Journal of Forensic Sciences And Criminal Investigation
ISSN: 2476-1311

Juniper Publishers key to the Researchers

Research Article
Volume 18 Issue 2 - January 2024
DOI: 10.19080/JFSCI.2024.18.555981

J Forensic Sci & Criminal Inves

Artificial Intelligence System for Value Added Tax Collection via Self Organizing Map (SOM)

Felix Bankole[1]* and Zama Vara[2]

University of South Africa, South Africa

HarpDome Technologies Inc, Canada

Submission: January 02, 2024; **Published:** January 19, 2024

***Corresponding author:** Felix Bankole, Professor, University of South Africa South Africa, HarpDome Technologies Inc, Canada

Abstract

Findings: Based on our experiments, our approach is an effective instrument for hot spots and heat map exploration since it employs visualizations techniques that are easy to understand. In the ANN-SOM similarity Heat map we observe that VAT vendors or entities, with similar VAT return characteristics, are grouped in the same area or node. Generally, in business, users are more interested in "abnormal clusters" or hot spots. That is, clusters of VAT vendors who have more suspicious behavior than normal nodes or clusters. However, when interpreting the ANN-SOM Heat map the abnormal clusters are those that have a smaller number of entities. That is, these nodes are composed of suspicious VAT vendors. Such VAT vendors require detailed human verification by VAT audit specialists. The results show that detection of suspicions VAT declarations is a very challenging task as VAT declarations datasets are extremely unbalanced in nature. Furthermore, the tax fraud domain is full of unlabeled data, which in turn makes it difficult to use supervised learning approaches. VAT fraud or suspicious behavior can be differentiated by observing VAT return form attributes such as VAT Liability, Exempt supplies, Diesel Refund and Input VAT on Capital Goods purchased.

Research, Practical & Social implications: The article highlights the use of SOMs in exploring hot spots in a large real-world data set from the taxation domain. The approach is an effective tool for hot spots exploration since it offers visualizations that are easy to understand for tax administration users. Tax auditors can select abnormal clusters for further investigation and exploration. The framework and method are designed with the objective of assisting with the VAT audit case selection. Furthermore, we envisaged that the model would improve the effectiveness and efficiency of the revenue collection agencies in identifying anomalies on VAT returns filed by the taxpayers. Moreover, Tax authorities may be able to select the most appropriate unsupervised learning technique from this work having considered other alternatives, their operational requirements and business context. Thus, leading to a multitude of available Artificial Intelligence aided VAT fraud detection algorithms and approaches. Additionally, the techniques proposed in this paper will help tax administrations with precise case selection using an empirical and data-driven approach, which does not depend upon labelled historic VAT datasets. Furthermore, we envisage the approach will result in a high hit ratio on suspicious VAT returns, and thus improve tax compliance due to the likelihood of detection.

Originality/value: The value of the study is that in as much as this paper's focal point is on VAT fraud detection, we are confident that the present model may just as well be applicable to other tax types, like Company Income Tax and Personal Income Tax for instance. This research outcome shows the potential of artificial intelligence techniques in the realm of VAT fraud and criminal investigation. Furthermore, this review put forward high-level and detailed classification frameworks on VAT fraud detection. Additionally, the framework proposed herein presents tax auditors with a systemic case selection guide of suspicious VAT returns. Furthermore, it is crucial to have an all-encompassing view on detecting tax fraud in general and VAT fraud. This is to broaden the understanding and knowledge of the VAT fraud phenomenon among researchers.

Keywords: Self-organizing map; Cluster analysis, Anomaly detection; VAT fraud detection; Artificial intelligence; Robotics process automation; Algorithms and Machine learning; Criminal investigation

Introduction

In the Information Systems field, IS or IT business strategies and modelling could be referred to as an act or science of initiating a transaction or exchange through a predetermined series of actions (e.g., organizational management, planning, or technology processes). Research on digital platforms (or multisided markets) originated in IS economics and has been pronounced in the

strategy field since early 2000s [1,2]. The adoption of Internet and mobile

phone services have enabled industries to introduce a platform enabling business models commonly referred to as disrupting industry structures [3,4]. For instance, in transportation, lodging and meal delivery sectors (such as Uber, Airbnb and Mr. Delivery).

AI digital platform (e-platform) is an ICT value creation which facilitates transactions between several groups of users including buyers and sellers [2]. For example, content and search engine optimization, social media marketing and optimization are augmenting consumer buying powers as more and more consumers are voicing their opinions. Furthermore, customers express their views about the industry, their brands and related product attributes. Ineffective delivery of products and services regarding customer requirements could impact the corporate brand, image, loyalty, and values. This could lead to customer discontent thereby causing disengagement with products and brands through eWom [5]. Overall, digital technologies present considerable opportunities for enterprise leaders to rethink their business to create better experience for customers, employees, partners, and as well lower cost of services [6].

In this research, we explore how an Artificial Intelligence digital platform framework could be employed to explore value added tax fraud prediction in the revenue service sector. Notably, AI digital platforms appear to influence effective and productive revenue collection strategic decisions. The recent work and trends in the field of AI digital platforms varies namely, Apple Siri enabled smart mobile searches, the web search and capture of keywords, google duplex for hair-grooming appointments, restaurant reservations, voice tone and language patterns duplex are hardly distinguishable with human voice [7]. Recently, Amazon entered a partnership with Marriott International Inc. wherein Amazon Flywheel and Amazon Alexa Voice enabled platforms perform the task of assisting hotel guests from room servicing to housekeeping [8].

In spite of these developments, we suggest that very little research has existed around the use of AI in computer information systems that explores digital platforms designed to aid the efforts of revenue collection and the identification of tax fraud and evasion.

VAT fraud and as well as VAT criminal investigation can be explained as a deliberate misrepresentation of information in VAT returns or declarations to decrease the amount of the tax liability [9]. VAT fraud is a major problem for tax administrations across the world. It is carried out by criminals and organized crime networks. VAT fraud can occur in many sectors including electronics, minerals, cars, and carbon permits. The most attractive goods for fraudsters have been those of high value and low volume such as mobile phones or computer chips, which generate huge amounts of VAT in the lowest number of transactions and in the shortest possible time [10]. At the heart of the VAT system is the credit mechanism, with tax charged by a seller available as a credit against their liability on their own sales and, if more than the output VAT due, refunded to them. According to Keen & Smith [11], this creates opportunities for several types of fraud characteristic of the VAT namely: False Claims for Credit or Refund; Zero-rating of Exports and Misclassification of Commodities; Credit Claimed for VAT on Purchases that are not Creditable; Bogus Traders; Under-reported Sales; Failure to Register; Tax Collected but not Remitted.

The development of AI digital platform (e-platform) for VAT fraud detection is required to ensure that large amounts of revenue that could be used by the government for the much-needed socio-economic public services such as hospitals, schools and road infrastructure are generated. Artificial neural networks (ANN), when trained properly can work like a human brain. They learn by example, like people and are known to be exceptionally good classifiers. Furthermore, the neural network is preferred in this study due to its ability to solve classification problems [12]. Machine Learning algorithms are very likely to produce faulty classifiers when they are trained with imbalanced datasets. Fraud datasets are characterized by imbalanced datasets. An imbalanced dataset is one where the number of observations belonging to one class is significantly higher than those belonging to the other classes. Other algorithms tend to show a bias for the majority class, treating the minority class as a noise in the dataset. In many standard classifier algorithms, such as Naive Bayes, Logistic Regression, and Decision Trees, there is a likelihood of the wrong classification of the minority class. ANN are well suited to imbalanced datasets [13]. Hence, this research proposes a Self-Organizing Map (SOM) Neural Network algorithm to detect VAT fraud.

A self-organizing map (SOM) or self-organizing feature map (SOFM) is a type of artificial neural network (ANN). The SOM is trained using unsupervised learning to produce a low dimensional map. It is a discretized representation of the input space of the training samples, called a map. Self-organizing maps differ from other AI algorithms in that they use competitive learning instead of error-correction learning [14]. In a sense, it uses a neighborhood function to preserve the topological properties of the input space [15].

Background

The background for this research is multifold, that is, to create a VAT fraud detection AI framework as well as the application of a Self- Organizing Map (SOM) algorithm to detect VAT fraud. This paper focuses on VAT fraud detection. The fraud detection arena is characterized by extremely large amounts of unlabeled structured and unstructured data. Unsupervised machine learning algorithms

How to cite this article: Felix B, Zama V. Artificial Intelligence System for Value Added Tax Collection via Self Organizing Map (SOM). J Forensic Sci & Criminal Inves. 2024; 18(2): 555981. DOI: 10.19080/JFSCI.2024.18.555981

are well suited to unlabeled datasets. Hence, we herein propose an unsupervised machine learning approach for detecting VAT fraud. We begin by describing the SOM algorithm. Thereafter, we discuss data collection and data preparation [9]. Additionally, we elaborate on issues relating to VAT variables feature selection, followed by an exploratory data analysis. Furthermore, we explain the machine learning technique and statistical algorithm employed in this study. Finally, we present results from actual digital platform-based experiments conducted on taxpayer level VAT dataset.

Justification

The traditional way for improving tax fraud detection by tax audit is costly and limited in terms of scope given the vast population of taxpayers and considering the limited capacity of tax auditors. Auditing tax returns is a slow and costly process that is very prone to errors. Conducting tax audits for example, involves costs to the tax administration as well as to the taxpayer. Furthermore, the field of anomaly and fraud detection is characterized by unlabeled historical data. To this end, the writers suggest the use of unsupervised machine learning (ML) algorithms which are well suited to unlabeled datasets. There is little research comparing the effectiveness of various unsupervised learning approaches in the VAT fraud realm. The detection of tax fraud can be constructively approached by techniques based on supervised ML techniques. However, those methods require enormous training datasets containing data instances corresponding to both verified tax fraud cases and compliant taxpayers. Many previous studies reviewed allude to the scarcity of labelled datasets in both developed and developing countries. The second problem with supervised ML approaches could be that only a small number of frauds are identified by tax administrations that are recorded in the training dataset. Thus, recorded fraud cases are not representative of the entire population. Therefore, a trained supervised models will be biased; however, it will have a high fraud hit ratio, but a low recall.

Consequently, the lack of the availability of labeled tax fraud is usually dealt with by unsupervised ML methods based on anomaly detection algorithms. Therefore, unsupervised methods are suitable as a decision support or selection tool in tax fraud systems. Therefore, unsupervised algorithms enable better and faster prioritization of tax audit cases, thus improving the effectiveness and efficiency of tax collection. Secondly, tax fraud cases based on accurate unsupervised learning may lead to a more efficient use of resources.

In the study various approaches are considered including Principal Component Analysis (PCA), k-Nearest Neighbors (kNN), Self-Organizing Map (SOM) and K-means, as well as deep learning methods including Convolutional Neural Networks (CNN) and Stacked Sparse AutoEncoder (SSAE). This paper can serve as a guideline to provide useful clues for analysts who are going to select ML methods for tax fraud detection systems as well as for researchers interested in developing more reliable and efficient methods for fraud detection. In this study, the VAT datasets obtained are from the South African tax administration. In particular, a dataset of the mining industry is chosen. This is because the South African diesel rebate scheme is very prone to abuse and VAT fraud. Additionally, in South Africa the mining sector is very important. It employs more than 464,000 people and accounts for 8.2% of GDP [16].

Objective of the Work

The objective of the work is to determine what type of AI technique or framework could be applied to improve tax collection. In particular, the present study explores the use of AI in VAT fraud detection. The main purpose of the study is to determine how corporate VAT fraud could be detected in real time. Corporates and private businesses, primarily use artificial intelligence to influence business models, sales processes, customer segmentation, strategy formulation as well as to understand customer behaviour, in order to increase revenue [4]. There is substantial research on the influence of AI on business strategies with the objective of increasing revenue [2]. However, there is limited research on the use of AI in information systems research to assist in the efforts of revenue collection and VAT fraud detection.

Detection of suspicions VAT declarations is a very challenging task as VAT declarations datasets are extremely unbalanced in nature. Furthermore, the tax fraud domain is full of unlabeled data, which in turn makes it difficult to use supervised learning approaches. In this research paper, we proposed an unsupervised learning approach. Regardless, it is crucial to have an all-encompassing review on detecting tax fraud in general.

Unsupervised algorithms are well suited to unlabeled historical datasets, common in the fraud detection or classification arena. The authors conduct experiments using an unsupervised Neural Network algorithm to classify suspicious Value Added Tax declarations. This algorithm can assist in the efforts of tax audits made by tax administrations. Consequently, it is envisaged that the chances of detecting fraudulent VAT declarations will be enhanced using AI techniques, proposed in this paper.

Literature Review

In the age of big data, detecting fraudulent activities within tax returns is analogous to finding a needle in a haystack. Anomaly detection approaches and ML techniques that focus on interdependencies between different data attributes, have been increasingly used to analyze relations and connectivity patterns in tax returns to identify unusual patterns [17]. In the surmise of Molsa [18] Artificial Intelligence & automation are poised to reshape the digital platform function. Phua, Alahakoon & Lee [19] in their paper, tabulate, compare and summarize fraud detection methods and techniques that have been published in

How to cite this article: Felix B, Zama V. Artificial Intelligence System for Value Added Tax Collection via Self Organizing Map (SOM). J Forensic Sci & Criminal Inves. 2024; 18(2): 555981. DOI: 10.19080/JFSCI.2024.18.555981

academic and industrial research during the past 10 years. This is done in the business context of harvesting the data to achieve higher cost savings. Phua et al. [19] presents a methodology and techniques used for fraud detection together with their inherent problems. In their research, they juxtaposed four major methods commonly used for applying a machine learning algorithm. Firstly, supervised learning on labelled data. Secondly, hybrid approach with labeled data. Thirdly, semi-supervised approach with non-fraud data, and lastly, unsupervised approach with un-labeled data. Meanwhile, Shao et al. [20] describe the building of a fraud detection model for the Qingdao customs port of China. The model is used to provide decision rules to the Chinese custom officials for inspection of goods based on historical transaction data. The objective is to improve the hit rate. The model is appropriately named 'Intelligent Eyes' and has been successfully implemented with high predictive accuracy [20].

Tax administration agencies must use their limited resources very judiciously whilst achieving maximal taxpayer compliance albeit at the lowest cost of revenue collection. Whilst, at the same time, adhering to lower levels of taxpayer intrusion. The Quantitative Analytics Unit of the Securities Regulation Institute in Coronado, California, USA, developed a revolutionary new statistical-based algorithm application called "NEAT," which stands for the "National Examination Analytics Tool" [21]. With NEAT, securities examiners can access and systematically analyze massive amounts of trading data from firms in a fraction of the time it has in the previous years. In one recent examination, NEAT was used to scrutinize in 36 hours exactly 17 million transactions executed by one investment adviser. Among its many benefits, NEAT can search for evidence of probable insider trading by comparing historical data of significant corporate activity like mergers and acquisitions against the companies in which a registrant is trading. They then use this information to analyze how the registrant traded at the time of those notable events. NEAT can review all the securities the registrant traded and quickly identify the trading patterns of the registrant for suspicious activities [21].

Theoretical Background

Artificial intelligence (AI) digital platform

An AI e-platform uses artificial intelligence techniques to make automated decisions based on data collection, data analysis and data scrutiny. An AI digital platform serves as a Computer Information Systems platform that showcases economic trends that may impact system automation efforts. AI techniques like ML use customer data, to learn how to best interact with customers thereby providing insights that could serve those customers with tailored messages at the right time without intervention from external factors to guarantee effective, efficient, and impactful product development and communication. In the current circumstances, this study endeavored to place AI digital platform development in the context of systemic developments which could be well-thought-out as digitalization of industries IT resources [22].

Additionally, an AI e-platform performs repetitive, routine, and tactical tasks that require less human intervention. It use-cases may include data analysis; media buying; automated decision making; natural language processing; generation of content; real-time personalized or tailored messaging [22]. Accordingly, AI digital platforms hold a vital role in helping managers to understand ML algorithms like k-nearest neighbor, Bayesian Learning and Forgetting, Self- Organizing Maps, Artificial Neural Network Self-Organizing Maps. These forms of algorithms help to gain a comprehensible understanding of how amenable and responsive a customer is to a specific product offering effort. Therefore, AI e-platform frameworks are required to process expansive and extensive data sets that can potentially unveil hidden knowledge and insights about products and their customers. Thus, enabling organizations to derive significant revenue growth, whilst strengthening customer relationships [23]. There is significant research on the impact of AI on business processes in information systems to increase revenue, however more research is needed to explore its potential in aiding revenue collection by tax administrations [24]. Consequently, in this current study we use unsupervised models to detect Value Added Tax fraud in order to improve tax compliance, and thus enhance revenue collection.

Social behaviours on tax fraud and compliance

Earlier on, we explained that an AI e-platform is required to understand customer needs to create appropriate and personalized messages and product offering. In the same vein, a tax administration's understanding of its taxpayers is key to effective tax administration and revenue collection. The taxpayers' attitude on compliance may be influenced by many factors, which eventually influence a taxpayer's behavior. Those factors which influence tax compliance behavior differ from one country to another, and from one individual to another [25]. Namely, taxpayer's perceptions of the tax system and Tax Authority [26]; peer attitude, norms and values; a taxpayers 's understanding of the tax system or tax laws [27]; motivation such as rewards [28]; punishment such as penalties [29]; cost of compliance [30]; enforcement efforts such as audit; probability of detection; differences between cultures; perceived behavioral control [31]; ethics or morality of the taxpayer and tax collector; equity of the tax systems; demographic factors such as sex, age, education and size of income and use of informants [32].

Therefore, tax fraud detection, enforcement and the behavior of others affect taxpayer compliance [33]. The IRS Commissioner Charles Rossotti noted that when the number of audits is reduced, honesty suffers as fears of policing decline.

Additionally, if taxpayers begin to believe that others are cheating, then the temptations to shave their own tax burdens

How to cite this article: Felix B, Zama V. Artificial Intelligence System for Value Added Tax Collection via Self Organizing Map (SOM). J Forensic Sci & Criminal Inves. 2024; 18(2): 555981. DOI: 10.19080/JFSCI.2024.18.555981

may become irresistible. Commissioner Rossotti's observations recognize that tax fraud detection and enforcement affect social behaviors, and that these behaviors can, in turn, affect taxpayers' compliance decisions [33]. Accordingly, the probability that a taxpayer will escape their tax obligations increases when the taxpayer suspects that his associates, colleagues, and acquaintances are evading taxes [34].

Vat Fraud Detection AI e-Platform Framework

(Figure 1)

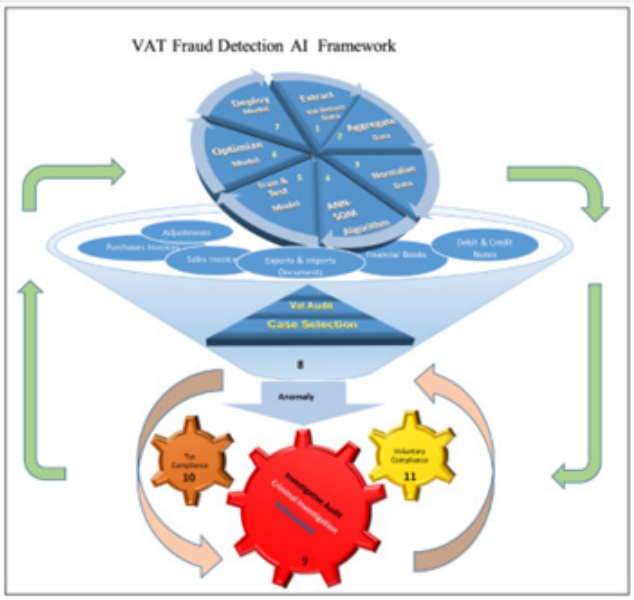


Figure 1: VAT fraud detection AI e-platform framework, source: (Bankole and Vara, 2023.

Overview of the framework

The VAT Fraud Detection AI e-platform framework employs a Self-Organizing Map (SOM) neural network. The framework is designed with the objective of assisting with the VAT audit case selection. Furthermore, we envisage that the model should improve the effectiveness and efficiency of the revenue collection agencies in identifying anomalies on VAT returns filed by the taxpayers. The framework hereunder classifies and segregate taxpayers into clusters or categories that have the greatest likelihood of committing fraud. Thus, the framework selects taxpayers for audit on the probability that they have committed fraud. The VAT Fraud Detection AI Framework proposed herein, is an amalgam of a typical industry standard machine-learning life cycle and tax authorities' VAT auditors' standard guide.

Flow of the framework

Task 1 – Extract VAT return data: According to the industry standard machine learning life cycle, this task is conducted under the data gathering phase of the life cycle. This step's goal is involved with the collection of data and the integration of data obtained from various sources such as files, database, the internet, or mobile devices. It is one of the most important steps of the life cycle. The quantity and quality of the collected data will determine the efficiency of the output. The more data we collect, the more accurate will be the classification or prediction.

Task 2 – Aggregate data: After collecting the data, we need to prepare it for further steps. In the ML life cycle this task is completed under the Data preparation phase. Data preparation is a step where we put our data into a suitable database or files and prepare it to use in our machine learning training. In our framework, for each VAT dealer, we aggregate all numerical continuous variable obtained from the return. In this study the summary values. over a period of 6 years, and are calculated for each individual VAT vendor. This effectively allows the algorithm to have a longer-term view of the vendor behaviour as opposed to monthly or yearly scrutiny. During this task will also conduct data pre-processing and exploratory data analysis.

Task 3 – Normalize data: This task is normally undertaken under the Data preparation phase of the Machine Learning Lifecycle. During data preparation we use a technique called normalization or standardization, to rescale our input and output variables prior to training a neural network model. The purpose is to normalize the data to obtain a mean close to zero. The review of the literature reveals that normalization could improve performance of the model [35]. Normalizing the data generally speeds up learning and leads to faster convergence. Accordingly, mapping data to around zero produces a much faster training speed than mapping them to the intervals far away from zero or using unnormalized raw data.

How to cite this article: Felix B, Zama V. Artificial Intelligence System for Value Added Tax Collection via Self Organizing Map (SOM). J Forensic Sci & Criminal Inves. 2024; 18(2): 555981. DOI: 10.19080/JFSCI.2024.18.555981

Task 4 – ANN-SOM Algorithm: Formally this stage is about selecting an appropriate Machine Learning Algorithm. This is an iterative process. During this study, we identified multiple machine learning algorithms applicable to our data and VAT fraud detection challenges. Therefore, as mentioned previously we shall use an unsupervised learning approach which is appropriate for unlabeled data. The algorithms we evaluated were K-means and Self Organizing Maps (SOM). According to Riveros et al. [36] the model trained with SOM outperformed the model trained with K-means. In their study they found that the SOM improved detection of patients having vertebral problems [36]. Likewise, after a few iterative processes, comparing the SOM and K-means performances, we chose the ANN-SOM algorithm.

Task 5 – Train and Test model: This stage is concerned with creating a model from the data given to it. At this stage we split the dataset into training and test datasets: 20% for testing and 80% for training. Herein, the training process is unsupervised. The remaining dataset is then used to evaluate the model. These two steps are repeated a number of times in order to improve the performance of the model [36].

Task 6 – Optimize model: A model's first results are not its last. The objective of the optimization or tuning to improve performance of the model. Tuning a model involves changing hyper parameters such as learning rate or optimizer [37]. The result for tuning and improving the model should be repeatability, efficiency and to reduce the training time. Someone should be able to reproduce the steps one has taken to improve performance.

Task 7 – Deploy model: The aim of this stage is the proper functionality of the model after deployment. The models should be deployed in such a way that they can be used for inference as well as be updated regularly [37].

Task 8 – VAT audit Case selection: The cohort of VAT vendors with return declarations that have been identified by the SOM as suspicious land up in the "funnel" for further scrutiny. This step is comprised of human verification. This audit is merely a general audit of cases selected for further scrutiny. This contrasts with an Investigative Audit, which is concerned with the auditing of cases by a specialist auditor.

Task 9 –Investigative audit, criminal investigation, and enforcement: Investigative audits are different from other tax audits in that a centralized specialist team conducts them. Task 9 is undertaken based on the results obtained from the previous audits conducted in Task 8 above, where audit officers have identified evidence of serious fraud.

Task 10 - Tax compliance: The tax compliance task is involved with the scrutiny of compliance related attributes like filing returns on time, timely payments, accurate completion of returns and timely registration with the tax authority, among others.

Task 11 – Voluntary compliance: The aim of the VAT fraud detection AI framework is to increase voluntary compliance. The level of audit activity and frequency of audit will be dictated by the availability of staff resources. The convenience of the AI framework suggested herein, is that it will ensure that the available staff resources are deployed judiciously with the twin objectives of maximizing both revenue collection and voluntary compliance by VAT dealers.

The "filter" or "funnel" described in the task 8, symbolizes the audit process, which involves a detailed human verification and validation of lading. This in turn assists in the independent verification financial records such as sales invoices, purchase invoices, customs documents, and bank cash deposits. However, the scope of the human verification is limited to the subset of taxpayers that have been flagged as anomalies by the SOM algorithm we propose. Once human verification has confirmed the presence of suspicious VAT declarations, such cases are then dealt with in task 9. Task 9 is a depiction of the work performed by investigative audit, criminal investigation, and enforcement teams, on confirmed cases. With this framework we envisage, that the effectiveness and efficiency of this AI assisted compliance framework will enhance detection of suspicious VAT vendors. Consequently, we anticipate that tax compliance will improve as the fear of detection increases (Task 10). Voluntary compliance will be a consequence of an improved, effective, and efficient AI based case selection technique (Task 11).

Material and Methodology

Data collection

We employed a rich data set, that is, the totality of VAT returns covering 6 years from 2013 to 2018. In order to delineate our data collection techniques, we have chosen to concentrate on only one type of industry, that is, mining. We have collected VAT returns for the complete list of registered vendors for six tax years, that is, 2013 to 2018. The firms have been anonymized so that we cannot link them with any publicly available data, however, they have been assigned identifying numbers so that we can follow a firm over time. The data contains detailed information on the line items in the returns, which is the VAT 201 declaration form of the South African tax administration. For instance, from the VAT return we managed to acquire 35 continuous variables. For ethical and confidentiality reasons, we shall not list all 35 variables, but only a subset of variables (Table 1).

Data preparation

According to Peck et al. [38] data preparation is the cleansing and organizing of real-world data, which is known to consume more than 80% of the time of a data scientist's work. Real-world data or raw data is dirty, full of missing values, duplicates and in some cases incorrect information [38]. Most machine-learning algorithms cannot deal with missing values. Hence, the data needs

How to cite this article: Felix B, Zama V. Artificial Intelligence System for Value Added Tax Collection via Self Organizing Map (SOM). J Forensic Sci & Criminal Inves. 2024; 18(2): 555981. DOI: 10.19080/JFSCI.2024.18.555981

to be converted and cleansed. In handling missing values, we dropped rows and then applied linear interpolation using mean values. Depending on the importance of the variable or feature the number of the missing values, any one of these solutions can be employed [38].

Table 1: Structure of the mining industry dataset.

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	.	.	.	X_{35}
Taxpayer ID	**Sales**	**Cost of Sales**	**Gross Profit**	**Output VAT**	**Input VAT**	**VAT Liability**	**VAT Refund**	**Diesel Refund**				
1	R1212446 978 150.00	R735 660 770 375.28	R R476 786 207 774.72	R137 044 069 237.86	R123 624 686 634.83	R 317 778 373.28	R 0.00	R 13 101 604 229.75	.	.	.	
2	R967 910 073 273.00	R 3 836 945 543.68	R R964 073 127 729.32	R 3 450 325 920.12	R 868 493 579.41	R2581 832 340.71	0	R 0.00	.	.	.	.
3	R817 769 471 754.00	R 353 988 805 850.56	R 463 780 665 903.44	R52 232 186 710.36	R45 914 952 221.56	R6317 234 488.80	R 0.00	R 0.00	.	.	.	.
4	R748978318 405.00	R400 865 955 921.92	R 348 112 362 483.08	R 59 559 939 667.48	R50 755 990 092.94	R8 803 949 574.54	0	R 0.00	.	.	.	.
5	724 874 587 388.00	(572 778 725 336.32	R 2152 095 862 051.68	R77 918116 069.42	R73 777 853 637.93	R4140 262 431.49	R 0.00	R 0.00	.	.	.	.
6	R 1674728 264 919.00	R 109 670 438 986.80	R (565 057 825 932.20	R (107 055 487 533.24	R94973 554317.75	R12 081 933 215.49	R 0.00	R 0.00	.	.	.	.
7	R638 832 370 992.00	R 5 324 404 259.92	R 133 507 966 732.08	R2 418 995 010.17	R1161 380 452.89	R1 257 614 557.28	R 0.00	R0.00	.	.	.	.
8	R 633 879 259 064.00	R299 756 674365.44	R 334 122 584 698.56	R47311 110 861.46	R37 469 584 295.68	R 9 841 526 565.78	R 0.00	R 0.00	.	.	.	.
9	R 629 586 422 707.00	R81 072 848 173.44	R 548 513 574 533.56	R7 060 139 836.59	R10975 416 041.30	R 0.00	R3915 276 204.71	R 0.00	.	.	.	.
10	R R606010 046 652.00	(484 204 023 786.24	R 121 806 022 865.76	R64996 149561.37	R62 051 600 429.02	R2944 549 132.35	R 0.00	R 0.00	.	.	.	.
11	R493 196 143 122.00	163 033 715 238.48	R 1330 162 427 883.52	R54112 112 844.71	R29 301 007 301.74	R23 435 241 483.49	R 0.00	R 1 375 864 059.48	.	.	.	.
12	R460 896 686 318.00	R 37 466 602 444.56	R 423 430 083 873.44	R3 573 617 053.50	R4 713 647 419.40	R0.00	R1 140 030 365.90	R 0.00	.	.	.	.
13	R422 003 518 129.00	R274 858 415 633.60	R 147 145 102 495.40	R8423 482 525.95	R34362 307 011.54	R 0.00	R 25 938 824 485.59	R0.00	.	.	.	.
14	R419 665 074 236.00	R217 144 516 505.04	R 202 520 557 730.96	R51 131 050 886.13	R 32957 036 016.75	R 18174 014 869.38	R 0.00	R0.00	.	.	.	.
15	R417 831 407 856.00	R52 314 970 087.52	R365 516437 768.48	R5 873 623 352.41	R 7 229 643 384.15	R 0.00	R1356 020 031.74	R 0.00	.	.	.	.
16	R417 005 827 949.00	1270 400 612 136.64	R 1146 605 215 812.36	R7 566 014 278.70	R36965 679 402.26	R 0.00	R30 194 949 175.46	R R 795 284 051.90	.	.	.	.

How to cite this article: Felix B, Zama V. Artificial Intelligence System for Value Added Tax Collection via Self Organizing Map (SOM). J Forensic Sci & Criminal Inves. 2024; 18(2): 555981. DOI: 10.19080/JFSCI.2024.18.555981

17	R 1398 680 526 777.00	R 6 526 719 179.44	R R392 153 807 597.56	R2217 063 208.09	R 1 145 154 756.92	R 1 071 908 451.17	R 0.00	R0.00	.	.	.	.
18	R R387 782 553 188.00	R60 109 908 420.00	R 327 672 644 768.00	R 6 389 824 208.18	R 8 108 921 654.57	R 0.00	R1719 097 446.39	R0.00	.	.	.	.
19	R 1387 151 365 653.00	1387 359 323 750.72	R 0.00	R20 233 017 741.27	R49399 701 197.34	R 0.00	R29166 683 456.07	R 0.00	.	.	.	.
20	R378 042 694350.00	R 326 584 344 108.40	R51 458 350 241.60	R30 365 638 345.91	R41 475 902 990.86	R 0.00	R11110 264 644.95	R 0.00	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.
5065	.	.	.	.	.	.	.	.	.	.	.	.

Consequently, we are fortunate in that we obtained a clean and high-quality dataset. However, the VAT, return dataset we obtained was at monthly level. We subsequently summed up all variables to annual values. The aggregation of all numerical variables of the VAT returns spans a six-year period from in 2013 to 2018. Thereafter the rand value amount was converted into ratios for ease of comparison. Nevertheless, as stated before the details of some of the variables that we used in this study could not be reported herein, due to the confidential nature of the tax audit process. Doing so can increase the potential for reverse engineering of the audit process, which is clearly undesirable and unlawful. However, each VAT ratio is designed from a point of view that a significantly higher or lower ratio value in relation to the rest of the sample or observations could arouse suspicion. In the opinion of Pamela Castellón González & Juan D. Velásquez fraud cases are most likely to occur among the extreme values of variables [39].

Dataset

The dataset consists of 5065 observations with 35 continuous variables. The observations are of Value Added Tax declarations or returns filed with the South African tax administration. However, the structure of the dataset showing sample variables can be seen in Table 1. In addition, we do provide aggregate indicative results that demonstrate the effectiveness of our approach. Thus, the 35 variables or attributes we have selected for this study are: gross income profile, income source, expense profile, source of purchases, tax payable or refundable, sales destination, imports or export purchases, accuracy of the declarations, overdue payments of taxes due to the tax authority, market segments, taxpayer industry, demographics, and the size of the firm.

Exploratory data analysis

In this section we use graphs, visualization, and transformation techniques to explore the VAT dataset (Table 2) in a systematic way. Statisticians call this task exploratory data analysis, or EDA for short. EDA is a repetitive cycle to firstly, give rise to questions about our data. Secondly, during this phase, we look for answers by visualizing and transforming the dataset population and lastly, we use what we have learnt to refine the questions and/or generate new questions. EDA is not a formal process with a strict set of rules [38]. We hope this initial data analysis will provide insight into important characteristics of the data.

Furthermore, we anticipate that EDA can provide guidance in selecting appropriate methods for further analysis. Additionally, we shall use summary statistics to provide information about our dataset. During this stage, we envisage that the summary statistics will tell us something about the values in the dataset. This includes where the average and median lies and whether our data is skewed or not. According to Peck & Devore [38], summary statistics fall into three main categories. That is, measures of location, measures of spread and graphs [38].

The measures of location will tell us where the data is centered. It will also tell us where a trend lies. Therefore, we shall use Mean, Median and Mode. The arithmetic mean, also called the average, is the central value of a discrete set of numbers. Specifically, it is the sum of the values divided by the number of values. The median is the middle of a data set. The mode of the data set tells us which value is the most common. On the other hand, measures of spread will tell us how spread out our data set is. According to Peck & Devore [38], the Range (inclusive of the Interquartile range and the Interdecile range), the Standard deviation, the Variance and the Quartiles are examples of measure of spread. The Range depicts how spread-out, is our data. The Interquartile range will tell us where the middle 50 percent of our

How to cite this article: Felix B, Zama V. Artificial Intelligence System for Value Added Tax Collection via Self Organizing Map (SOM). J Forensic Sci & Criminal Inves. 2024; 18(2): 555981. DOI: 10.19080/JFSCI.2024.18.555981

data is located. Whilst the Quartiles will illustrate the boundaries of the lowest, middle, and upper quarters of the dataset [38]. A correlation matrix was used to quantify dependencies between 35 continuous variables. For this, a Pearson correlation matrix was calculated for all 35 variables. A correlation value between two variables that has an absolute value greater than 0.7 is considered as high and therefore the variables are highly closely related to each other. The objective of this analysis is to establish whether two variables are correlated to each other, and not that they are necessarily causal. The sign of the actual value, which is either positive or negative, provides information about whether two variables are positively or inversely related to each other [40,41]. For example, the correlation value between sales and input vat is 0.98, meaning that there is a direct positive relationship between the two variables. This is rational, because input vat is charged on all purchases of goods and services, which will later become sales of goods and services by the entity. (Table 2).

The correlation matrix and heat maps generated across the 35 variables are a valuable visual representation of VAT data set trends. While the correlation matrix and the heat maps produce the same conclusion, the heat maps can provide further information about the distribution and localization of correlated variables. Our method of generating heat maps can visualize the correlations between multiple variables, providing a broader analysis than using a correlation matrix.

Table 2: VAT dataset variables descriptions.

Variable	Description
Sales	Sales is the amount paid by the customers for goods or services supplied by the VAT Vendor. The Sales amount is inclusive of Output VAT
Cost of Sales	Cost of Sales refers to the direct costs of producing the goods or services sold by a company or VAT Vendor. It excludes indirect expenses, such as distribution costs and sales administration costs. Cost of Sales is also referred to as "Cost of goods sold (COGS)". Cost of Sales is inclusive of input VAT
Gross Profit	Gross profit is the profit a company or VAT vendor makes after deducting the costs associated with making and selling its products or services.
Output VAT	Output VAT is the value added tax the VAT Vendor charges on their own sales of goods or services both to other businesses and to ordinary consumers. This is the amount that the Vendor must pay over to the Tax Administration
Input VAT	Input VAT is the value added tax added to the price when the VAT vendor purchase goods or services that are liable to VAT. The Vendor can deduct the amount of VAT paid from their settlement with the tax authorities.
VAT Liability	VAT liability is the final amount payable to the tax administration after all allowable deductions
VAT Refund	A VAT refund is an amount of VAT that is payable by the tax administration to a VAT vendor, where. the total amount of input tax less allowable deductions, exceeds the total amount of output tax in a particular tax period, or a vendor has paid an amount of VAT, in excess of the amount that should have been paid to the tax authority.
Diesel Refund	Vendors who are registered for the Diesel Refund Scheme can deduct the diesel rebate from Net VAT. Net VAT is Output VAT less Input VAT.
.	.
.	.
.	.
x_{35}	.

How to cite this article: Felix B, Zama V. Artificial Intelligence System for Value Added Tax Collection via Self Organizing Map (SOM). J Forensic Sci & Criminal Inves. 2024; 18(2): 555981. DOI: 10.19080/JFSCI.2024.18.555981

Summary statistics (Table 3)

Table 3: Summary statistics.

Output VAT		Input VAT		Sales		Cost of Sales		χ_{35}
Mean	R 59 912 932.40	Mean	R 91 731 312.69	Mean	R 1 028 844 838.79	Mean	R 662 200 299.66	.
Standard Error	R 7 324 388.00	Standard Error	R 12 859 071.46	Standard Error	R 142 377 028.27	Standard Error	R 95 159 622.47	.
Median	R 885 430.17	Median	R 784 570.07	Median	R 7 905 662.00	Median	R 4 813 261.36	.
Mode	R 0.00	Mode	R 0.00	Mode	R 0.00	Mode	R 0.00	.
Standard Deviation	R 521 268 002.97	Standard Deviation	R 915 164 856.98	Standard Deviation	R 10 132 804 156.97	Standard Deviation	R 6 772 397 414.28	.
Kurtosis	R 962.53	Kurtosis	R 767.67	Kurtosis	R 734.23	Kurtosis	R 774.23	.
Skewness	R 25.96	Skewness	R 24.31	Skewness	R 23.35	Skewness	R 24.57	.
Range	R 23 743 554 071.51	Range	R 36 965 679 402.26	Range	R 417 005 827 949.00	Range	R 270 400 612 136.64	.
Minimum	R 0.00	Minimum	R 0.00	Minimum	R 0.00	Minimum	R 0.00	.
Maximum	R 23 743 554 071.51	Maximum	R 36 965 679 402.26	Maximum	R 417 005 827 949.00	Maximum	R 270 400 612 136.64	.
Sum	R 303 459 002 601.77	Sum	R 464 619 098 750.49	Sum	R 5 211 099 108 494.00	Sum	R 3 354 044 517 797.61	.
Observations	5065	Observations	5065	Observations	5065	Observations	5065	5065

Correlation matrix (Table 4)

Table 4: Correlation matrix.

	Output VAT	Input VAT	VAT Refund	VAT Payable	Sales	Cost of Sales	Gross Profit	χ_{35}
Output VAT	100%	76%	37%	49%	72%	75%	56%	.
Input VAT	76%	100%	88%	15%	98%	100%	81%	.
Diesel refund	34%	53%	58%	7%	61%	51%	70%	.
VAT refund	37%	88%	100%	-1%	90%	88%	81%	.
VAT payable	49%	15%	-1%	100%	22%	14%	33%	.
Input VAT on capital goods	61%	81%	74%	21%	84%	77%	83%	.
Zero rated plus _exempt sales	40%	88%	98%	3%	93%	87%	90%	.
Sales	72%	98%	90%	22%	100%	97%	91%	.
Cost of sales	75%	100%	88%	14%	97%	100%	79%	.
Gross profit	56%	81%	81%	33%	91%	79%	100%	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
χ_{35}	χ_{35}	χ_{35}	χ_{35}	χ_{35}	χ_{35}	χ_{35}	χ_{35}	χ_{35}

How to cite this article: Felix B, Zama V. Artificial Intelligence System for Value Added Tax Collection via Self Organizing Map (SOM). J Forensic Sci & Criminal Inves. 2024; 18(2): 555981. DOI: 10.19080/JFSCI.2024.18.555981

Correlation plots (Figure 2 -7).

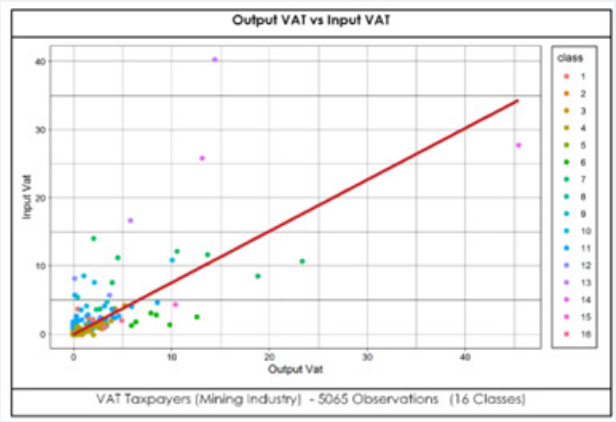


Figure 2: Output VAT vs Input VAT.

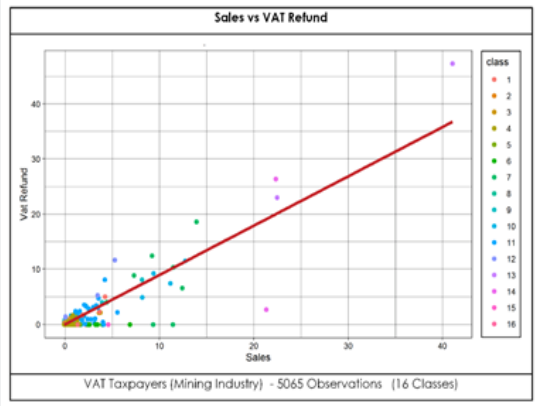


Figure 3: Sales vs VAT refund, source: prepared by authors (2022).

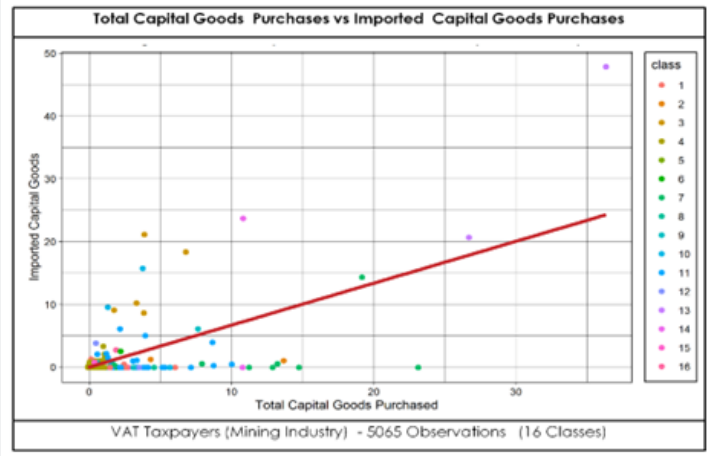


Figure 4: Total capital goods vs imported capital goods purchases. source: prepared by authors (2022)

How to cite this article: Felix B, Zama V. Artificial Intelligence System for Value Added Tax Collection via Self Organizing Map (SOM). J Forensic Sci & Criminal Inves. 2024; 18(2): 555981. DOI: 10.19080/JFSCI.2024.18.555981

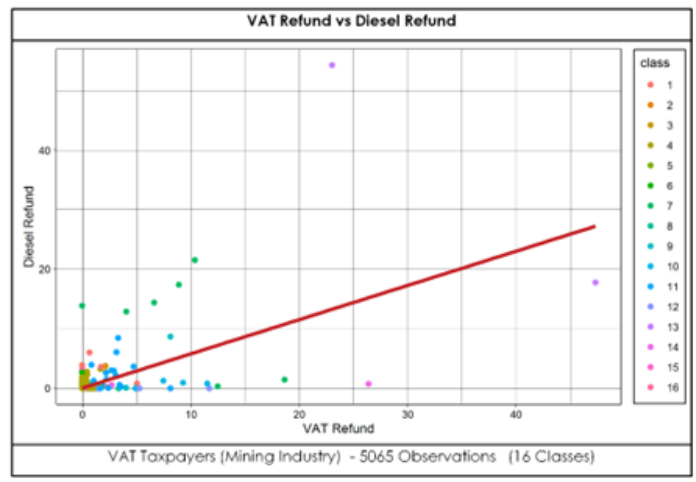


Figure 5: VAT refund vs diesel refund. source: prepared by authors (2022).

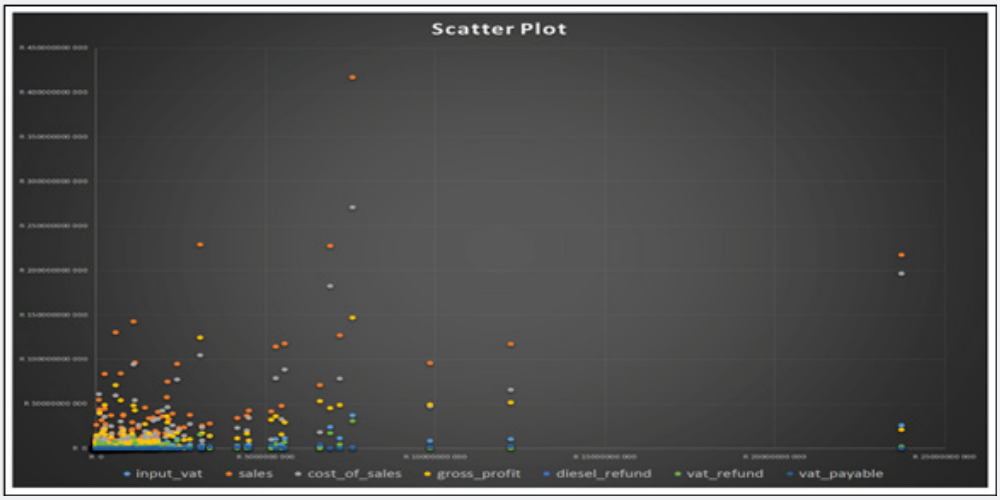


Figure 6: Scatter plot. source: prepared by authors (2022)

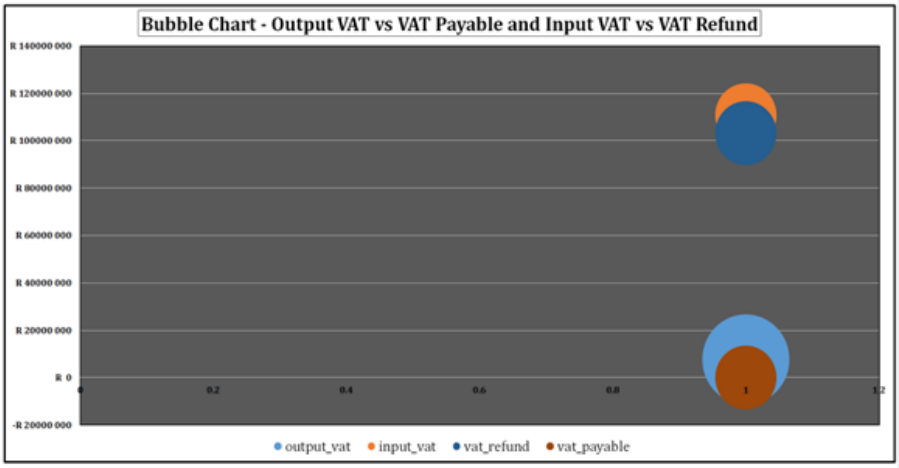


Figure 7: Bubble chart source: prepared by authors (2022).

How to cite this article: Felix B, Zama V. Artificial Intelligence System for Value Added Tax Collection via Self Organizing Map (SOM). J Forensic Sci & Criminal Inves. 2024; 18(2): 555981. DOI: 10.19080/JFSCI.2024.18.555981

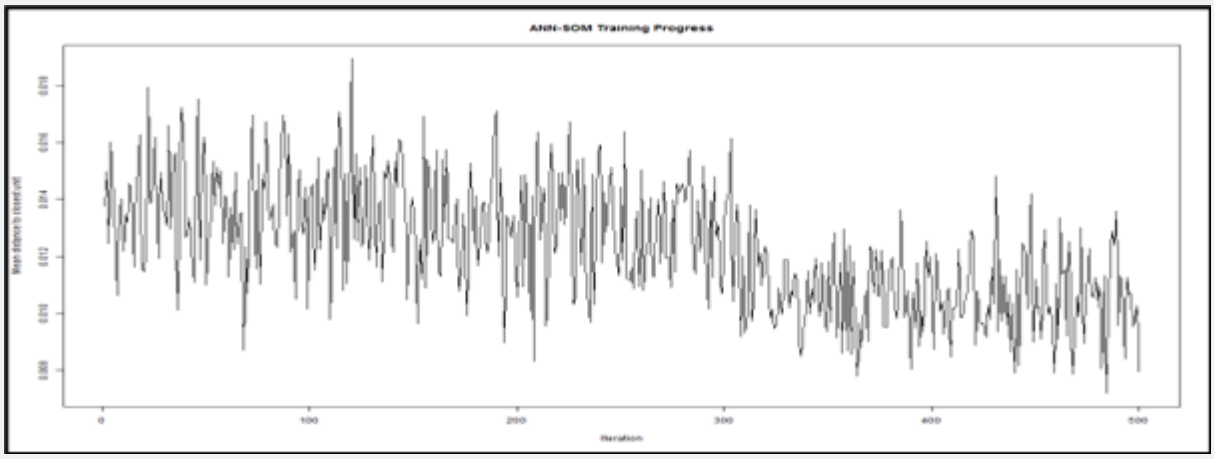


Figure 8: Artificial neural network–SOM training progress. source: prepared by authors (2022).

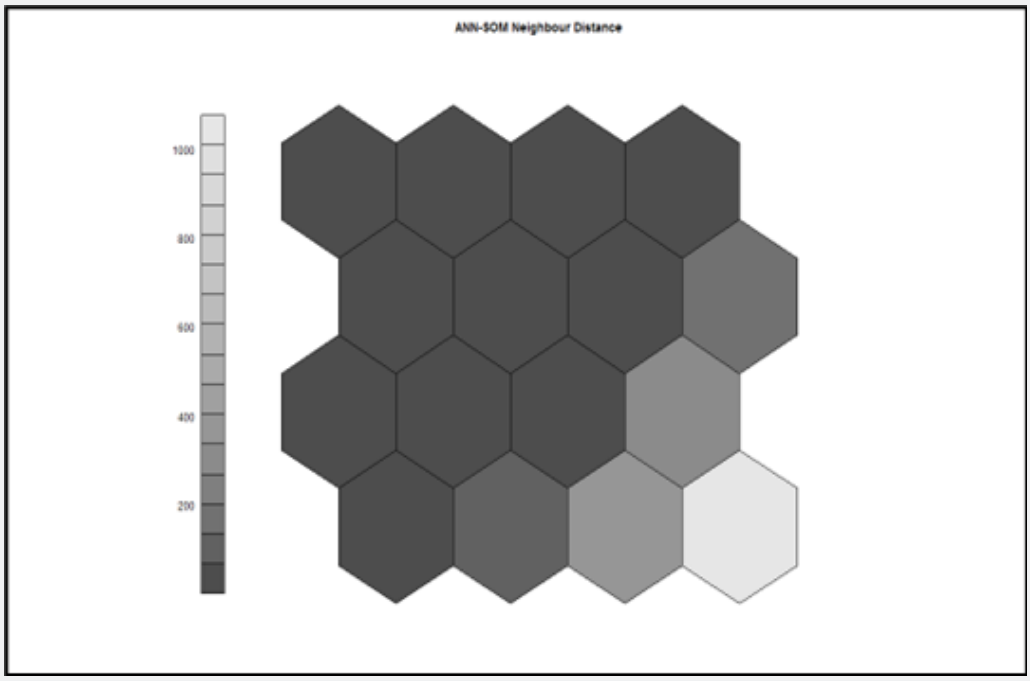


Figure 9: Artificial neural network-SOM neighbour distance. source: prepared by authors (2022).

Results and Discussion

Data pre-processing

Normalizing the data as mentioned in Task 3, generally speeds up learning and leads to faster convergence. Accordingly, mapping data to around zero produces a much faster training speed than mapping them to the intervals far away from zero or using un-normalized raw data. Academic researchers [35] point to the importance of data normalization prior to the neural network training to improve the speed of calculations and obtain satisfactory results in nuclear power plant application. In the opinion of various authors statistical normalization techniques enhance the reliability and the performance of the trained model [42].

SOM training algorithm

In training algorithm, the SOM map is trained iteratively by taking training data vectors one by one from a training data vector

How to cite this article: Felix B, Zama V. Artificial Intelligence System for Value Added Tax Collection via Self Organizing Map (SOM). J Forensic Sci & Criminal Inves. 2024; 18(2): 555981. DOI: 10.19080/JFSCI.2024.18.555981

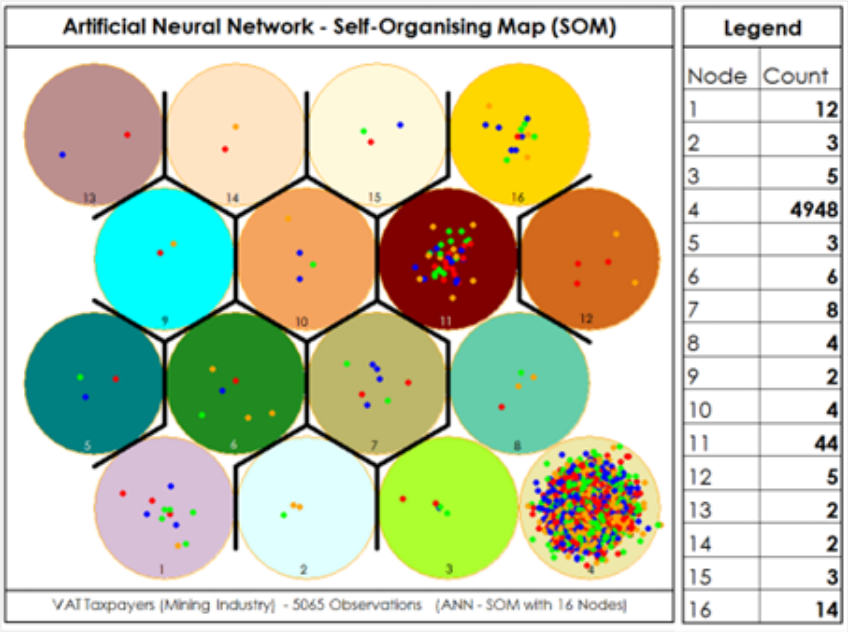
sequence, finding the Best Matching Unit (BMU) for the selected training data vector on the map and updating the BMU and its neighbors closer toward the data vector. This process of finding the BMU and updating the prototype vectors are repeated until a predefined number of training iterations or epochs is completed. The SOM training progress plot is depicted in figure 8 below.

SOM neighbour distance

The neighbor distance is often referred to as the "U-Matrix". This visualization is of the distance between each node and its neighbors. Typically viewed with a grayscale palette, areas of low neighbor distance indicate groups of nodes that are similar. Areas with large distances indicate the nodes are much more dissimilar. Furthermore, they indicate natural boundaries between node clusters. Them SOM Neighbor distance plot (Figure 9) [43-49].

Artificial neural network-SOM heat map

The ANN-SOM heat map is the outcome of the Neural Network Self-Organizing map (SOM) algorithm we trained on the VAT dataset, which has 35 continuous numeric variables. The heat map shows the distribution of all variables across the SOM. We stated before that the dataset used in this experiment spans a period of 6 years from 2013 to 2018. The outcome is a grid of 16 Nodes from 5065 observations belonging to the mining industry (Figure 10) [50-55].



Legend	
Node	Count
1	12
2	3
3	5
4	4948
5	3
6	6
7	8
8	4
9	2
10	4
11	44
12	5
13	2
14	2
15	3
16	14

Figure 10: Artificial neural network-SOM similarity heat map. source: prepared by authors (2022).

Results and Conclusion

As already mentioned, our VAT dataset consists of 35 continuous variables. The continuous variables are made up of variables like Output VAT, Input VAT, Sales, Cost of Sales, VAT Refund, VAT Payable, to name just a few. The sample size consists of 5065 different VAT taxpayers all belonging to the mining industry. In our experiments, the ANN-SOM map size is 4x4. The SOM cluster heat map contains sixteen distinct clusters and the dots on the cluster represent individual taxpayers or entities. In the ANN-SOM Heat map above we observe that VAT vendors or entities, for example with similar VAT return characteristics, are grouped in the same area or node. In business, users are more interested in "abnormal clusters" or hot spots. That is, clusters of VAT vendors who have suspicious behaviour rather than normal nodes or clusters. We use three approaches to identify hot spots. That is, by using the ANN-SOM Heat map. Distance matrix Visualization and domain experts' feedback based on component plane visualizations. Using distance matrix visualizations, homogenous clusters (low variation) will have shorter neighbor distances (the white regions) compared to high variation clusters (the dark regions) as shown in Figure 9. The value of a component in a node is the mean value of entities. (VAT Vendors) in the node and its neighbors. The average value of entities is determined by the neighborhood function and the final radius used in the final training (Figure 8). The color coding of the map is created based on the minimum and maximum values of the component of the map. In this research paper, we use the grey color map where the maximum value is assigned black, and the minimum value is assigned white.

How to cite this article: Felix B, Zama V. Artificial Intelligence System for Value Added Tax Collection via Self Organizing Map (SOM). J Forensic Sci & Criminal Inves. 2024; 18(2): 555981. DOI: 10.19080/JFSCI.2024.18.555981

However, when interpreting the ANN-SOM Heat map, the abnormal clusters are those that have a fewer number of entities. That is, these nodes are composed of suspicious VAT vendors. Such VAT vendors require detailed human verification by VAT audit specialists. Node 4, for example, has the largest number of entities at 4948. The entities clustered in Node 4 are homogeneous in nature, and thus depict VAT entities with normal behaviour. VAT fraud or suspicious behaviour can be differentiated by observing VAT declarations form attributes such as VAT Liability, Exempt supplies, Diesel Refund, and Input VAT on Capital Goods purchased. Detection of suspicions VAT declarations is a very challenging task as VAT declarations dataset are extremely unbalanced in nature. Furthermore, the tax fraud domain is full of unlabeled data, which in turn makes it difficult to use supervised learning approaches. In this research paper, we proposed an unsupervised learning approach. Nevertheless, it is crucial to have an all-encompassing review on detecting VAT fraud. This is to broaden the understanding and knowledge of the VAT fraud phenomenon among researchers and in the government marketing domain. Remarkably supervised learning algorithms have proved to be limited in the arena of VAT fraud detection, since the tax administrations have extremely low to non-existent labelled historic data. This in turn cripple the efficacy of supervised learning approaches.

In as much as this paper's focal point is on VAT fraud detection, we are confident that the present model may just as well be applicable to other tax types, like Industry Income Tax and Personal Income Tax for instance. This research outcome shows the potential of AI techniques in the realm of VAT fraud. Furthermore, this review put forward high-level and detailed digital classification frameworks on VAT fraud detection. Additionally, the e-platform framework proposed present tax auditors with a systemic case selection guide of suspicious VAT returns. Consequently, combining the two frameworks into a single hybrid approach can improve the success of detecting other VAT fraud schemes. Tax administration may be able to select the most appropriate unsupervised learning technique from this work having considered other alternatives, their operational requirements and business context. Thus, leading to a multitude of available AI aided VAT fraud detection algorithms and approaches. Additionally, the techniques proposed in this paper should help tax administrations with precise case selection using an empirical and data-driven approach, which does not depend upon a labeled historic VAT dataset. Furthermore, we envisage the approach should result in high hit ratio on suspicious VAT returns, and thus improve tax compliance due to the increased likelihood of detection. We have demonstrated the use of ANN-SOM in exploring hot spots in a large real-world Value-Added Tax domain. Based on our experiments, our approach is an effective instrument for hot spots and heat map exploration since it employs visualizations techniques that are easy to understand. In future different profiling or clustering algorithms and sampling techniques can be applied to further improve the performance of the proposed approach. Notwithstanding, compared with supervised approaches, unsupervised methods are less precise. That is, they will not only identify tax fraud cases, but will also indicate taxpayers with irregular and suspicious tax behavior and dishonest taxpayers. Future research studies using hybrid algorithms may produce higher quality research outcomes.

References

McIntyre DP, Srinivasan A (2017) Networks, platforms, and strategy: Emerging views and next steps. Strategic Management Journal 38: 141–160.

1. McIntyre DP, Srinivasan A (2017) Networks, platforms, and strategy: Emerging views and next steps. Strategic Management Journal 38: 141–160.
2. Stallkamp M, Schotter APJ (2019) Platforms without borders? The international strategies of digital Platform firms. Global Strategy Journal 11(1): 58-80.
3. Hagiu A (2009) Two-sided platforms: Product variety and pricing structures. Journal of Economics and Management Strategy 18: 1011–1043.
4. Bankole FO, Bankole OO (2017) The effect of Cultural Dimension on ICT Innovation: Empirical Analysis of Mobile Phone Services. Telematics and Informatics 34 (2017): 490-505.
5. Bankole F, Taiwo A, Claim I (2022) An extended digital forensic readiness and maturity model. Forensic Science International: Digital Investigation 40: 301348 1-13.
6. Bankole F, Vara Z (2023) A Comparison of SOM and K-Means Algorithms in Predicting Tax Compliance. Encyclopedia of Data Science and Machine Learning. IGI Global 1-3149.
7. Huete-Alcocer N (2017) A Literature Review of Word of Mouth and Electronic Word of Mouth: Implications for Consumer Behavior. Front Psychol 8: 1256.
8. Bendor-Samuel P (2018) What is a digital platform? The Enterprisers Project. A community helping CIOs and IT leaders solve problems.
9. Haselton T (2018) Google's Assistant is getting so smart it can place phone calls and humans think it's real.
10. Dignan L (2018) The AI, machine learning, and data science conundrum: Who will manage the algorithms? Artificial Intelligence.
11. Vanhoeyveld J, Martens D, Peeters B (2019) Value-added tax fraud detection with scalable anomaly detection techniques. Applied Soft Computing Journal 86: 1-20.
12. Madongo I (2017) What is VAT fraud and why is the EU worried? KYC360 News.
13. Keen M, Smith S (2007) VAT Fraud and Evasion: What do we know, and what can be done ? International Monetary Fund, Fiscal Affairs Department. International Monetary Fund.
14. Georgieva, Markova, Pavlov (2019) Using Neural Network for Credit Card Fraud Detection. AIP Conference Proceedings 2159.
15. Kelechi (2018) Handling Imbalanced Datasets: Predicting Credit Card Fraud.
16. Kohonen T (1982) Self-Organized Formation of Topologically Correct Feature Maps. Biological Cybernetics. 43: 59-69.

How to cite this article: Felix B, Zama V. Artificial Intelligence System for Value Added Tax Collection via Self Organizing Map (SOM). J Forensic Sci & Criminal Inves. 2024; 18(2): 555981. DOI: 10.19080/JFSCI.2024.18.555981

17. Bansal M, Suman (2014) Credit Card Fraud Detection Using Self Organised Map. International Journal of Information & Computation Technology 4(13).

18. Marketing Evolution (2021).

19. Velampalli S, Eberle W (2017) Novel graph based anomaly detection using background knowledge, Proceedings of FLAIRS 2017, AAAI Press pp. 538–543.

20. Molsa M (2017) Success factors when implementing AI-powered marketing solutions.

21. Phua CWC, Alahakoon D, Lee VCS (2004) Minority Report in Fraud Detection: Classification of Skewed Data. SIGKDD Explorations 6(1): 50 - 59.

22. Shao H, Hong Z, Chang GR (2002) Applying Data Mining to Detect Fraud Behavior in Customs Declaration. Proceedings of the International Conference on Machine Learning and Cybernetics 3.

23. White MJ (2014) The SEC in 2004. 41st Annual Securities Regulation Institute. Coronado, California.

24. Larsson O (2021) AI & Digital Platforms: The Markets. In AI and Learning Systems-Industrial Application and Future Direction. IntechOpen Book Series.

25. American Marketing Association (2017) Definitions of Marketing. Retrieved May 30, 2021.

26. Smith WH (1970) Information Systems in Tax Administration. In: Proceedings of the Annual conference on Taxation under the Auspices of the National Tax Association 63 (1970): 267-283.

27. Kirchler, E (2007) The economic psychology of tax behaviour. Cambridge University Press, United Kingdom.

28. Ambrecht (1998) Increasing Taxpayers Compliance : A discussion of the Negligence Penalty. Retrieved May 29, 2021.

29. Silvani C (1992) Improving Tax Compliance. In: Improving Tax Administration in Developing Countries. Eds: Richard Bird and Milka Casanegra de Jantscher. International Monetary Fund, Washington, DC, USA.

30. Feld L, Frey B, Targler B (2006) Rewarding Honest Taxpayers? Paper presented on April 9-11, 2006 on "Managing and Maintaining Compliance ".

31. Allingham M, Sandmo A (1972) Income Tax Evasion: A theoretical analysis. Journal of Public Economics 1(34): 323-338.

32. Slemrod J (1992) Do Taxes Matter ? Lessons from the 1980's. American Economic Review. 82(2): 250-256.

33. Furnharn, T (1983) The Protestant Work Ethic, Human Values and Attitudes towards Taxation. Journal of Economic Psychology 3(2): 113-128.

34. Murphy K (2004) An examination of taxpayers' attitudes towards the Australian tax system: findings from a survey of tax scheme investors. Australian National University, Research School of Social Sciences, Canberra, Australia: Centre for Tax System Integrity, Working Paper No. 46.

35. Weisman J (2001) IRS audits at record low: Officials cite decade-long staff decrease, old computers, law defending taxpayers. USA Today.

36. Bird RM, Milka CD (1992) Improving Tax Administration in Developing Countries. International Monetary Fund, United States.

37. Sola J, Sevilla J (1997) Importance of input data normalization for the application of neural networks to complex industrial problems. IEEE Transactions on Nuclear Science 44(3): 1464-1468.

38. Riveros NA, Cardenas BA, Pico EE (2019) Comparison between K-means and Self-Organizing Maps algorithms used fordiagnosis spinal column patient. Informatics in Medicine Unlocked 16.

39. Bennett KP, Parrado Hernandez E (2006) The Interplay of Optimization and Machine Learning Research. Journal of Machine Learning Research 7: 1265-1281.

40. Peck R, Olsen C, Devorea L (2015) Introduction to Statistics and Data Analysis (5th Edn.), Brokes Cole, United States.

41. González PC, Velásquez JD (2013) Characterization and detection of taxpayers with false invoices using data mining techniques. Expert Systems with Applications 40(5): 1427-1436.

42. Ratner, B (2009) The correlation coefficient: Its values range between + 1 / - 1, or do they? International Journal of Computer Theory and Engineering 17: 139-142.

43. Serra P (2003) Measuring Performance of Chile's Tax Administration. National Tax Journal 56(2): 373-378.

44. Jayalakshmi T, Santhakumaran A (2011) Statistical normalization and back propagation for classification. International Journal of Computer Theory and Engineering 3(1): 89-93.

45. Aleksander I (2017) Partners of humans: a realistic assessment of the role of robots in the foreseeable future. Journal of Information Technology 32 (1): 1-9.

46. Araujo E, Silva C, Sampaio D (2008) Video Target Tracking by using Competitive Neural Networks. WSEAS Transactions on Signal Processing 4(8): 420-431.

47. Baesens B, Veronique, Vlasselaer V, Verbeke W (2015) Fraud analytics using descriptive, predictive, and social network techniques: a guide to data science for fraud detection. John Wiley & Sons, USA.

48. Bandara KG, Weerasooriya WM (2019) A Conceptual Research Paper on Tax Compliance and Its relationships. International Journal of Business and Management 14(10).

49. Bergman M, Nevarez A (2006) Do Audits Enhance Compliance? An Empirical Assessment of VAT Enforcement. National Tax Association 59(4): 817-832.

50. Flach P (2012) Machine Learning: The Art and Science of Algorithms that Make Sense of Data. University of Bristol, Cambridge University Press, United Kingdom.

51. Krishnamurthya R, Desouza KC (2014) Big data analytics: The case of the social security administration. Information Polity 19(1): 165-178.

52. Mahadevan, B (2000) Business Models for Internet based E-Commerce: An Anatomy. California Management Review 42(4): 55-69.

53. Nagadevara V, Kumar A (2006) Development of Hybrid Classification Methodology for Mining Skewed Data Sets – A Case Study of Indian Customs Data. Proceedings of the 4th ACS/IEEE International Conference on Computer Systems and Applications. United Arab Emirates.

54. Siddharth M, Hao L, Jiabo H (2019) Robust geo mechanical characterization by analyzing the performance of shallow-learning regression methods using unsupervised clustering methods.10.1016/B978-0-12-817736-5.00005-3. Gulf Professional Publishing, United States.

55. Smola A, Vishwanathan S (2008) Introduction to Machine Learning. Cambridge University Press, United Kingdom.

56. Swarnajyoti B, Lorenzo P (2014) A novel SOM-SVM based active learning technique for remote sensing image classification. IEEE Transaction on Geoscience and Remote Sensing 52(11): 6899-6910.

57. Tripathi A (2017) Practical Machine Learning - Cookbook. Packt Publishing Ltd, United Kingdom.

How to cite this article: Felix B, Zama V. Artificial Intelligence System for Value Added Tax Collection via Self Organizing Map (SOM). J Forensic Sci & Criminal Inves. 2024; 18(2): 555981. DOI: 10.19080/JFSCI.2024.18.555981

DOI: 10.19080/JFSCI.2023.18.555981

How to cite this article: Felix B, Zama V. Artificial Intelligence System for Value Added Tax Collection via Self Organizing Map (SOM). J Forensic Sci & Criminal Inves. 2024; 18(2): 555981. DOI: 10.19080/JFSCI.2024.18.555981