts of this work.

- **Artificial Intelligence (AI)**

AI generally enables a computer or machine to replicate the capability of the human mind to perceive, learn, solve problems and make decisions (IBM, 2020). It is a field of computer science that sees the development of a system with embedded technologies (consisting hardware, software, and algorithms) to facilitate smart decision making (Soviany, 2018). AI utilizes a digital computer to perform tasks commonly associated with human beings such as "the ability to reason, finding out meaning for specific concepts, generalisation from examples or learning from past experiences to make smart decisions in new (unseen) situations" (Soviany, 2018). AI is generally classified into

Artificial Narrow Intelligence (ANI), Artificial General Intelligence (AGI) and Artificial Super Intelligence (ASI) (Soviany, 2018). As experiential learning is part of the associated tasks of AI, but falls in the category of Machine Learning, ML is said to be a subset of AI as seen in Fig 3 below:

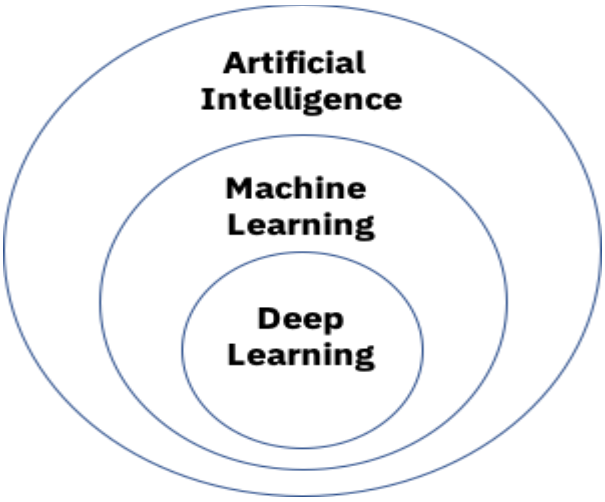

Figure 2. Relationship between AL, ML and Deep Learning. Source: (IBM, 2020).

- **Deep Learning**

Models created under this category are done by the creation of many layers of neural networks (Ge et al., 2017). While being created with performance improvement of highly sensitive data in mind, they are also regarded as 'black-box' algorithms in which both their features and decisions cannot be explained at the various system levels (Soviany, 2018). The 'deep' in deep learning is derived from the depth of layers involved in the neural network (Kavlakoglu, 2020). Typically, any neural network with above three layers is considered a deep learning algorithm (Kavlakoglu, 2020).

In deep learning, there are multiple 'hidden' layers of neural networks, each of which further refines the conclusions of the previous layer (IBM, 2020). The term 'forward propagation' refers to the movement of calculations through the hidden layers of the neural network to the output layer (IBM, 2020). Backpropagation also exists, where errors are identified, assigned weights and pushed back to previous layers to refine or train the model (IBM, 2020). Deep learning can work with both supervised and unsupervised ML methods. It has the ability to use both labelled data and large sets of unlabelled data in unsupervised methods to train itself and subsequently cluster and classify inputs (Kavlakoglu, 2020). More data points will be required in deep learning to improve its

accuracy and is especially useful for complex use cases like fraud detection (Kavlakoglu, 2020).

2.1.6 Fraud detection using data mining.

The rapid increase in the amount of data available and computing power has made it possible to obtain useful insights into previously unknown knowledge from that dataset. Models can be trained using informative data samples to reach great levels of sensitivity and accuracy according to previously defined Key Performance Indicators (KPI's) (Soviany, 2018). Data complexity issues which may be seen in big data samples can be handled with additional tasks such as, "data structuring discovery, hidden patterns discovery, clustering and anomaly detection from large data amounts, data mining, feature generation and selection" (Soviany, 2018).

A useful application of AI or ML techniques is in the area of financial fraud detection and management (Soviany, 2018). This is especially the case in online payments where models have the ability to detect hidden patterns at a much faster rate and human intervention is greatly reduced. Here, Fraud Detection Rates (FDR) are determined at the cost of a high False Alarm Rate (FAR) in typical legacy rule-based systems which lead to the denial of genuine online payments (Soviany, 2018). Its application is not limited to online payments, however. Many studies have been carried out based on its abilities to discover anomalies in accounting transactions, for instance in (Schreyer et al., 2017) and also detect fraudulent patterns in credit card transactions e.g. (Dal Pozzolo et al., 2014).

As (Gedde & Sandvik, 2020) investigate the use of Machine Learning in tax return anomaly detection, they highlight that, firstly, not all tax returns submitted are correctly done, and, therefore, not all can be separated for further audit. Accordingly, the selection of tax returns which require further investigations is the major focus of the tax administration. This forms the initial concerns of data imbalance.

As the models to be explored come from either a predictive or detective basis, it is necessary to note that the success of one method over another stems from the specific targets sought to be achieved, as some studies merely seek subclassification of datasets while others point to results in predictive capability (Dutta et al., 2017). As an initial probe at results descriptions, Schneider et al. (2015)'s hybrid approach to fraud detection highlighting varied levels of patterns is given in table 1 below:

Table 1. Hybrid approach to tax detection. Source: Schneider et al., 2015)

Suitable for associative link patterns	**Suitable for known patterns**	**Suitable for unknown patterns**	**Suitable for complex patterns**
- Social Network Analysis - Knowledge discovery through associative link analysis	- Rules - Type business and activity to filter fraudulent behaviours	- Anomaly Detection - Detect individual and aggregated abnormal patterns	- Predictive models

Schneider et al., (2015) further denote that a tax evasion model should be created taking into consideration "individual's and his environment's social analysis, business rules, anomaly characteristics and typical fraud monitoring models" in use. They also highlight other factors that pertain to probabilities for tax evasion such as tax morale, the high susceptibility of small businesses to tax fraud due to fewer controls and the short shelf life of tax models.

2.1.7 Systematic review of Machine learning and Data Mining models

In this section, there will be reviews of the topics of Unsupervised, Semi-supervised and Supervised ML methods for implementation in tax administration. Numerous studies have been conducted on the uses of ML in fraud detection (Dal Pozzolo et al., 2014; Soviany, 2018) amongst others. These studies have been conducted using different models and show the usefulness of one model over the other and what circumstances under which this can be used. They majorly focus on the application of one (or combination of some) models in solving the financial fraud and tax evasion challenges in specific cases. This is not unusual as tax administration is a country-specific phenomenon.

In this study, seven models will be explored from their use in tax cases in specific and financial fraud in general from relevant literature. Notable applications of these methods through different techniques as executed in some countries will be highlighted, with discussion on what features are necessary for adoption, and other factors to be considered for implementation. A model (or combination) useful for the Nigerian case will be derived.

A summary of the reviewed literature on the subject matter is given in table (2) below:

Table 2. Summary of reviewed literature highlighting models used and findings. Source: Author.

	Author and paper	Models	Data	Purpose	Model function	Results/Findings
1	Bayesian Networks on Income Tax Audit Selection - A Case Study of Brazilian Tax Administration (da Silva et al., 2016)	-Naïve bayes -Tree augmented Naïve bayes	Real data- Brazilian personal income tax	Increase specificity in cases marked for audit	Predictive function	Naïve bayes gives good taxpayer segmentation with improvement on the current system
2	Detecting financial restatements using data mining techniques (Dutta et al., 2017)	-Decision trees -ANN -Naïve Bayes -SVM -Bayesian Belief Network (BBN).	3,513 Financial restatement incidences (2001 – 2014)	Highlight all cases of financial restatements (both erroneous and intentional).	Detective function	ANN outperforms other algorithms used, consistently.
3	Characterization and detection of taxpayers with false invoices using data mining techniques (González & Velásquez, 2013)	For clustering :-SOM and -neural gas; For building: -ANN -Decision trees -Bayesian networks	Anonymous VAT data - Chilean tax administration	Characterize and detect potential users of false invoices in a given year	Detective models	Combination of supervised methods recommended: ANN labels fraud correctly, highest odds selected by Bayesian networks & Decision Tree

	Author and paper	Models	Data	Purpose	Model function	Results/Findings
4	Data Mining in Tax Administration - Using Analytics to Enhance Tax Compliance (Martikainen, 2012)	-	Project in Finnish tax administration OECD conference insights.	Enhancing tax compliance	-	Data mining methods improve tax compliance
5	Detection of Anomalies in Accounting Data using Deep Autoencoder Networks (Schreyer et al., 2017)	Deep Auto-encoder neural networks	Real datasets of journal entries	Detect anomalous journal entries	Detective approach	Approach is relevant to capturing highly relevant account entries
6	Unsupervised Machine Learning on Tax Returns (Gedde & Sandvik, 2020)	K-means clustering DBSCAN One-Class SVM Autoencoders Boosted-trees benchmark	Real anonymous tax returns, Simulation of the real dataset	Detect anomalous tax returns	Detective models	Methods improve upon manual techniques, but not suitable for stand-alone solutions
7	Optimizing Tax Administration Policies with Machine Learning (Battiston et al., 2020)	Neural networks, Random Forest, Decision Trees Linear models	Income data on self-employed and sole proprietors in certain regions, for 2007 – 2011 from Italian Revenue Agency	Determine predictions for sole proprietors and individuals with a view to enhancing the design and implementation of tax administration policies	Predictive models	Quality approach derived with application in 'proactive policy' in nudging taxpayer behaviour

	Author and paper	Models	Data	Purpose	Model function	Results/Findings
8.	Big Data Analytics for Tax Administration (Mehta et al., 2019)	Benford's analysis, ratio-analysis, Spectral clustering, time series analysis, Logistic regression	VAT and GST* data from Indian Tax Administration. Further data from online marketplaces, other government agencies and banks	Identify tax return defaulters	Detective function	Increased tax compliance levels, 16% increase in tax revenue collected
8	Transaction aggregation as a strategy for credit card fraud detection (Whitrow et al., 2009)	Random Forest, SVM, Logistic regression, KNN	Real historically labelled credit card transaction data from two banks	Assessing the effectiveness of transaction aggregation as against transaction-level detection with a view to reduce cost of preprocessing transaction level data in financial fraud	Detective models	Transaction aggregation effective in many, but not all cases. It is most effective when Random Forest is used for classification
9	Using data mining technique to enhance tax evasion detection performance (Wu et al., 2012)	Association rules	VAT data from Taiwan Revenue Authorities	Tackle VAT evasion	Detection performance	Association rules applied enhance tax evasion detection
10.	Using deep Q-learning to understand the tax evasion	-Q-learning -Deep neural network	Greek (corporate) tax returns	Determine taxpayer behavior to help form tax		Models developed can aid tax policy

	behaviour of risk-averse firms			policy decisions		formulation.
	Author and paper	Models	Data	Purpose	Model function	Results/Findings
11.	Tax Fraud Detection for Under-Reporting Declarations Using an Unsupervised Machine Learning Approach (de Roux et al., 2018)	Spectral clustering Kernel Density estimation	1,367 tax declarations of building projects in Bogotá, Colombia	Determine the effectiveness of unsupervised methods in tax evasion detection	Detective models	On previously specified premise, models derived do not miss on identifying faulty tax returns
12.	An Empirical Method for Discovering Tax Fraudsters: A Real Case Study of Brazilian Fiscal Evasion (Matos et al., 2015)	Association rules Dimension reduction methods (PCA and SVD)	Taxpayer data from Brazilian Fiscal Agency	Rank taxpayers according to their propensity commit fraud	Predictive model	Fraudulent tax payers are identified with 80% accuracy
13.	Signaling tax evasion, Financial Ratios and Cluster Analysis (Dias et al., 2016)	K-means clustering	Portuguese companies operating in construction minerals sector.	Classify taxpayers according to possible mix of tax evasion	Predictive models	Classificatory capacity exists taking
14.	Fraud Detection in Tax Declaration Using Ensemble ISGNN (K. Zhang et al., 2009)	Ensemble ISGNN (Iteration Learning Self-Generating Neural Network)	Financial data of sampled enterprises	Detecting legitimacy of declared enterprise tax	Detective model	Proposed approach is effective with high f1 scores

The data and purpose columns in the table are added to give greater insight as to the investigations being done in these papers, to guide the deductions of this work. Also,

highlighting the purposes will help determine if a model is suitable for further investigation if the purpose of the reviewed model closely matches the aim of this study.

It is noted that the evaluation metrics given in the summary above is not uniform. Some papers delivered actual percentages of improved detection, while others gave a summary of findings without publishing details of their findings. We will make generalisations based on what is obtainable and base our recommendation from there.

2.1.8 Selected Models Explained

The OECD (2004) report on tax compliance described neural networks and regression trees as 'typical' data mining techniques used in tax fraud detection. From the review on literature exploring tax fraud detection using Data mining/ ML methods, it was observed that there is a dominance of supervised methods over unsupervised. The need for unsupervised methods has been observed as occurring because of a general lack of historical data for tax administrations who do not capture audit information automatically and also because of the high cost of obtaining historically marked data (de Roux et al., 2018; Gedde & Sandvik, 2020).

Seven models were observed as the most frequently occurring in literature. Therefore, two unsupervised and five supervised methods with highlights of certain models developed within these methods will be explored. The techniques are reviewed and perspectives for learning are obtained on data analysis in general and tax administration specifically. The models used in tax fraud detection are explained below:

- **Self-Organizing Maps (SOM)**

This model is a category of the Artificial Neural Network (ANN) in unsupervised learning approach used to "produce a low-dimensional, discretized representation of the input space of the training samples" (Ge et al., 2017). Each SOM is comprised of nodes and each node associates itself with a "weight vector of the same dimension as the input data vectors, and a position in the map space" (Ge et al., 2017). Typically, SOM mapping is done from a high-dimensionality to a low-dimensionality data space (Ge et al., 2017). The nature of SOM makes it useful for data visualization, dimensionality reduction etc and these features are crucial in unsupervised approach to tax evasion detection.

SOM, was used in conjunction with Neural Gas in (González & Velásquez, 2013) as a method to group similar taxpayers from small medium and large companies according to behavioural patterns. Decision trees were applied subsequently for classification.

- **Artificial Neural Networks (ANNs)**

These are a consortium of related models, inspired by the structure of biological neural networks. They are described as structures of interconnected "neurons" that transmit messages with each other (Ge et al., 2017). Numeric weights are allocated to each neuron and calibrated based on experience which makes them "adaptive to inputs and capable of learning" (Ge et al., 2017). Neural Networks learn from observed data and can use this to approximate any linear/non-linear function (Ge et al., 2017). Neural networks have been considered the "most successful ML method by both academia and the industry due to its versatility" (Battiston et al., 2020). Although some uses of ANN lies in unsupervised learning approach, most of its applications lie in supervised learning approach examples of which are regression modelling and data classification (Gedde & Sandvik, 2020). Fig 4 below illustrates a three-layer neural network.

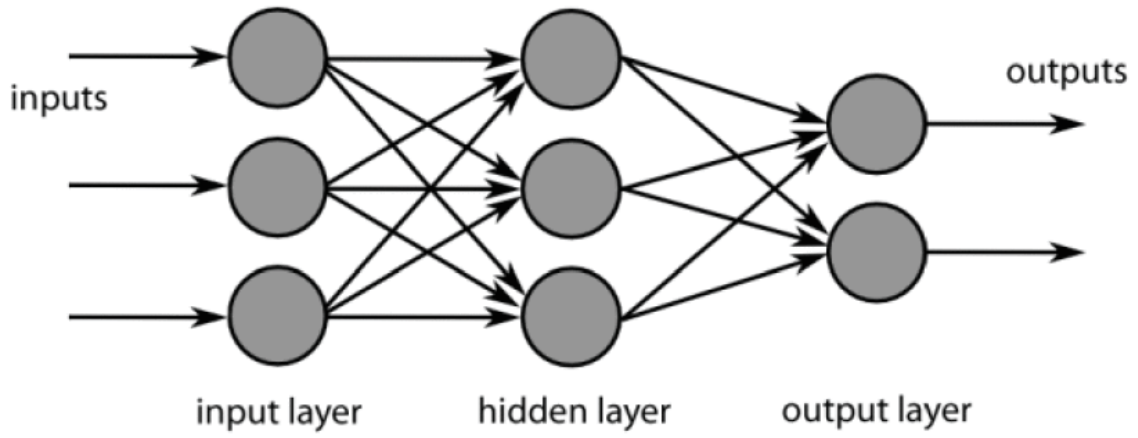

Figure 3: Structure of a three-layer neural network (Ge et al., 2017)

Neural networks have been applied to a wide array of uses and models have been built based on this technique to tackle various issues of financial fraud including tax evasion in both supervised and unsupervised approaches.

In this review, ANN was used in a supervised manner in (Dutta et al., 2017; González & Velásquez, 2013). while models developed from this approach in an unsupervised manner is seen in the ISGNN model implemented by (K. Zhang et al., 2009) and the SOM used in (González & Velásquez, 2013). Another unsupervised variation of this technique is the

auto-encoder neural network. (Gedde & Sandvik, 2020; Schreyer et al., 2017) used it an unsupervised fashion to detect anomalous journal entries and taxpayer information, respectively. It uses a specific "feed-forward multilayer neural network that can be trained to reconstruct its output" (Schreyer et al., 2017).

- **Decision trees**

Decision Tree models form a model of decisions based on the actual values of data attributes (Brownlee, 2019). It serves as a decision support tool with which a tree-like structure is used to show the connection between variables and subsequently make decisions (Ge et al., 2017). "They are used to determine a strategy most likely to meet the aim". They are useful in classification and regression problems and their speed and accuracy makes them resorted to often in ML (Brownlee, 2019). (Dutta et al., 2017; González & Velásquez, 2013) utilize this approach for classifications with (González & Velásquez, 2013) applying it after implementing unsupervised approaches.

- **Random Forest**

Ho (1995) proposed Random Forest in 1995 as a means to improve the accuracy of unseen data in the face of complexity when dealing with single decision trees. Typically, RDF model is formed using multiple decision trees based on the raining dataset an gives different outputs when used for different purposes (mode class when used for classification and mean prediction value when used for regression) (Ge et al., 2017). This technique has proved extremely effective in the investigations executed by (Battiston et al., 2020; Whitrow et al., 2009) in their detective ability, and, as such, is highlighted in this research as one of the models to be considered in tackling financial fraud.

- **Bayesian Networks**

Here, Bayes' Theorem for classification and regression problems are applied (Brownlee, 2019). Naïve Bayes as used in (da Silva et al., 2016; Dutta et al., 2017; Whitrow et al., 2009), although regarded as a more traditional method of classification, is the preferred option over a general Bayesian Network, as training the latter proves to be complicated and Naïve Bayes has the reputation of being performant "despite the violation of its independence assumption" (Whitrow et al., 2009).

- **Support Vector Machines (SVMs)**

This class of supervised learning models which produce a set of hyperplanes in a high dimensional space based on the different categories of data samples that can be partitioned (Ge et al., 2017). Here, kernel functions are used to "project input data onto high-dimensional feature spaces, wherein it's easier to separate instances linearly."(Demush, n.d.). This feature makes them a good option for fraud detection problems. As a method, it is originally suited for supervised learning in two-class categorization (Gedde & Sandvik, 2020). Although, it was superseded by other techniques in the literature reviewed in this study (Dutta et al., 2017; Whitrow et al., 2009) in its supervised form, it remains a viable technique which can be applied to classification or regression problems.

However, it can also be adapted to both semi-supervised and fully unsupervised learning approaches. A form of semi-supervised learning approach known as One Class Support Vector Mechanism (OC-SVM) where the goal is to differentiate "one minority class from the majority of observations, as opposed to distinguishing two classes from each other" (Gedde & Sandvik, 2020). In this study, the OC-SVM was considered a better approach on the initial analysis because of its relatively high precision despite reduced number pf anomalies detected (Gedde & Sandvik, 2020). Considering the high cost of false positives, it is considered as a cost effective method (Gedde & Sandvik, 2020).

- **Clustering**

(Jain, 2010) defines Cluster analysis as, "the formal study of methods and algorithms for grouping, or clustering, objects according to measured or perceived intrinsic characteristics or similarity". It is a form of unsupervised Machine Learning method which groups unlabelled data subjects as an initial attempt to understand a dataset (*What Is Clustering?*, n.d.). It is also applied in fraud detection (Mehta et al., 2019). Two types of clustering will be explored in this research as seen in the relevant literature.

1. K-means clustering:

This is one of the most popular models used for clustering in datasets. It divides the dataset into k-clusters so that each partition is located closest to the cluster with the nearest mean (Louridas & Ebert, 2016; Ge et al., 2017). K-means clustering was explored in taxpayer groupings by (Gedde & Sandvik, 2020) in their method comparison and also in

(Dias et al., 2016) where a cluster analysis methodology was proposed to apportion elements into homogenous groups to enable effective identification of at-risk companies.

2. Spectral clustering:

This technique is used when "finding maximal subgraphs or cliques in weighted graphs" (de Roux et al., 2018). (de Roux et al., 2018; Mehta et al., 2019) use this method in an initial exploration of their datasets.

2.1.9 Review summary to recommendation

A mind map of the selected techniques and methods with the results sought by their implementation is given below:

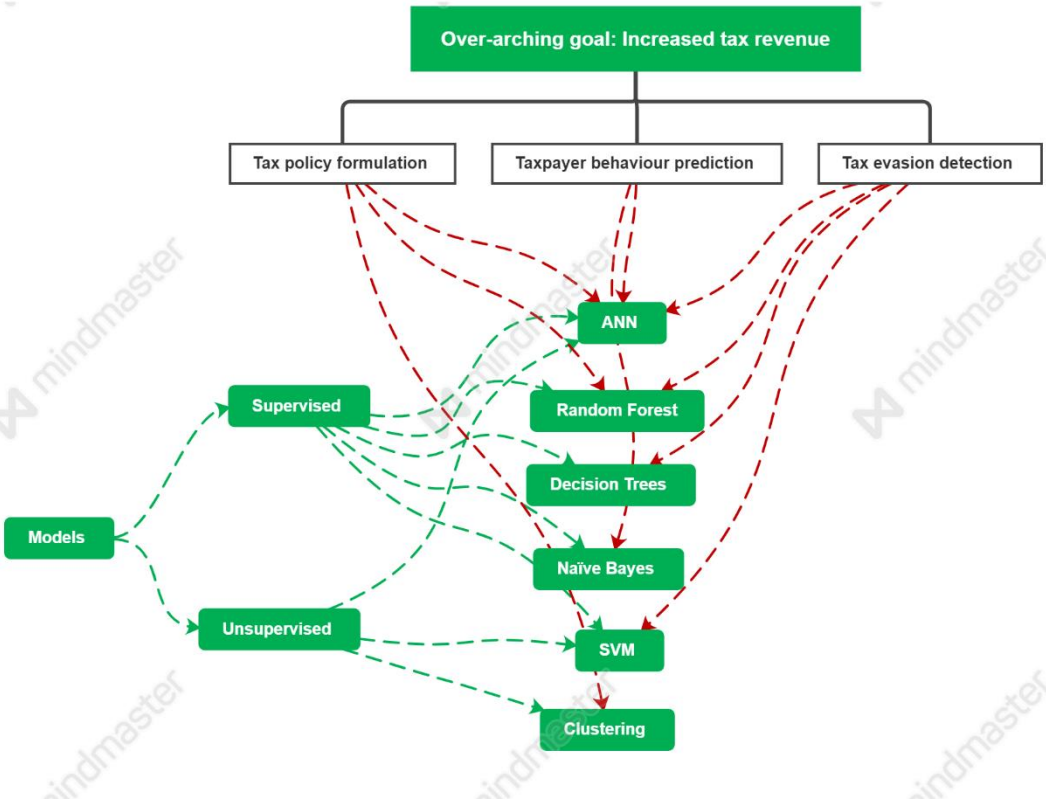

Figure 4. Model summary recommendation depending on tax objective. Source: Author

The first inference of this research is on supervised learning models. In the presence of marked (labelled) data, a semi-supervised approach is proposed. Initial data clustering

using spectral clustering is advised. Subsequently, a model based on ANN is to be used to conclude the analysis. Spectral clustering is preferred over K-Means for of the purpose for anomaly detection. It has been used in literature as a pre-classification method in both supervised and unsupervised learning approaches. In cases where the preferred results pertain to tax policy formulation, as seen in (Dias et al., 2016), K-Means clustering method can be resorted to. Again, ANN is the recommended method because of its versatility and adaptability to different datasets.

For datasets without historically marked data, the SOM approach is preferred. This is because of its high dimensionality reduction capabilities and its visualizations. As it is a variation of ANN, this research further recommends unsupervised modes of Neural Networks as it proves highly effective and is selected for its adaptability to unbalanced data and ability to give high accuracy results despite noise level.

Independent methods for tax evasion detection are not proposed under unsupervised technique in this research. This is because the works from literature do not provide sufficient evidence to show that they can be applied as stand-alone methods. Therefore, the methods proposed here are to be used as support mechanisms to assist tax officials in better targeting corporations to investigate further.

These propositions are given taking into consideration three goals as seen from literature: detecting faulty returns, predicting taxpayer behaviour and formulating tax policies. The overarching goal of increasing tax revenue is also considered and forms the basis for the recommendations given.

2.2 Theoretical framework

2.2.1 Role of Tax consultants in tax compliance and evasion

The theoretical framework seeks to explore the Tripartite relationship between the taxpayer, the tax authority, and the tax advisor/consultant in this study. To the best of our knowledge, there has been limited work done in investigating this relationship. More importantly, this work intends to inquire into the role played by the tax consultant in tax evasion. This is especially the case in corporate income tax to which this work is based upon as most corporate entities utilise the services of tax consultants in their tax returns preparation.

The traditional approach to viewing tax evasion stems from the bilateral relationship between the taxpayer and the tax authority where the tax authority determines the "audit strategy and punishment policy" and the taxpayer, in a fairly uncertain state, has to determine the probability that he will be audited (Battaglini et al., 2019). This approach, however, excludes the crucial role of the tax consultant who, as an expert in the assisting the taxpayer to implement the varied tax rules, mediates the tax compliance decision (Battaglini et al., 2019).

Battaglini et al. (2019) conducted a study to examine the relationship between tax evaders and their tax consultants. They highlight studies that show the possibility of the influence of tax professionals on "the formation of expectations on enforcement probability and helping to shape ethical standards of their clients". Their study shows a strong relationship between levels of tax evasion by companies being served by the same tax accountant. Models developed by them shows two main possibilities: Tax professionals may serve either as tax evasion facilitators or as information hubs.

Although these propositions are merely suggestive, they have strong policy implications and can guide the tax authority in better directing their tax compliance efforts. On the one hand, if the tax consultant serves as an information hub (Battaglini et al., 2019), compliance levels could significantly increase as it is assumed that the taxpayer views the tax professional as an expert and such information will boost his propensity to comply with tax regulations. On the other hand, if tax professionals act as tax evasion facilitators, the tax authorities can also act strategically in this regard. Battaglini et al. (2019) recommend that the tax authorities could utilize the tax advisor's role as information hubs to disseminate tax regulation effectively and achieve higher levels of tax compliance.

2.2.2 Adoption of ML techniques in Tax administration

(Hauptman et al., 2014) note that voluntary compliance can be achieved if the right communication strategy is adopted by tax officials. They base their research on the tendency for strategic communication on punishable actions to increase the perceived detection likelihood and, therefore, deter tax evaders. This being done with the aim of achieving tax compliance at a reasonably lower cost. Models will be explored here where communication is done with the purpose of preventing incidences of collusion.

2.2.2.1 Excerpts of successful applications.

Several countries have used ML techniques in their tax administrative with various levels of success, the most notable of these, as seen in the revenue obtained, are Australia, Norway, and Canada. We will explore the models used by these countries and the preconditions necessary for its implementation.

- **Australia: Model for deep analysis and communication with taxpayers**

The Australia Tax Office (ATO) implemented a series of ML techniques designed to assess taxation in real time with capabilities for error detection, and treating this by communicating with tax payers using automated tools (Mendez, 2020). Prefilled tax returns are used in the Australian tax administration (Office, n.d.). Supervised ML model, 'K-type nearest neighbour' is used to compare a taxpayer's return with those of people in similar financial situations. Any claim significantly different from what is expected will have the taxpayer receive a nudge to recheck their submissions (Office, n.d.). The tax year for 2017 filings saw about 240,000 adjustments with increase in tax revenue to the tune of $133 million (Office, n.d.).

In summary, they were able to execute a tax system where risk detection and analysis was done, including communication with taxpayers and treatment- all at a distance from the taxpayer (Mendez, 2020). The positive impact of this is shown in the increase in revenue obtained and the reduction in tax official hours spent in the tax administrative process. The use case is especially useful in the Nigerian case where corruption exists predominantly in some steps of tax filing. The implementation of such techniques which make all these steps possible at a distance from the taxpayer will significantly reduce all incidences of collusion, and an increase in tax revenue may be achieved.

In a qualitative observation carried out by (Martikainen, 2012) in his masters' thesis in 2012, where he attended a tax conference organized by the OECD. In this conference, the ATO already showcased a robust data mining and analytics capability based on "predictive risk models, large-scale visualizations, and the use of third-party data and social network analysis" (Martikainen, 2012).

A further deconstruction of the use of analytics at the ATO by the Commissioner Michael D'Ascenzo in the same conference highlighted the models used by the ATO in their tax administration activities:

- Lodgement models
- Debt models
- Models to detect high-risk refunds ("which includes a scalable social network discoverable algorithm")
- Clustering techniques

A further use case presented in this conference by the Irish tax administration drew from the use of data mining methods in tax administration in a manner analogous to the banking sector where data mining models are used to determine customers who are at most risk of loan default and using similar models to deduce taxable entities who are susceptible to tax evasion and determining this through audit profiles of past cases. The deployment of this method showed a 75% success rate in predictive capabilities (Martikainen, 2012).

- **Norway: Income and income tax deduction prediction model.**

Norway's tax administration, SKATT, developed a project to identify tax deductions due to individual taxpayers. SKATT's dataset was used along with an ML technique which permitted the estimation of "the level of income, its composition, the level of debts and the family situation of the individual taxpayer in order to establish the origin of the legal deductions in the annual income tax return" (Mendez, 2020). Also included in this ML model was the use of classification models which were predicated on 'sentiment analysis' and 'feeling analysis' algorithms to "perform text analysis and natural language processing available in social networks to classify words or phrases as positive, negative or neutral" (Mendez, 2020).

During the development of the Pilot Project, the ML algorithms used include: Linear regression, Decision tree with random forest, and Neural networks (Mendez, 2020). The results obtained from individual uses of these models was unsatisfactory as it did not meet the accuracy standards of the administration. Consequently, an 'assembled model' was developed using elements from "'neural network', 'random forest', 'gaussian processes', 'two stage random forest' algorithms" with an added focus on predicting what taxpayers

are entitled to deductions in addition to the amount of deductions that was the earlier focus (Mendez, 2020).

The possibility afforded by the 'assembled model' was the reduction in the effect of variables with relatively insignificant presence in the datasets with a preference for datasets related to deductions (Mendez, 2020). A major warning resulting from this model was the 'black-box' feature it represents, as it was difficult to determine what occurred within the assembled model. Despite this, "the proof of concept established that the model correctly predicted the income level in 86% of the cases and in the case of deductions in 76% of the cases" (Mendez, 2020).

2.3 Summary

This chapter considered earlier studies on tax evasion, e-tax process, the current e-tax situation in the FIRS and the challenges faced in its execution. It progressed to explain the definition of Machine Learning and its related concepts. It also showed the different categories of ML techniques with systematic review of some techniques used in tax administration to combat tax fraud, the features which were present to give the results that were obtained in the different cases in which they were used. Table 2 gives a summary of this data. A recommendation framework was also proposed from reviewed literature and considering tax revenue objectives. This is seen in Figure 2.

The theoretical review explored the tripartite relationship between the taxpayer, tax authority and the tax consultant, highlighting the role tax professionals play in tax compliance. It also examined the adoption of ML techniques in taxation, after exploring the current ML techniques used by other tax administrations in the first section. It advanced to investigate the tax administrations that attained notable successes experienced by some countries. The models employed in the tax analysis, the features present and the results obtained were given.

3. Methodology Description

3.1 Introduction

This chapter aims to provide a thorough explanation of the research methodology and design selected for this study after a detailed review. A case study has been selected to provide empirical basis for this research as it is understood that the systematic review provided in the preceding chapter is inconclusive in deriving the research objectives. Therefore, this chapter will provide a comprehensive description of the selected design as well as give the necessary conditions for an effective research.

This chapter is divided into six sections. The first section connects the focal objective of the study to the research questions. The second section elaborates on the case and subject selection. It also specifies the definition of case study research and gives and its suitability to the current study. The third section thoroughly explains the data collection procedures to be used and the rationale for their selection. The fourth section gives the data analysis and validity procedures. Finally, the concluding section gives a chapter summary.

3.2 Research Questions

The central objective of this thesis was to discover the applicability of Machine Learning models to the improvement of the tax administration process- specifically to aid the detection of evasive tax returns for further audit and the advancement of a framework for its implementation. Furthermore, this research aims to investigate the tripartite relationship between the tax official, taxpayer and the tax professional. Finally, it aims to give recommendations to the Nigerian state on the applicability of these techniques having highlighted the elements necessary for its implementation, possible challenges to its implementation and expounded upon the crucial role of the tax professionals in compliant tax administration.

The literature review exposed the usefulness of using data mining methods in financial fraud and anomalous returns detection. It also revealed the practicality of using the large amounts of data accessible to the tax officials to uncover patterns that may help in tackling tax fraud. As the literature encompasses different models explored in different countries

and situations with different objectives, it became necessary to extract a framework for implementation based on the specific objective sought by a particular revenue body. This formed the major research question for this study.

RQ 1. How can the application of Machine Learning techniques provide benefits in the tax administration process for companies?

In order to carry out a comprehensive research, this main research question is divided into four sub-research questions:

- RQ1.1. How to examine the influence and effect of ML in tax administration in companies?
- RQ1.2. How the use of ML techniques in tax administration will be received by the stakeholders for implementation?
- RQ1.3. How to determine the key criteria that will measure the benefits obtained from the implementation of ML techniques in tax administration?
- RQ1.4. What ML model(s) can be used to achieve the tax objective for corporations?

Further questions arising from these sub-research questions are given below. These questions will form the basis for the interview questions.

The first sub-research question here has been highlighted in chapter 2 of this work. It is further broken down into the following questions which would form the basis for the interview questions to be used in this work:

- What is the current tax administration process in companies preceding the application of ML models?
- What features represent an ideal tax administration?
- What elements inhibit or advance the implementation and usability of this ideal system?

These questions help understand the current state of the tax administration, highlighting all loopholes present, to identify what an ideal system should entail and finally aims to identify what factors may impede or accelerate its attainment.

The RQ1.2. is developed into the following questions which would also form part of the interview questions:

- What are the roles of key stakeholders in tax compliance?
- What are the current features of those stakeholder roles?

- What is regarded an improvement in stakeholder activity in the tax administration process?

In examining stakeholder roles, it is necessary to highlight its present characteristics with a view to understanding what improvements need to be made to reach an ideal tax system. The RQ1.3. is disaggregated to the following questions:

- What are the criteria that establish the effectiveness and efficiency of ML application in tax administration?
- What factors (metrics) show that ML in tax would give benefits (reduce tax leakages and increase financial revenue) in tax?
- What would make a 'significant' improvement in the tax administration process?

Finally, following RQ1.4, as this research aims to provide a framework for the application of ML models in tax administration, the ML models that can be adopted considering the tax system described in the case will be investigated. Also, the tax objectives which must be specified before initiating changes are examined. Consequently, RQ1.4 has the following divisions:

- What are the specific tax objectives that underlie tax evasion?
- What models can be used to best tackle tax evasion?

These questions give insight to the impact of ML application in achieving optimal tax compliance and have been answered in the systematic analysis done in the preceding chapter.

Based on a further objective of examining tax professionals' role, a second main research question is proposed:

RQ 2. How can tax professionals facilitate or prevent tax evasion in corporations?

Again, to provide a thorough exploration of this question, a sub-research question is proposed:

- RQ2.1 How to determine the influence of the tax consultants in the submission of compliant tax returns by corporations?

Further questions arising from this sub-research question is given:

- What is the perceived role of the tax consultant in the tripartite relationship?
- What is the perception of the tax consultant in the eyes of the corporation?
- What factors hinder or encourage the tax evasion facilitator role in tax consultants?

This sub-research question gives a deep dive into the perceived and actual role of the tax consultant and explores what elements are responsible for this assumed role in the crucial relationship. This research question will be investigated through qualitative analysis.

3.3 Case Study Selection

As a qualitative approach is relevant to providing answers to the research questions, the case study design was essential to hone the research results and derive meaningful insights from this study. This is because case study permits the close examination of data "within a specific context" (Zainal, 2007). Case studies examine data at the "micro level" by examining only a small proportion of elements of interest (Zainal, 2007). Case study research method is suitable to recognize patterns from complex phenomena and generally contribute to knowledge in that phenomena (Krusenvik, 2016). Zainal (2007) describes it as a robust method suited to "a holistic and in-depth investigation" of a specific phenomenon. It is, therefore, useful in this research to use case study in closely examining the subject of interest to derive meaningful inferences.

Case study methodology has been criticized for lack of depth in deriving results (Krusenvik, 2016), and as a result, a careful development of the case study design is imperative (Zainal, 2007). Runeson et al. (2012) further opines that appropriate context must be given when developing a case, so as to make the "protocol" available to other researchers as this will give generalizations that will be useful for other researchers. A single case or multiple case can be selected depending on the research interest. Gustafsson (2017) gives the identification rule by stating that multiple case study method is applied "when a study includes more than one single case". One major disadvantage of the single case study, however, is its ineffectiveness in giving a solution which can be generalised (Zainal, 2007). In this case, single case studies can be combined with other methods to provide requisite results (Zainal, 2007). This concept, known as triangulation, is also recommended by Runeson et al. (2012) who state that it provides a broader picture, strengthens the validity of the research and is imperative when using qualitative data. It is based on this that this research will use a case study approach which is relevant to the study being carried out.

(Krusenvik, 2016) notes that case studies, similar to other research methodologies, complements the advantages and drawbacks of other techniques and its choices of use

depends on the phenomenon being researched and the results sought. Three methods of case study research as given by Yin (2018) are exploratory, descriptive and explanatory case studies. More suited to this research is exploratory case study approach which investigates the patterns in data with a view for initial examination of a topic of interest, which then sets the pace for more in-depth research subsequently (Zainal, 2007).

It is imperative to avoid juxtaposing case study with qualitative research as case study can also be used with quantitative data (Yin, 2018). An important factor in case studies is its understandability which should exist to such a high degree as to enable other readers implement the research in their respective studies (Stake (1995) in Gustafsson, 2017). Case studies are "open ended" and are suitable in cases where obtaining an exact solution is unachievable (Gustafsson, 2017). In qualitative research, the subjective views of the researcher are functional to the results obtained and it therefore has an "interpretative paradigm" (Starman, 2013). This then makes a case study more qualitative in nature even though it could be both qualitative and quantitative (Starman, 2013).

Case study research has the advantage of not being limited to a pre-existing datasets because of their ability to "analyse qualitatively complex events because they do not require many cases" (Starman, 2013). Single case studies also have the advantage of a detailed examination of "causal mechanisms in individual cases" and are also able to "accommodate complex causal relations" (Starman, 2013).

Apart from the problem of lack of the ability to generalize, case studies also have the purported disadvantage of producing results that conform to the researchers preconceived ideas (Starman, 2013). This is because, while there are general doubts towards objectivity in scientific research in general, case study is suspected to provide more room for subjectivity (Starman, 2013). A counter point to this lies in the thorough explanation of the research, especially the analysis process as it is in this phase that concepts are determined (Starman, 2013).

Runeson et al. (2012) opine that the case study method provides research results which are difficult to obtain using other research methods. They further state the applicability of case study research in situations where the technology is introduced or refined to improve a particular process is suitable.

On this premise, case study research method is selected to answer the research questions. The interviewees were referred from the researcher's personal network. The interviewee selection was also done with an attempt at region variation to get robust results.

3.4 Data Collection Methods

Data collection is simply defined as the means by which the data relevant to the research are collected. Research validity is to be obtained by adequate coverage of data sources to enable data triangulation (Runeson et al., 2012). Data triangulation is the use of multiple sources of data which provide additional sources to evidence to produce unbiased results (Runeson et al., 2012).

This research will use semi-structured interviews and document review as sources of evidence.

3.4.1 Interview

(Runeson et al., 2012) describe interviews as common and, nonetheless, important sources of data in case study under software engineering. As this is one of the major sources of data in this research, it is imperative that the people to be interviewed are appropriate to give relevant input to the research (Runeson et al., 2012). To ensure proper data coverage, it is also important to have a selection of interviewees who represent a sufficient representation of the subject under study. In this research, attempts are made to interview FIRS officials and tax consultants from more than one geopolitical zone in Nigeria. Informed consent to record these interviews were obtained to maintain ethical coherence in this interview process as suggested in (Runeson et al., 2012).

In this research, as the interviewees cannot answer each research question, the interview questions are derived from the research questions (Runeson et al., 2012). Also, a semi-structured interview is executed in this research, as the interview questions were planned beforehand, but not necessarily asked in the same manner as was planned (Runeson et al., 2012). "Saturation" is used as a basis for selecting number of interviewees. (Runeson et al., 2012) define saturation as the lack of new information with the addition of new interview subjects.

Interviews are done to uncover knowledge known only to the interviewees that the researcher is unaware of. Interviews are conducted to understand phenomena from people actually working in the context of the case (Runeson et al., 2012). For this research, the interview was necessary to validate the information derived from the earlier executed document review.

Semi-structured interview was executed in this research, with preliminary questions guiding the interview and the interviewer asking follow-up questions, as necessary. This was important to delve deeply into the case and understand its features and challenges. The questions stated were given to provide adequate coverage of the research objective. The researcher exercised minimal control over the questions to maintain the interview structure and ensure data validity.

Ten (10) interviews were conducted with Tax consultants and FIRS officials. The interviews were conducted after scheduling appointments and both parties meeting at the scheduled time. Interviews were conducted using (recorded) zoom video conferencing software and WhatsApp voice call using an external device for recording. Interviewees were asked for consent prior to recordings. A transcription software (Otter.ai) was utilized for transcription purposes as all interviews were conducted in English. The computer-generated transcripts were subsequently reviewed and corrected for accuracy. The interview responses were transcribed, analysed, and coded using NVIVO software.

3.4.2 Document review

This data collection method involves examining existing documents which may include business documents like financial statements, letters, emails and personal messages and administrative documents like newsletters and minutes of meeting. As a precursor to the interview, a systematic review was done to summarize the existing work on the use of Machine Learning and data mining in tax administration. Kitchenham & Charters (2007) give one of the reasons for conducting systematic reviews as “to provide a framework/background in order to appropriately position new research activities”. It is on this premise that the review was done on this research.

In presenting findings from case study research, it is important to avoid data that is irrelevant to the research question. In addition, the results obtained are to be compared with published literature or existing data (Gustafsson, 2017). Document review is advantageous in that when consistent results are obtained, they provide evidence that the concept being researched is robust (Kitchenham & Charters, 2007). On the other hand, when there are inconsistencies in the findings, the variations discovered can be further researched.

The review done in this research focused on the usefulness of one ML technique over the other in this context of tax fraud detection. As there are relatively few publications done on ML in tax administration, it became imperative to extend the context of the document collection to accommodate studies on fraud detection in general. This study utilized academic articles, journals, government publications to expound upon the theoretical aspects of the study. The review covered topics on the various ML methods, tax evasion and fraud, the current tax system in Nigeria, digital tax systems and the application of ML models to areas of tax evasion detection, taxpayer behaviour prediction and tax policy formulation. The review also explained on the influence of tax officials in facilitating tax evasion.

The review gave a detailed analysis of the tax evasion problem faced in Nigeria. It also gave case studies of countries which had successful implementation of ML models in their tax administration process. There was a lack of theoretical evidence as to the state of CIT administration in Nigeria and this precipitated the need to gather empirical evidence by the conduction of interviews to know the depth of the situation.

3.3 Data Analysis

In single case study research, the two major ways of analysing data include the use of "statistical inferential analyses" and the "visual inspection of graphed data" (Nock et al., 2007). Proper data analysis is important in deriving appropriate solutions under case study research. In this research, interviews are used, multiple quotes are used, which may be used to answer more than one research question, it is important to not "complicate the chain of evidence" (Runeson et al., 2012).

This phase uses data collected to understand what occurred in the studied phase. RQDA qualitative tool (for analysing data in text) is employed for analysis as stated earlier. For qualitative data, data analysis is to be carried out simultaneously with the data collection process (Runeson et al., 2012). This way, it is easier to observe what additional data is needed and to spot when "saturation" has been reached. The interview done is transcribed into text using the software otter.ai for further analysis and coding.

To create a descriptive form of data analysis, Open Coding was used to analyse the data. It involves attaching labels to the ideas (codes) generate from the interview and further dividing them into various categories.

For proper analysis of qualitative data in case study, (Runeson et al., 2012) propose the following steps (which may be iterated):

- Data collection: In this case, through interviews.
- Data coding: A code representing a theme is assigned several pieces of text. Each piece of text can also be allocated to more than one code.
- Hypothesis generation: Identified after examining material received in the above steps.
- Generalization: With further data collection and repeat of the above steps in an iterative process, a "small set of generalizations can be formed".
- Reporting of the knowledge derived.

For exploratory case studies, hypothesis generation techniques, which extract hypothesis from the data is recommended (Runeson et al., 2012). In this case, prior definition of multiple hypothesis is not encouraged, rather hypothesis are to be discovered from the data (Runeson et al., 2012). This research will analyse text from interview transcripts and generate hypothesis from there.

3.4 Data Validity

Validity of a study reflects how trustworthy the research is and the extent to which the study is not influenced by the researcher's bias (Runeson et al., 2012). It is a factor which

must be considered during the entire course of the case study (Runeson et al., 2012). Construct validity and reliability are used to provide validity in this research.

Construct validity is a test of how much the operational controls which determine data collection and analysis represent the mindset of the interviewer and interviewee (Azogu, 2018). Reliability entails producing similar results if the research was conducted by another researcher (Azogu, 2018). Therefore, the use of interview enables the construct validity and reliability construct procedures for this research.

3.5 Summary

This chapter gave a through explanation of the research design and methodology to be adopted in this study. The research objective was clearly stated, and the data collection and analyses methods clearly articulated. Furthermore, the validity of the data collection and analyses tools were examined. The discussion in the next chapter will be derived from the results of this chapter.

4. Results

4.1 Introduction

This chapter gives detailed explanation of the case and subject selection. It also provides analysis and inferences from the interviews conducted. The NVIVO software was employed to thoroughly analyse the interview outcomes. Finally, the chapter ends with a summary.

4.2 Case and Subject Selection

Taxation in Nigeria is administered by the Federal, State and Local governments of the country. The FIRS is the tax collection body of the Federal tax service and is the tax authority for the Companies Income Tax (CIT) which is the specific case of this study. Despite the FIRS being the principal tax collection body, the administrative system is generally inefficient as a large percentage all tax returns submitted by companies in the "tax net" are subject to manual investigations and reviews. To begin with, despite the prevalence of the compulsory Tax Identification Number for all companies in Nigeria, the FIRS has difficulty including all registered companies in the tax net.

A major factor contributing to this challenge is improper data capture and management methods. Different laws have been enacted in recent times to promote the adoption of e-taxation in Nigeria as a proper e-tax system is necessary for data mining to be based on. Although these laws are significant in achieving the tax objective, the issue of improper data management inhibits the gains arising from the execution of these laws. Scalability is also a major challenge to the current e-taxation system as the servers become overloaded during the tax filing period and some companies have resort to manual methods of tax submission.

Furthermore, corruption is a major culprit in the excessive leakages experienced in the tax system. The absence of adequate data makes it possible for corrupt FIRS officials to engage in nefarious activities that undermine the general tax collection purpose. Adequate data collection, analysis and management would reduce the likelihood of these practices. Appropriate punishable measures could also be established to serve as a deterrent to tax evasion, but that is beyond the scope of this work.

As tax revenue increase is the objective of tax administrations, a way of achieving this is to reduce the costs associated with the inefficiencies identified in the current system. Increasing tax revenue is imperative to fund the government's increasing activities while stemming budget deficits. It can therefore be deduced that using ML models trained with data will reduce the amount of human interaction in the system and reduce corruption and all leakages that exist. If this is the case, there are little to no field audits, costs will be greatly reduced, and an increase in revenue will be achieved.

The main objective of this research is to discover the importance of ML methods in tax administration with a view to increasing the tax revenue obtainable in the current system. It was also to ascertain the role the tax consultants play in facilitating/inhibiting tax evasion. This research was carried out using interviews as a primary means of data collection. The interviews were done to provide adequate understanding of the studied case, to obtain evidence on the research and to have direct contact with the current stakeholders in the tax administration system- tax consultants and FIRS officials. A total number of Ten (10) interviews were conducted from tax consultants around Nigeria and FIRS officials representing various areas in the country. It was imperative to get views from both the tax consultants and the FIRS officials to get rich perspectives from the two ends of the case being studied- the FIRS administration of the CIT. This was done with the aim to get their views on the current system and its challenges and also to understand what they would propose as improvement to it.

4.3 Presentation of findings

The findings in this section will detail the data derived from the interviews. This data will be aligned according to the research questions and given in the subsections of this chapter. The first subsection will highlight the background of the interviewees with a means to gauging the validity of their responses. The other subsections address important issues that answer the research questions given with a view to realizing the research objective. These subsections provide detailed explanations for the observations made during the interview.

The hybrid approach, which combines inductive and deductive methods of coding, was used as the method of analysis in this research. This means that, for our data analysis, there was a combination of codes pre-established based on the research questions of this

work (inductive) and codes generated after a rigorous review of interview transcripts (deductive). Code categories, which constitute a combination of two or more codes with close features, were developed during the course of the analysis.

4.3.1 General Description of the Respondents

This section will give the background of the interview respondents and explain part 2 of the interview guide. To establish the validity of the research, it was necessary to ensure that the respondents had requisite knowledge of the current system- both advantages and challenges.

The interview respondents included 7 tax consultants and 3 FIRS officials. All the respondents had experience exceeding 3 years in the Nigerian tax administration system with two respondents having over 10 years' relevant experience. Therefore, it can be concluded that the respondents for this study have enough experience appropriate to the information needed for this study.

4.3.2 Machine Learning and Tax Administration System for Companies

One of the major research goals was to determine the effect of ML methods application in tax administration. This section pertains to RQ1.1 which is, "How to determine the influence and effect of ML in tax administration in companies?" The interviews conducted gave good insights to answer this question and other questions that were derived from it. This section is divided into sub-sections based on the questions it pertains to and the codes used in our analysis. These sub-sections are given below:

4.3.2.1 Current Tax Administration System in Nigeria – FIRS's Administration of CIT

The question to be answered in this section is "What is the current tax administration process in companies preceding the application of ML models?". The literature review provides evidence of the level of human involvement in the process and also the maturity of the e-taxation system which gives a view to the data collection and management system. The typical tax audit system is highlighted below:

The tax audit process starts with self-assessment. Companies are allowed to prepare their own tax returns and submit this to the tax office before the deadline. For any given tax year, the company is typically given up to 6 months after the company's financial year

end to file their taxes. This tax paid in conjunction with the tax returns is examined by FIRS officials in what is known as a 'desk review' "*which is done basically for arithmetic accuracies*" (Respondent). When officials feel that the tax is under-declared, they send mails to the company asking for more documents to back up their submitted returns. If they are still not satisfied with the evidence provided in the documents presented, they proceed to the tax audit stage. If they are satisfied, however, the tax submission process ends at this point.

The tax audit stage is characterized by two meetings which is divided into the "audit planning meeting" at the start and an "exit meeting" at the end of the process. The former involves the FIRS asking the companies questions related to the business activities and also tax related questions. They also state their findings from the desk review and give the company the opportunity to provide "source documents" to back up their self-assessment. Typically, these documents are sent physically to the FIRS offices. The process may end after the company provides this, but for most companies, there will be frequent back and forth on the figures in the tax returns and the documents being presented. This oscillating dialogue can last anywhere from a few weeks to a month. The duration generally depends on how fast the company being audited is able to comply with the request for source documents.

The tax consultants upon being notified of the planned audit would then carry out a "health check" on their clients to determine if the FIRS' position is justified. A respondent clearly stated that sometimes, this may lead to the company having tax credits and other times, the company may need to pay more tax than was previously declared. In this case, the tax consultants can, "*write a letter saying, 'after carrying out a review, a review of all our books, we realize that we underpaid to the tune of 'this' amount*'" (Respondent).

The tax consultants clearly stated the importance of their presence during these tax audits. Due to lack of data and wide interpretation of the tax laws, the FIRS officials may want to impose the highest liability on the companies. "*They will have an assumption in their head and think this is how it is, but well because you are on ground it helps us to reduce whatever assumption they have. And it gives us an idea of what the liability will look like, because we've already done a health check*." (Respondent).

On a graver note, if the FIRS suspect the authenticity of the documents being sent or have a higher cause to believe that the company is involved in tax fraud, they may arrange physical visits to the company premises to carry out the tax audit. The purpose is to physically inspect the accounting books of the company and get greater access to documents that may not have been obtainable if done at a distance. These visits are referred to as "field audits" and can last anywhere between two to five business days per company. The FIRS officials typically request the source documents to back up the figures shown in the tax returns.

What makes it more cumbersome is the fact that when audits are carried out, the tax auditors are authorized by law to audit a company for up to 6 years prior to the current tax year in question. They can do this based on their judgement that such company may have not been audited in prior years or may have been involved in tax fraud in those prior years. This is sometimes problematic because of lack of proper data keeping by the FIRS when they send a notice of audit covering years that they have audited before. "*Well, because of how we have some loopholes in Nigeria, there are instances whereby FIRS doesn't have full knowledge of the years which you've been tax audited and the reason for that is not far-fetched. The reason is that, on average, they most times move officers around the offices from one location to another location and the person handling the file of a company in a tax year, may not properly document those years, so that there is expected to be doubt for future audits*". He however stated that in such situations, the company can send the FIRS the correspondences and the assessment that was raised and that closes the issue.

Finally, the FIRS arrive at a position based on this and this is communicated to the company. If under-declaration of tax is detected, the company is given the revised tax position with the fines inclusive. The company may either accept this revised position or refute it. If they refute it, they issue an "objection letter" to FIRS coupled with further documentation. A "reconciliation meeting" ensues next from which, generally, both parties reconcile and come to one position, "*and the client is not willing to pay whatever you see without documentation, we can't do anything. And also there are always back and forth, back and forth. Later on, there will be a level of compromise*" (Respondent).

In special cases, the tax audit process may get to a final stage of "Tax Investigation". The tax auditors, if they feel that they could not gather enough evidence to reach the true

financial position for tax purposes may refer the case to tax investigators. These tax investigators have more powers than the tax auditors. Tax investigators can request access to a company's internal systems while a tax auditor relies mostly on evidence presented by the companies themselves. Tax investigators can also extend past the 6-year period provided for the tax auditors and can investigate for an unlimited number of years as they deem fit and as is necessary for them to come to a "true and fair position" of a company's financials.

The tax laws in Nigeria have broad meanings, with numerous opportunities that can be taken advantage of to engender tax avoidance. This feature of the tax law makes interpretation malleable as FIRS officials may interpret it in a certain way to ensure maximum tax payment, while tax consultants may interpret in another way that enables them to legally pay the minimum amount possible. These differing interpretations are usually reconciled in the final reconciliation meeting where all parties: the FIRS officials, tax consultants, and the companies they represent all settle and finalize all aspects of the tax audit that was just conducted.

From the interview, it was also obvious that the beginning of digital tax is being experienced in the chosen case. some respondents stated that the COVID-19 pandemic forced some aspects of the Tax audit to be moved online. "*Right now, since corona came, a lot of reconciliation meetings are now held virtually*." (Respondent). Also, "*Now they are trying to move digital in the sense that some of these documents they ask for, they allow us send soft copy while previously we have to print everything*." (Respondent).

Therefore, from the above description, there are large number of instances where human intervention is utilized in tax assessment. "Best of judgement" positions arrived at are not based entirely based on data, but partly on the FIRS officials' perceptions and conclusions derived from the documents are presented to them and from experience. While this in itself is not bad, it is extremely cumbersome and not scalable. It is also prone to corruption and leakages.

4.3.2.2 Features of an Ideal Tax Administration

This section seeks to answer the question, "What features represent an ideal tax administration?" It explains what an ideal CIT process is according to the respondents. The interview results showed that the major feature of a good tax administration system

is the automation of the tax returns and submission process. This was expressed by the respondents both from the angle of speed and convenience in submitting tax returns and also from the angle of avoiding the hassle of the field audit process especially when audit is being done for prior tax years.

Another characteristic pointed here was proper data collection and management procedures. Most of the tax consultants interviewed stated that if there was proper data management system which was transparent on both sides of the officials and the corporations, their jobs would be easier as there would be less pushbacks from the self-assessment stage of the tax process. Even if the FIRS officials wanted clarification on something, there would always be data to back up all items featured in the tax returns.

Proper knowledge of tax laws was a feature mentioned by one respondent as a must-have in an ideal CIT system. According to her, this feature would ensure be fewer incidences of tax evasion as there are multiple opportunities for tax avoidance in the current tax law. In addition, a few respondents highlighted the lack of infrastructure faced in the country as a major culprit in all attempts to evade tax. If people believed that their taxes would be put to good use, they would be more willing to pay it. Therefore, a proper tax system would provide accountability for the taxes paid.

A respondent stated that an ideal tax system would involve the FIRS giving incentives for companies to want to pay tax. She described a system where it would be imperative to register with the tax authorities as the law can prevent a tax compliant company from doing business with other non-compliant companies. This may force the non-compliant companies to be more responsive to their tax obligations.

Another major feature needed to have a good tax system, is one where there is a uniform accounting/invoicing procedure for all companies backed by law, so that "source documents" needed can be easily retrieved. As one of the respondents put it, "*Only multinationals and big companies have highly automated systems with proper data collection methods and so they rarely get audited if at all and even so, they can easily retrieve data*." To this end, the digital responsiveness of the companies to proposed digitized initiatives in the tax area is also a significant factor of an ideal tax system.

Finally, a respondent noted that an ideal situation is one where corruption is reduced to the minimum level possible. According to him, "*And I actually mean in a place like Nigeria, there's a lot of tax leakages, I mean, instances whereby you have companies that should pay huge amounts of taxes, but because of inefficiency in place, the FIRS couldn't even track the money. There are instances whereby a company is expected to pay 500 million as tax liability, but because their staff (FIRS) went behind the government, and go and negotiate (with the companies), 'you know what? you pay 100 million to us pay 200 million to government and forget about the remaining' Well, with a level of digital taxation, or machine learning exercise, there won't be anything like that, because everything will be real-time. So that's the future. And I pray we get there very soon*."

4.3.2.3 Factors That Inhibit and Advance the Adoption of ML Methods in Tax for Corporations in Nigeria.

This subsection is dedicated to answer the question, "What elements inhibit or advance the implementation and usability of this ideal system?" Considering the fact that certain factors have to be present for the adoption of ML methods in tax as seen from the e-tax system described in chapter 2 of this study, this section aims to examine these factors in detail according to the responses from interviewees.

The inhibiting factors are given as follows:

- Lack of technical skilled labour force in the FIRS
- Lack of proper legislation.
- Paper filing processes in less tech-savvy companies.
- Lack of proper data management system by the FIRS
- Biased data fed to algorithms from the start.
- Fear of human intervention despite new measures adopted due to corruption.
- Infrastructure (Machinery and power supply, network).
- Lack of political will.
- Commitment challenges

One of the respondents noted the scalability issues in regard to e-taxation. She said, "*Sometimes, when companies try to file their tax returns online, they may have to file their tax returns manually before the deadline*". This is because most companies have 31st December as their accounting year end date and therefore have similar deadline for tax return submission and the servers may be overloaded during such periods. The online

system may not be obtainable in this instance and companies may have to revert to manual systems.

Another major point that was noted was the lack of uniformity in accounting methods in companies. If more data intense methods were to be employed in tax administration, then the businesses themselves would need to have some level of digital maturity to be able to submit data electronically for further review by the tax officials.

A challenge repeated by most of respondents was the issue of corruption. One of the respondents clearly noted that if there was to be any form of human input, there was likely to be corruption as the algorithms cannot feed themselves into the system. This points to a data bias. In addition, another respondent noted the issue of lack of stakeholder/management buy-in. This problem exists because of systemic corruption, as the management who should see the benefits of a system that will breed convenience and facilitate all the benefits highlighted above would not approve it for implementation or frustrate its efforts if they benefit from the inefficiencies of the current system. According to him, "*the challenge we have is implementation in our area, as you try to implement a system, somebody is there trying to see how we can create loophole into it in the same system, so implementation will be a challenge*." Despite this, many respondents stated that the newly elected FIRS chief was (digital) development oriented and getting management buy-in from a leader like him may not prove to be a challenge.

Furthermore, skilled labour in the FIRS for such implementation is lacking. Proper legislation is another factor that would give proper credence to an ideal system. Data management practices by the FIRS was lamented by the tax consultants as a major problem even in the current state, talk more of in a system employing data mining methods. Companies, during tax audits, are sometimes asked to provide documentation that was provided before when being audited for prior tax years and this process is tedious for companies. In this vein, the following measures were proposed as factors necessary to promote the use of ML in tax:

- Proper data management methods
- Adequate infrastructure
- Training of FIRS staff
- Proper legislation backing such implementation.

This section highlighted the factors that may prevent and facilitate the adoption of ML methods in the Nigerian tax administration. The literature review gave general pointers to the fact that there may be elements hindering effective implementation of the proposed method. To discover actual elements which may prevent its implementation, interviews were conducted with professionals who had significant experience in the working system of the FIRS's administration of CIT. In addition, we asked for their ideas on what an ideal system would entail and most of them had a fair notion of what an efficient tax administration represents.

It can be seen from the above that all the subsections give answers to RQ1.1. The responses obtained indicate the kind of data collection and management system in use by the FIRS. From the literature review, particularly the countries highlighted for their stellar data management practices, it can be observed that excellent data management serves as a good foundation for the implementation of the proposed method. We can therefore conclude that scalable data collection and management practices have to be adopted for proper implementation of these methods.

4.3.3 Effect of The Adoption of ML in Tax on The Activities of Relevant Stakeholders

This section aims to answer RQ1.2. "How the use of ML techniques in tax administration will be received by the stakeholders for implementation?" The goal is to see if the application of ML methods in tax will be beneficial to the stakeholders in the company tax administration process. These stakeholders include the FIRS officials, tax consultants and the businesses in question. To analyse this, we had to understand what their current roles are, and what they consider as an improvement in this function due to the adoption of the proposed technique.

The majority of respondents for this study were tax consultants as we had seven (7) of them. The tax consultants had similar job roles while two (2) tax auditors and one (1) tax investigator from the FIRS were interviewed in this study. The tax consultants explained that their duties involve preparation and submission of the tax returns and filing it. They also represent their clients during tax audits and investigation. They were basically meant to ensure that their clients complied with the tax laws and spent the minimum time possible on tax activities while focusing on their main business activity.

The Tax auditors interviewed stated their roles involves reviewing tax returns done by companies and also conducting tax audits in situations where they have reason to believe the documents do not represent a "true and fair view" of the company's financial position. The tax investigator interviewed has his role as an extension of the tax auditor. While he conducts audits, he has greater access to a company's system and does not just rely on what documents are presented to him. Therefore, it is believed that this greater access afforded by the tax investigator can make him come to a more accurate position of a company's financials.

It was imperative at this point to inquire of them what would represent an improvement in their duties in the tax administration process for companies. The tax consultants stated vividly that an improvement for them would be an automated process of tax returns submission. For them, their roles would be greatly improved if the FIRS had access to all the data they required and not request financials of previous years which proves difficult to obtain in most cases and makes their tasks arduous. Finally, an improvement would entail having the FIRS come to a position on the company's financials from a standpoint backed by data, and not just their "hunches". For them, if all their clients (the companies) possessed great data management systems, retrieving source documents when requested by the FIRS would not be a challenge anymore, and they would complete tax audits, where necessary, faster.

On the other hand, the FIRS officials indicated that their tasks would improve if more data intensive methods were employed. On the one hand, they would have better knowledge of the companies under their dispensation, and it would enable them to add more companies to the tax net and thereby increase revenue. On the other hand, their tasks would be easier as they can focus their efforts on select companies highlighted by the ML algorithms and also, back and forth encountered during tax audits would reduce, as they would have more data at their disposal through automation.

Generally, it was evident that the inefficiencies characterised in the current system such as the possibility for prior year audit, the long delays in normal tax audits due to data retrieval and high likelihood of corruption due to the level of human interaction in the current system were frequent occurrences in the current system and this fact has been

highlighted in the literature. It was obvious, as a result, that there was ample room for improvement in these methods.

Overall, it was observed that the respondents were keen on the idea of using Machine Learning and more data -centric methods in tax administration as they observed that using such would shorten the time spent in tax submission and, most importantly, in instances of tax/field audits.

4.3.4 Indicators Measuring the Effectiveness and Efficiency of Adopting ML in Tax Administration in Nigeria

This section will detail the criteria that can be used to measure the effectiveness of adopting ML methods in tax in both quantitative and qualitative measures. This is an important part of this research because it is imperative to know if the adoption of ML in tax would make any positive difference to the current CIT process. RQ 1.3 was dedicated to assessing this impact. The first subsection will give details on this in a quantitative manner using simulations derived from the interview data and online sources. The second section will explain the qualitative indicators explained by the interviewees themselves.

4.3.4.1 Empirical Indicators for Measuring the Advantages of using ML in tax Administration for Companies

This subsection will simulate a scenario, based on interview responses, that show the value to be obtained from the adoption of ML methods in tax administration. This value is comprised of both cost reduction and increase in revenue after the initial IT set up. This section is aimed at answering the question, "What factors (metrics) show that ML in tax would reduce tax leakages and increase financial revenue?"

Seven out of Ten respondents clearly stated that about 90% of all companies in the tax net make it to the 'tax audit' stage while one respondent outrightly stated, "*it's not about **if** a company will get audited by FIRS, it's about **when***." By this, she clearly meant that companies which may not be audited in a certain tax year, may be subject to such audit in subsequent years since they have the authority to audit tax returns of up to 6 years prior. This research will attempt to draw out a cost benefit analysis of adopting this method over the existing system. This is to show undeniably that adopting this method will provide

benefits to all the stakeholders in the tax administration process, both monetarily and otherwise.

Features of the current system:

Three major cost points of the current system are highlighted below:

- Quantitative benefits as regards salaries of people who carry out the audits
- Salaries of officers whose duties comprise seeking new companies to draw into the tax net; and finally,
- Assessing the leakages due to inefficiencies in the current system.

However, the first point alone will be used for this analysis.

Drawing out an average of the field audit cases, as described by the respondents, the following factors can be simulated on average:

- Two FIRS officials conduct the audit per company.
- The audit lasts anywhere between two (2) weeks and three (3) months, depending on the year being audited and data management method of the firm in question. For the purpose of this analysis, we will assume an average audit period of one month. This is inclusive of reconciliation meetings.
- Assuming two officials working on the audit spends about 20 hours each on tax audit per company per month.
- The average tax auditor earns N240,000/month (*FIRS Salaries in Nigeria*, n.d.) and therefore, within an approximate 200-hour month earns NGN1,200/hour.
- There are about 30,000 companies in the tax net and about 90% of these get audited in a given tax year. This equals 27,000 companies used in this analysis.

Therefore, the FIRS expends NGN1,200 x 40hours (= NGN48,000) on one company in a given tax year. Other administrative costs such as transportation and power costs have been excluded to give simplicity to our analysis. We can safely say that a total of N48,000 x 27,000 companies (estimated in the tax net) (= NGN1.296 Billion) is expended yearly on audits.

Features of the proposed system:

Analysis will be done based on costs for initial set-up and costs for continuous maintenance. What is proposed is the establishment of one more large data centre to meet their data needs located at the major administrative building. All the other tax offices can

access data from this location through cloud services. Currently, one data centre is in use (system, 2012) by the FIRS.

In this, the cost of audit will be included, but only 20% of the companies will be included in the yearly costs this time as it is assumed that the ML methods employed will help to streamline the percentage of companies that get audited but may not completely replace audits. Also, 50% of the initial costs will be allowed for maintenance and training in subsequent years. Following this, we have the following set out for the proposed system:

Initial costs		Amount NGN'000
Computers		226,000
Data centre	Hardware	75,000
	Power	40,000
	Networking measures	4,000
	Disaster recovery measures	34,000
	Software	12,000
	Land and building	20,000
Expert consultation and Training		54,000
Software licences		31,000
Audits (20%)		259,200
Miscellaneous Expenses		5,000
		760,200

5 Computers x 226 offices 1 computer = N200,000

Analysis and Conclusion

It is currently not possible to calculate all the money lost to corruption. Also, efforts can be made to further reduce costs in subsequent years. However, a simple payback period for 5 years will be used below to illustrate the economic viability of the proposed system. As it is difficult to estimate (possibly increased) inflows, it has been left out of this analysis. From estimated expenses above, we have the following comparison:

Current system

Year	Amount NGN'000
1	1,296,000
2	1,296,000
3	1,296,000
4	1,296,000
5	1,296,000
Total	6,480,000

Proposed system

Year	Amount NGN'000
1	760,200
2	380,100
3	380,100
4	380,100
5	380,100
Total	2,280,600

Following this, the economic savings that can be generated from the adoption of the proposed system is immediately obvious. It is pertinent to add that the gains are seen despite the fact that increased inflows in the proposed system further leakages in the current system are not accounted for. It can safely be concluded that the adoption of this system should be the priority of the FIRS.

4.3.4.2 Descriptive Indicators for Measuring Effectiveness of ML in Tax Administration in Nigeria.

This subsection seeks to answer research questions, "What are the criteria that establish the effectiveness and efficiency of ML application in tax administration?"
From the interviews, the following indicators were gathered:

- Increase in Tax revenue
- Reduction in leakages due to corruption
- Increase in companies captured in the tax net.
- Increased convenience in tax submission and assessment.
- Greater resource utilization.
- Decrease in the number of court cases.

The greatest indicator for the consultant is the increased convenience it affords measured in the time it takes them to assess and get cleared of tax successfully per given year. If possible, to never see the FIRS officials again and conduct their business at a distance. For the FIRS officials, this metric is increased revenue. This, according to them, is the perfect indicator that can measure the impact of ML in tax, it can be measured often and also reviewed occasionally.
To answer the question, "What would make a 'significant' improvement in the tax administration process?", respondents clearly stated proposed benefits of adopting ML in tax. One clear indicator that sprang up from the respondents is automation of the tax returns process. If this can be achieved, productivity and efficiency will be achieved, and the government revenue will also increase.

4.3.5 Proposed Models to Achieve Tax Objectives for Corporations in Nigeria

A detailed review was conducted in chapter two of this study and the tax objectives highlighted from the results are: Tax policy formulation, Taxpayer behaviour prediction

and tax evasion detection. Also, the models proposed under each of these objectives were given in that review. It can be concluded that RQ1.4. has been covered in this review.

4.3.6 Role of Tax Consultants in Facilitating Tax Compliance/Tax Evasion

There were varying opinions about their role. Some consultants were of the opinion that it is a thing of pride for the consultants if their clients were never flagged for tax audit, so they do all in their power to ensure the company is fully compliant with tax laws. While others stated clearly that tax consultants may sometimes bend to the will of the clients to pay minimum tax and so assist in adjusting the books of accounts to reflect a false financial position.

An interviewee was of the opinion that some consultants do aid tax evasion, "*We have tax consultants that will ask the taxpayer, how much do you want to pay, so that we will prepare your returns based on what you want to pay. So, if you are unlucky, the tax people will come and review your system, and you have to pay more. But if you are lucky, they may not visit you. And if they don't visit you, so be it*." He mentioned this while highlighting the fact that if such incidents of tax evasion occur before the "statute of limitation" period of 6 years, and no tax investigation is executed, such companies may indeed get away with tax evasion.

There was also the speculation that consultants may collude with officials to issue a tax clearance certificate for falsely declared tax where both parties share in the differential. Another opinion was that the influence the tax consultants have on the tax compliance process stems from their vast knowledge of the tax law and, therefore, their ability to represent their clients to comply with these laws as best possible.

Finally, one respondent gave the distinction of large companies and smaller companies in how companies may respond to tax consultant's influence towards tax evasion. He stated that larger corporations tend to use the Big Four to represent them in their tax affairs and may not respond to any consultant's nudge to evade tax because they are bigger corporations and would like to protect their reputation at all costs. For smaller corporations, "*Because when it comes to tax, a client's tax is as good as its consultants. So we have in Nigeria, right being your case study, we have instances whereby all those midsize company, the small company, not the IOC's, not the International Organization, they don't like to approach big four, likes of PwC, KPMG when it comes to managing*

their tax compliance activity, because they know we follow the rule of law, follow the tax law to the letter." "*So and they don't want That bad reputation because that will definitely affect their brand. So they try as much as possible to shy away from that. Companies that may engage in this are mostly indigenous companies that don't have anything to lose*…" (Respondent).

In all of these, the most prevailing point remained viewing the influence of tax professionals on compliance from the aspect of the law, how well it is understood by the tax consultants and how well they can translate it to ensuring full compliance while paying minimum costs. The first case of tax evasion mentioned above stems from the apathetic attitude to tax generally, the second case points to corruption. While the role of the consultants in aiding tax evasion cannot be denied, it can be deterred by proper legislation and structure around the tax system in general and also the size of the company.

4.4 Summary

This chapter gave a detailed explanation of the interview results with a view to answering the research questions. It started with a brief overview of the case and subject selected for this study. Additionally, to give detailed answers to the derived questions of this study, divisions highlighted from the research questions were formed and further divided into subsections to give explanatory answers. Finally, reference was also given to the parts of chapter 2 of this study where the research questions were answered as well.

5. Discussion, Conclusions and Future Work

5.1 Introduction

The preceding chapter clearly shows the advantage of applying ML methods in tax administration in Nigeria. The respondents also agreed that the use of this method would improve the results obtained from their operations and improve their tasks considerably. Despite this, there are some limiting factors to the establishment of this system in Nigeria. This chapter will therefore discuss the major findings and of this research and suggest a framework for the adoption of ML methods in the tax administration of companies in Nigeria. In the next section, the summary of findings will be given where the current research will be tied to current evidence while the following sections will treat the impact, advance limitations, and propose future work.

5.2 Summary of the Findings

Consequent to careful analysis of the interview results and also practices from model countries in the adoption of ML methods in tax administration(Australia and Norway) as given in the literature review of this work, it became evident in this study that the FIRS needs to upgrade its current e-taxation system and also incorporate mechanisms that automate the tax administrative process and reduce human interaction in the system to tackle the problems of corruption and difficulty in tax submissions and audit.

In view of this, the researcher proposes a framework that will guide the implementation of this method in Nigeria taking into consideration possible challenges that may be encountered on the way to implementation.

5.2.1 Proper data management system

The objective of this study involves making profitable use of data to detect tax evasive returns and therefore to reduce human hours involved in this process and reach more accurate predictions. The heart of attaining this objective lies in a proper data management system. It is off this platform that ML methods can be applied and all the advantages to be obtained from this can be achieved.

The following points further explain this point in different aspects:

- **Upgrading the current e-Taxation system to incorporate e-taxation methods:** Applying any form of data mining in tax has to work from an existing mature e-Taxation system. The model cases described in chapter 2 of this study is a good testament of the fact. Both countries listed have a robust e-taxation system from which data mining methods can be applied. This is because of the automation e-Taxation affords. Handling large amounts of data would be difficult and cumbersome if a high level of automation was absent.

  Therefore, to apply this to the Nigerian case, the current e-taxation system would have to be upgraded to allow for the application of ML methods. This can be done by improving the existing servers and data centres to allow for load, especially during peak tax submission periods. Generally, this pertains to having proper consideration for scalability from the building process. Tax offices can also be equipped with proper computers that allow for increased demand of data analysis and management.

- **Increased number of companies in the tax net:** With the rule of assignation of TINs to new companies instituted in July 2020, the FIRS could spring from this platform to maintain proper records of new companies being registered. Also, efforts can be made to add more companies to this "tax net". However, these methods should not be human centric as is obtainable presently. Through Proper data analysis, other companies that deal with the companies in the tax net can be derived and added to the tax net from the backend. Legislation can also support this, but this will be discussed in detail in another section.

- **Accurate data of companies currently in the tax net:** This can be seen from two ways. One way is having proper record of all the information (including source documents) submitted about a company. This way, the ML algorithms can give accurate predictions for the company itself and also in comparative analysis for other companies. Also, companies would not have to go through the trouble of presenting documents for prior years because the FIRS already has that information. The other way involves an integration of the company's accounting system with the FIRS, to enable real time accessibility to data and proper data analysis to determine relevant tax position. However, this method will involve

having a mature e-management system in every company and this is currently not obtainable and may not be for a reasonable period of time.

- **Data cleansing:** A proper methodology has to be applied to the implementation of ML methods in tax in Nigeria. Using accurate data in this process is of utmost importance as anything short of this will defeat the purpose of using this method in the first place. On this note, proper care is to be taken in the data input process at the start and proper measures must be put in place to ensure continuous data quality.

- **Pre-filled tax forms:** With proper company data, pre-filled tax forms can be achieved. The companies with best practices in the use of data mining methods highlighted previously as well as other companies with mature data mining methods in tax use pre-filled tax forms in the submission of tax returns. What this entails has been described in chapter 2 of this study but has arisen here because of the framework being described for its applicability in the Nigerian case. Sending companies pre-filled forms automatically at the end of the company's accounting year would make it easier for these companies to submit their tax returns and therefore comply fully with tax laws. It may also deter evasion as pre-filled forms point to the fact that the FIRS have robust information about a company, and this will discourage attempts at tax evasion.

- **Tax administration at a distance:** The case described in the ATO shows automated communication with taxpayers in situations where their returns do not match what is expected, determined based on similar returns that have been submitted. The FIRS can adopt a similar method of communicating with taxpayers at a distance. They can employ either the K-type nearest neighbour algorithm to compare the tax returns to returns of similar values or perform a trend analysis to compare current returns to that of previous years. Either ways, a substantial deviation from what is expected will be automatically flagged by the system and a notification can be sent to the company to adjust its returns. From these, it is also proposed that it is only where a substantial deviation occurs a second time that an FIRS official can intervene in the situation.

5.2.2 Legislative support

Legal and regulatory backing is needed for the successful implementation of ML in tax administration in Nigeria. This also applies to any other initiative being introduced to a state. The FIRS is a semi-autonomous government body and to a certain extent, can decide on certain initiatives to adopt without further approval. That being said, the FIRS has been making strong strides towards e-Taxation as repeatedly pointed out by the respondents. This move can be supported by the adoption of ML methods in tax. Therefore, legislation is paramount to give legal backing to the implementation of this method. The following factors will require the enactment of modification of existing laws:

- Supporting the change in tax submission: Laws can be made to give legitimacy to the process of data mining in tax administration. This will enforce the method and can be applied once its advantage over manual audits has been established. Following this, FIRS officials may be removed from the manual tax audit system in a phased manner, until the proposed system is established. Punitive methods can then be meted out to officials who flout the new laws and still perform manual audits despite given directives.

- On the part of the companies, a way to bring more companies to the tax net would be to put laws in place that state that businesses should only transact with companies that are already in the tax net. Another aspect would be to enforce the e-accounting systems of these companies. The researcher is of the opinion that the private sector is typically quick to embrace new technologies, and this is even more the case where such change is mandated by the government. The government can make it compulsory for each company to have a certain level of digital maturity and to submit their tax returns entirely online. This legislation can be implemented in a phased approach with more digitally mature states given a stricter deadline and easing off into less digitally mature states with time. This way, the government can "nudge" companies to adopt more digital methods in order to submit their tax returns. Nevertheless, before such legislation can be advanced, there has to be proper infrastructure in place.

5.2.3 Establishment of adequate infrastructure

The respondents clearly stated the issues existing in the current e-Taxation system. Sometimes companies may face service timeouts on the FIRS website in their attempts to submit their tax returns online. Their turning to manual filing methods in such instances defeats the purpose of the automated system put in place and retracts the benefits to be obtained from its adoption. Therefore, proper infrastructure must be established to implement this framework successfully. Although the institution of this infrastructure may be costly to implement at the initial stage, the benefits to be derived will outweigh these costs and should serve as a major motivation for its implementation.

Some components of these were mentioned in the quantitative analysis done in the previous chapter. These components, coupled with some others are outlined below:

- Constant power supply
- High network
- Data centre establishment
- Hardware (Including computers in offices and as used in data centres)
- Software
- Servers

5.2.4 Continuous training

This factor is imperative for the successful implementation of this system. The current FIRS staff may be made to undergo compulsory training to learn the workings of the proposed system and therefore use it in a manner that will facilitate the benefits to be derived. In addition, the FIRS could employ staff with the right background to execute relevant tasks in the proposed system. Overall, it is necessary to carry the current staff along in the design of the system if possible. This will enable them ease into the workflow without much friction. It must be noted that this training must not be stand-alone but must be continuous and scheduled as need arises and as staff turnover is experienced.

Furthermore, each company (company representative) may also be subjected to training on the use of e-taxation methods. This is more important for less tech-savvy companies.

This training can be in the form of broadcast media like television, radio, and digital campaigns or through the organization of closed seminars and workshops where all relevant concepts and workflows are adequately explained. The latter option is more expensive but may be implemented if the perceived benefits to be achieved exceed the costs expended in arranging it.

5.2.5 Partnership and collaboration with experts

The FIRS must consider collaboration with data mining experts in the implementation of this framework especially at the initial stages. These experts may be responsible for the initial set-up and training activities for the FIRS officials. Collaboration with subject matter experts would enable them to avoid major pitfalls in the implementation of this system and help ensure that the execution is successful.

5.3 Discussion

Other research in this area show the relevance of one model over another. They consist of useful comparisons of different methods while highlighting the prevalence of one method over another. These models were derived and tested out using actual tax data (da Silva et al., 2016; Gedde & Sandvik, 2020; González & Velásquez, 2013). Their studies showed the impact of ML in other financial transactions (Whitrow et al., 2009). This study, however, created a framework for its application while using a systematic review to also propose models to be used depending on the tax objective being sought by the tax administration.

5.4 Research Impact

This study began with highlighting the problems and inefficiencies associated with manual review of tax returns and also the challenges encountered in the current e-taxation system. It proceeded to explore the different ML models used in similar situations in other countries and proposed some methods to be adopted based on the case's peculiarity. It explained how advanced ML methods can be employed to solve the highlighted issues and also advanced a practical framework for its implementation. This study gave a general description of an ideal tax administration system and gave pointers to how advanced data analysis and ML can help to attain this. This research therefore presents a novel method of achieving an efficient tax administration system for companies.

In furtherance to this, this study also explored more deeply, the tripartite relationship between tax officials, tax consultants and companies by highlighting the role of the tax consultants in tax compliance.

To conclude, the basic answer to the research question is that the use of data mining tools and machine learning methods will enhance the decision-making capacity of the tax officials to make more accurate decisions on what companies can be selected for further audit. Also, great costs will be reduced as a result of avoidance of tax leakages encountered in the current corrupt system. Therefore, the hypothesis advanced at the beginning of this study has been strengthened by its findings.

5.5 Limitations

In viewing the limitations of this study, one major factor is the constraint in generalizability and lack of rigour (Krusenvik, 2016). As this research employed a single case study, this limitation holds true. Therefore, the issue of bias emanates from this as the researcher's close involvement with the case may reduce the level of objectivity experienced (Yin, 2018). To address these issues, however, the researcher ensured that multiples sources of data were used. Also, the researcher ensured that the research was conducted according to the case study methodology to ensure that the research design remained valid.

One of the major criticisms of fraud detection research based on data mining, as noted by (Sinayobye et al., 2018) is the absence of publicly available real data with which to perform the experiments. A major limitation in this study was obtaining empirical data from the FIRS or any other tax administration in order to test out the models proposed. Upon request, the researcher was informed that the was classified under "tax secrecy" and could not be released to an independent researcher. It can only be used in research from within the respective tax administration body. As the models could not be tested, interviews were conducted to provide deeper insight to the current system and what an ideal system would entail.

The strong aspect of this study stemmed from the detailed overview and analysis of the selected case to provide the given framework. While this is advantageous, it can also serve as a limitation due to limited generalizability to other tax administrations. The

findings obtained however, can be applied to other tax administrations especially the tax models and tax objectives derived from the study.

5.6 Future work

From the findings and limitations of this study given previously, the following points are proposed for future research:

Firstly, the proposed models can be tested using real tax data. This will test the validity of the results obtained from the review and give pointers to what models are best to be adopted in consideration of the tax objectives sought.

Secondly, the reliability of the research method chosen could be tested. In this instance, a different researcher could conduct the same research to compare the results against each other. Furthermore, a different country can be used as a case study, or a multiple case studies could be conducted comparing two or more different countries' tax administrations and producing results which are more 'generalizable'. Again, as regards method, the data collection method can be extended to also employ quantitative data collection and analysis to provide robust insights from diversified responses.

Furthermore, qualitative methods have a "good chance of bias" and may be difficult to compare (Lips, 2021) and this study is not exempt from this weakness. However, this study gives proper context to the selected case to enable other researchers have proper "protocol" in further research (Runeson et al., 2012) as a means to forestall case study criticism in this research.

In addition, Dal Pozzolo et al. (2014) highlights the factors of incremental learning and unbalanced data in data mining efforts to combat fraud. For the element of unbalanced data, the importance of having an automatic data capture process is also emphasized in their study. This consideration is important because the number of fraudulent transactions may be (and is more often the case) smaller than the genuine transaction count. For incremental learning, they indicate that a static learning detection model is typically relearned at periodic intervals (e.g. per month/year) as against online learning models which are characterized by instant updating of new data as soon as they arrive (Dal Pozzolo et al., 2014). This study focuses on stationary model setting. However, future

studies could explore real time updating of tax information and its treatment considering the unbalanced data factor.

Again, other studies highlight the relevance of blockchain technology in combating tax evasion. Comparative studies can be done on this basis, using frameworks, to determine what option is more suitable for a particular administration.

Finally, a future work could analyse the impact of this system more concretely and give perspectives from the angle of the companies (or tax entity) in question. This would be useful in determining if the output of this thesis was achieved to start with.

References

Allingham, M. G., & Sandmo, A. (1972). Income tax evasion: A theoretical analysis. *Journal of public economics*, *1*(3-4), 323-338.

Azogu, I. N. (2018). *A Framework For The Adoption Of Blockchain Technology In Healthcare Information Management System: A Case Study Of Nigeria* [Master's Thesis]. Taltech.

Battaglini, M., Guiso, L., Lacava, C., & Patacchini, E. (2019). The true role of tax professionals. *VoxEU.Org*. https://voxeu.org/article/true-role-tax-professionals

Battiston, P., Gamba, S., & Santoro, A. (2020). Optimizing Tax Administration Policies with Machine Learning. *University of Milan Bicocca Department of Economics, Management and Statistics Working Paper*, *436*.

Bird, R. M., Martinez-Vazquez, J., & Alm, J. (Eds.). (2003). Public finance in developing and transitional countries*: Essays in honor of Richard Bird*. Edward Elgar.

Brownlee, J. (2019, August 11). A Tour of Machine Learning Algorithms. *Machine Learning Mastery*. https://machinelearningmastery.com/a-tour-of-machine-learning-algorithms/

Centre for Public Impact. (2019). *Artificial intelligence in taxation A case study on the use of AI in government*. https://www.centreforpublicimpact.org/assets/documents/ai-case-study-taxation.pdf

Chen, K.-P., & Chu, C. Y. C. (2005). Internal Control versus External Manipulation: A Model of Corporate Income Tax Evasion. *The RAND Journal of Economics*, *36*(1), 151–164.

da Silva, L. S., Rigitano, H., Carvalho, R. N., & Souza, J. C. F. (2016). Bayesian Networks on Income Tax Audit Selection-A Case Study of Brazilian Tax Administration. *BMA@ UAI*, 14–20.

Dal Pozzolo, A., Caelen, O., Le Borgne, Y.-A., Waterschoot, S., & Bontempi, G. (2014). Learned lessons in credit card fraud detection from a practitioner perspective. *Expert Systems with Applications*, *41*(10), 4915–4928. https://doi.org/10.1016/j.eswa.2014.02.026

Davia, H. R., Coggins, P. C., Wideman, J. C., & Kastantin, J. T. (2000). *Accountant's guide to fraud detection and control* [PhD Thesis]. Univerza v Mariboru, Ekonomsko-poslovna fakulteta.

David, F., & Abreu, R. (2013). Tax Evasion. In S. O. Idowu, N. Capaldi, L. Zu, & A. D. Gupta (Eds.), *Encyclopedia of Corporate Social Responsibility* (pp. 2497–2503). Springer. https://doi.org/10.1007/978-3-642-28036-8_302

de Roux, D., Perez, B., Moreno, A., Villamil, M. del P., & Figueroa, C. (2018). Tax Fraud Detection for Under-Reporting Declarations Using an Unsupervised Machine Learning Approach. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 215–222. https://doi.org/10.1145/3219819.3219878

Demush, R. (n.d.). Fraud Detection Machine Learning Solutions. *Solving Financial Fraud Detection with Machine Learning Methods*. Retrieved February 5, 2021, from https://perfectial.com/blog/fraud-detection-machine-learning/

Dias, A., Pinto, C., Batista, J., & Neves, E. (2016). Signaling Tax Evasion, Financial Ratios and Cluster Analysis. *BIS Quarterly Review*.

Dutta, I., Dutta, S., & Raahemi, B. (2017). Detecting financial restatements using data mining techniques. *Expert Systems with Applications*, *90*, 374–393. https://doi.org/10.1016/j.eswa.2017.08.030

Feinstein, J. S. (1991). An Econometric Analysis of Income Tax Evasion and Its Detection. *The RAND Journal of Economics*, *22*(1), 14–35. https://doi.org/10.2307/2601005

FIRS. (2018). https://www.firs.gov.ng/SiteApplication/Home/Home.aspx

FIRS Salaries in Nigeria. (n.d.). Glassdoor. Retrieved April 25, 2021, from https://www.glassdoor.com/Salary/Federal-Inland-Revenue-Service-Nigeria-Salaries-EI_IE702895.0,30_IL.31,38_IN177.htm

Ge, Z., Song, Z., Ding, S. X., & Huang, B. (2017). Data Mining and Analytics in the Process Industry: The Role of Machine Learning. *IEEE Access*, *5*, 20590–20616. https://doi.org/10.1109/ACCESS.2017.2756872

Gedde, N., & Sandvik, I.-S. (2020). *Unsupervised machine learning on tax returns: Investigating unsupervised and semisupervised machine learning methods to uncover anomalous faulty tax returns* [Master's Thesis] Norwegian School of Economics.

González, P. C., & Velásquez, J. D. (2013). Characterization and detection of taxpayers with false invoices using data mining techniques. *Expert Systems with Applications*, *40*(5), 1427–1436. https://doi.org/10.1016/j.eswa.2012.08.051

Gustafsson, J. (2017). *Single case studies vs. Multiple case studies: A comparative study*. 15.

Hand, D. J., Mannila, H., & Smyth, P. (2001). *Principles of data mining*. MIT Press.

Hauptman, L., Horvat, M., & Korez-Vide, R. (2014). Improving Tax Administration's Services as a Factor of Tax Compilance: The Case of Tax Audit. *Lex Localis - Journal of Local Self-Government*, *12*. https://doi.org/10.4335/12.3.481-501(2014)

Ho, T. K. (1995). Random decision forests. *Proceedings of 3rd International Conference on Document Analysis and Recognition*, *1*, 278–282 vol.1. https://doi.org/10.1109/ICDAR.1995.598994

IBM, C. E. (2020, July 1). *What is Artificial Intelligence (AI)?* IBM. https://www.ibm.com/cloud/learn/what-is-artificial-intelligence

ICAEW. (2019, Edition). *Digitalisation of tax: International perspectives*. https://www.icaew.com/-/media/corporate/files/technical/technology/thought-leadership/digital-tax.ashx?la=en

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, *31*(8), 651–666. https://doi.org/10.1016/j.patrec.2009.09.011

Jun Lee, S., & Siau, K. (2001). A review of data mining techniques. *Industrial Management & Data Systems*, *101*(1), 41–46. https://doi.org/10.1108/02635570110365989

Kavlakoglu, E. (2020, September 1). *AI vs. Machine Learning vs. Deep Learning vs. Neural Networks: What's the Difference?* IBM. https://www.ibm.com/cloud/blog/ai-vs-machine-learning-vs-deep-learning-vs-neural-networks

khlif, H., & Achek, I. (2015). The determinants of tax evasion: A literature review. *International Journal of Law and Management*, *57*(5), 486–497. https://doi.org/10.1108/IJLMA-03-2014-0027

Kitchenham, B., & Charters, S. (2007). *Guidelines for performing systematic literature reviews in software engineering*.

Krusenvik, L. (2016). *Using Case Studies as a Scientific Method: Advantages and Disadvantages*. http://urn.kb.se/resolve?urn=urn:nbn:se:hh:diva-32625

Lips, S. (2021, December 21). *Master Thesis Workshop I*.

Louridas, P., & Ebert, C. (2016). Machine Learning. *IEEE Software*, *33*(5), 110–115. https://doi.org/10.1109/MS.2016.114

Martikainen, J. (2012). *Data mining in tax administration-using analytics to enhance tax compliance*. [Master's Thesis]. Aalto University.

Matos, T., de Macedo, J. A. F., & Monteiro, J. M. (2015). An Empirical Method for Discovering Tax Fraudsters: A Real Case Study of Brazilian Fiscal Evasion. *Proceedings of the 19th International Database Engineering & Applications Symposium*, 41–48. https://doi.org/10.1145/2790755.2790759

Mehta, P., Mathews, J., Kumar, S., Suryamukhi, K., Babu, Ch. S., Rao, S. V. K. V., Shivapujimath, V., & Bisht, D. (2019). Big Data Analytics for Tax Administration. In A. Kő, E. Francesconi, G. Anderst-Kotsis, A. M. Tjoa, & I. Khalil (Eds.), *Electronic Government and the Information Systems Perspective* (pp. 47–57). Springer International Publishing. https://doi.org/10.1007/978-3-030-27523-5_4

Mendez, V. I. V. (2020, June 15). *Strengthening the Toolkit for Tax Compliance Management: Machine learning 1*. https://www.ciat.org/fortaleciendo-el-maletin-de-herramientas-para-la-gestion-del-cumplimiento-tributario-machine-learning-1/?lang=en

Momoh, Z. (2018). Federal Inland Revenue Service (FIRS) and Tax Compliance in Nigeria: Challenges and Prospects. *International Journal*, *1*(4), 5.

Nethone. (2019, March 21). *A Beginner's Guide to Machine Learning in Payment Fraud Detection Prevention*. Medium. https://medium.com/@Nethone_/a-beginners-guide-to-machine-learning-in-payment-fraud-detection-prevention-360c95a9ca54

Nigeria loses $15 billion annually to tax evasion, says FIRS. (2019, October 24). *The Guardian Nigeria News - Nigeria and World News*. https://guardian.ng/business-services/nigeria-loses-15-billion-annually-to-tax-evasion-says-firs/

Nock, M. K., Michel, B. D., & Photos, V. I. (2007). Single-case research designs. *Handbook of Research Methods in Abnormal and Clinical Psychology*, 337–350.

OECD. (2004). *Compliance Risk Management: Managing and Improving Tax Compliance*. https://www.oecd.org/tax/administration/33818656.pdf

OECD - Forum on Tax Administration. (2006, March). *Using Third Party Information Reports to Assist Taxpayers Meet their Return Filing Obligations—Country Experiences With the Use of Pre-populated Personal Tax Returns*. https://www.oecd.org/tax/administration/36280368.pdf

Office, A. T. (n.d.). *How we use data and analytics*. Australian Taxation Office. Retrieved December 12, 2020, from https://www.ato.gov.au/about-ato/managing-the-tax-and-super-system/insight--building-trust-and-confidence/how-we-use-data-and-analytics/?default

Olaghere, E., & Adewuyi, O. (2020, July 29). *INSIGHT: Improving Electronic Tax Administration in Nigeria*. https://news.bloombergtax.com/daily-tax-report-international/insight-improving-electronic-tax-administration-in-nigeria

Richards, N., & Ekhator, E. (2019). *Electronic taxation in Nigeria: Challenges and prospects*. *30*, 47–64.

Rudin, C., & Wagstaff, K. (2013). *Machine learning for science and society*.

Runeson, P., Host, M., Rainer, A., & Regnell, B. (2012). *Case study research in software engineering: Guidelines and examples*. John Wiley & Sons.

Sandmo, A. (2005). The Theory of Tax Evasion: A Retrospective View. *National Tax Journal*, *58*(4), 643–663.

Schneider, F., Raczkowski, K., & Mróz, B. (2015). Shadow economy and tax evasion in the EU. *Journal of Money Laundering Control*, *18*(1), 34–51. https://doi.org/10.1108/JMLC-09-2014-0027

Schreyer, M., Sattarov, T., Borth, D., Dengel, A., & Reimer, B. (2017). Detection of anomalies in large scale accounting data using deep autoencoder networks. *ArXiv Preprint ArXiv:1709.05254*.

Sinayobye, J. O., Kiwanuka, F., & Kyanda, S. K. (2018). A State-of-the-Art Review of Machine Learning Techniques for Fraud Detection Research. *2018 IEEE/ACM Symposium on Software Engineering in Africa (SEiA)*, 11–19.

Soviany, C. (2018). The benefits of using artificial intelligence in payment fraud detection: A case study. *Journal of Payments Strategy & Systems*, *12*(2), 102–110.

Starman, A. B. (2013). The case study as a type of qualitative research. *Journal of Contemporary Educational Studies/Sodobna Pedagogika*, *64*(1).

System, W. (2012, July 23). Federal Inland Revenue Service(FIRS) Network Infrastructure Project. *WECO Systems International Limited*. https://wecosysgroup.com/federal-inland-revenue-servicefirs-network-infrastructure-project/

Taxes in Estonia. (2019, May). Work in Estonia. https://www.workinestonia.com/working-in-estonia/taxes/

Umenweke, M. N., & Ifediora, E. S. (2016). The Law and Practice of Electronic Taxation in Nigeria: The Gains and Challenges. *Nnamdi Azikiwe University Journal of International Law and Jurisprudence*, *7*, 101–112.

What is Clustering? | Clustering in Machine Learning. (n.d.). Google Developers. Retrieved March 7, 2021, from https://developers.google.com/machine-learning/clustering/overview

Whitrow, C., Hand, D., Juszczak, P., Weston, D., & Adams, N. (2009). Transaction aggregation as a strategy for credit card fraud detection. *Data Mining and Knowledge Discovery*, *18*, 30–55. https://doi.org/10.1007/s10618-008-0116-z

Wu, R.-S., Ou, C. S., Lin, H., Chang, S.-I., & Yen, D. C. (2012). Using data mining technique to enhance tax evasion detection performance. *Expert Systems with Applications*, *39*(10), 8769–8777. https://doi.org/10.1016/j.eswa.2012.01.204

Yin, R. K. (2018). *Case study research and applications: Design and methods (6th ed.)*. Sage publications.

Zainal, Z. (2007). Case Study As a Research Method. *Jurnal Kemanusiaan*, *5*(1), Article 1. https://jurnalkemanusiaan.utm.my/index.php/kemanusiaan/article/view/165

Zhang, K., Li, A., & Song, B. (2009). Fraud Detection in Tax Declaration Using Ensemble ISGNN. *2009 WRI World Congress on Computer Science and Information Engineering*, *4*, 237–240. https://doi.org/10.1109/CSIE.2009.73

Zhang, X.-D. (2020). Machine Learning. In X.-D. Zhang (Ed.), *A Matrix Algebra Approach to Artificial Intelligence* (pp. 223–440). Springer. https://doi.org/10.1007/978-981-15-2770-8_6

Appendix 1 – Plain licence for allowing the thesis to be available and reproducible for the public

I, Bridget Kifordu (23.07.1994)

1. Allow the Tallinn University of Technology without any charges (Plain licence) my work: Machine Learning in Tax Administration – A case study of Nigeria,

supervised by Prof. Ben Sadok,

1.1. to be reproduced for the purpose of conservation and electronic publication, including the digital repository of the Tallinn University of Technology, until the end of copyrighted time limit;
1.2. to be available to the public through the Tallinn University of Technology online environment, including the digital repository of the Tallinn University of Technology, until the end of the copyrighted time limit.

2. I am aware, that all rights, named in section 1, will remain to the author.

3. I confirm that by allowing the use of the Plain licence, no intellectual rights of third parties will be violated as set in the personal data protection act and other legislation.

BRIDGET CHINYE KIFORDU (*Signature*)

04.05.2021 (*Date*)

Appendix 2 - Interview Questions (Guide)

Step 1: Introduction and dialogue on the interviewee's role and experience

- What is your role?
- How long have you worked in your current position?
- Can you describe your responsibilities in operational terms?
- What other person do you need to cooperate most with while executing your duties?
- From what part of Nigeria do you majorly execute your duties? Do you have clients in other states?

Step 2: Goals of the research

- Can you briefly describe how tax audits are done for companies in each tax year?
- Do you know anything about Data mining in tax? If so, what do you know?
- What is your opinion on moving from the current form of tax audit to using data mining and Machine Learning models in tax administration?
- Do you think using these models will affect your duties in tax returns preparation? YES/NO (If YES, please explain how. If NO, please give reasons).

Step 3: Questions about digitized tax system and tax audits

- How would you rate the e-tax system rolled out by the FIRS? Is it fully functional? Do you find it useful?
- In your experience, what is the frequency of audit visits? How efficient is the audit system?
- What is your experience in having FIRS officials conduct audits? Is information captured digitally?
- What do you see might be the challenges to having such systems (digitized tax systems) in place? Can you perhaps give examples if there are and what do you think should be done to address these challenges? Ethical concerns?

Step 4: Questions on the influence of tax professionals in tax compliance

- Would you agree that the tax consultants play a great role in tax compliance? If so, why. If not, why?

- In your experience, has any of the companies you represent been selected for tax audit by the tax officials? If so, does it affect the acceptance/treatment of subsequent tax returns by the tax officials?
- Do you think the tax consultants can serve (as information hubs) to ensure proper adherence to tax policies?

Step 5: Questions about the tax system in Nigeria as a whole

- What do you see as an ideal CIT tax process?
- What features need to be present to achieve this?
- Are the companies, in your opinion, satisfied with the current system in use? (YES/NO- give reasons)
- To your knowledge, are there steps being taken to address these challenges in the current system?
- How will this new method be received by stakeholders? How will it affect your job?

Step 6: Questions on the measuring criteria for effectiveness and efficiency in the system.

- Do you have metrics for measuring the success in using the current CIT tax system?
- Do you have metrics for measuring the productivity in using the proposed (ML) system?
- What do you think are the expected benefits of adopting adata mining in Tax administration?

Appendix 3 – Link to the interview audio recordings

Please click here or click or on the link provided below:

https://drive.google.com/drive/folders/1uSkLAKJqLA7708Og98otDGiqZm7P-LuW?usp=sharing

Appendix 4 – Thematic Map of Categories and Codes