# Detecting Tax Evasion and Financial Crimes in The United States Using Advanced Data Mining Technique

Md Rakibul Haque Pranto[1]*, Ismoth Zerine[2], Md Mainul Islam[2], Morium Akter[3], Tauhedur Rahman[4]

## Abstract

Background: Financial crimes and tax evasion pose a major threat to the financial security and integrity of the United States. While regulatory mechanisms exist, detecting fraudulent activity within massive and complex financial datasets remains a significant challenge. Advanced data mining techniques offer the potential to improve the identification of irregularities and strengthen oversight systems. Methods: This study applied advanced data mining methods—including decision trees, neural networks, logistic regression, and K-means clustering—to identify patterns of financial crime and tax evasion. The dataset consisted of one million financial records obtained from publicly available IRS and SEC reports between 2015 and 2020. Stratified random sampling ensured broad representation across industries. Analytical approaches included logistic regression, K-means clustering, and correlation analysis, with performance evaluated through p-values, sensitivity, specificity, and area under the ROC curve (AUC). Results: The analysis showed that lower tax payments and higher anomaly scores were strong indicators of tax evasion. A significant positive correlation was observed between transaction volume and anomaly score ($r = 0.52$, $p < 0.0001$). Logistic regression identified tax paid, income, transaction volume, and anomaly score as significant predictors of tax evasion, achieving excellent classification accuracy with an AUC of 0.91. K-means clustering further classified entities into three risk categories, with the high-risk cluster exhibiting substantially higher values for income, transaction volume, and anomaly scores compared to other groups. Conclusion: The findings demonstrate the effectiveness of advanced data mining techniques for detecting tax evasion and financial crimes in large-scale datasets. By identifying key predictive indicators and stratifying risk, these methods show strong potential for integration into financial oversight systems. Their implementation could significantly enhance the capacity of regulatory frameworks to detect, prevent, and mitigate financial misconduct in real time.

Keywords: Tax evasion, financial crimes, data mining, logistic regression, anomaly detection.

***Significance*** | This study advances proactive fraud detection by applying data mining techniques to combat U.S. tax evasion and financial crimes effectively.

*Correspondence. Md Rakibul Haque Pranto, Department of Management, St. Francis College, USA. Email: ismothmona7171@gmail.com

Editor Mohammad Sahabuddin, PhD, And accepted by the Editorial Board Mar 07, 2023 (received for review Jan 01, 2023)

## 1. Introduction

Financial crimes and tax evasion remain persistent and pressing challenges that compromise the integrity, stability, and resilience of financial systems worldwide. These illicit activities have particularly detrimental consequences for national economies, eroding public trust, undermining fiscal stability, and imposing severe

**Author Affiliation.**
[1] Department of Management, St. Francis College, USA.
[2] College of Graduate and Professional Studies, Trine University, Angola, Indiana, USA.
[3] Department of Information Technology and Project Planning Management, St. Francis College, USA.
[4] Dahlkemper School of Business, Gannon University, USA.

**Please cite this article.**
Pranto, M. R. H., Zerine, I., Islam, M. M., Akter, M., Rahman, T. (2023). "Detecting Tax Evasion and Financial Crimes in The United States Using Advanced Data Mining Technique", Business and Social Sciences, 1(1),1-11,10337

3067-8919/© 2023 BUSINESS & SOCIAL SCIENCES, a publication of Eman Research, USA.
This is an open access article under the CC BY-NC-ND license.
(http.//creativecommons.org/licenses/by-nc-nd/4.0/).
(https./publishing.emanresearch.org).

administrative and enforcement burdens on regulatory institutions (Ashtiani & Raahemi, 2023). In the United States, where the financial sector is among the most sophisticated and globally integrated, tax evasion and related financial crimes not only contribute to substantial fiscal deficits but also generate cascading effects that destabilize financial markets, reduce investor confidence, and obstruct the effective functioning of public governance. The persistence of these crimes, despite decades of regulation, legal reforms, and technological interventions, underscores the complexity and evolving nature of financial malfeasance in contemporary contexts.

Even after years of intense regulation, investigation, and the establishment of comprehensive compliance mechanisms, financial crimes remain difficult to detect and prevent. This difficulty primarily stems from the massive volume, intricate complexity, and dynamic fluidity of financial data that contemporary institutions must process on a daily basis (Popoola et al., 2023). Traditional detection systems—such as manual auditing procedures and rule-based monitoring frameworks—have proven inadequate in identifying novel or subtle patterns of illicit behavior (Elumilade et al., 2021). Their reliance on predefined heuristics makes them reactive rather than proactive, leaving them unable to address emerging fraudulent schemes that continuously adapt to exploit systemic vulnerabilities.

Furthermore, the sophistication of modern fraudulent practices, coupled with the exponential growth in financial information generated by digitized transactions, has created an urgent demand for more advanced detection and prevention instruments (Samuel, 2023). This evolution has given rise to an increasing reliance on computational tools and intelligent frameworks capable of processing and analyzing vast quantities of structured and unstructured financial data. Among these, **data mining techniques**—which employ algorithms to uncover hidden patterns and anomalies within large datasets—have emerged as particularly promising. Data mining allows for the identification of complex, non-obvious correlations that are otherwise imperceptible through conventional auditing or compliance approaches.

This study is specifically concerned with examining the potential of modern data mining tools and methods to identify tax evasion and financial offenses in the United States. By doing so, it aims to address a critical gap in current systems of ownership and control, which often lack the scalability and flexibility required to combat rapidly evolving fraudulent activities. Leveraging algorithms such as decision trees, neural networks, and clustering models, the research proposes a more robust, detailed, and scalable solution for fraud detection (Olayinka, 2019). Unlike traditional auditing, which is labor-intensive and constrained by human subjectivity, data mining empowers regulators and financial institutions to proactively uncover anomalies, identify hidden risks, and adaptively respond to emerging forms of financial misconduct (Dong & Zhang, 2023). As such, data mining is not merely a technical enhancement but a potential paradigm shift in how financial crimes are detected, prevented, and deterred.

Illegal tax practices and financial crimes have long been recognized as significant impediments to sustainable economic growth and national development. In the U.S. context, tax evasion alone costs the federal government billions of dollars annually, significantly constraining its ability to finance essential public services, social programs, and infrastructure development (Ozili, 2020). Moreover, financial crimes such as money laundering, corporate fraud, embezzlement, and insider trading exacerbate systemic vulnerabilities by distorting the efficiency of markets, eroding investor confidence, and enabling the proliferation of organized criminal networks. Collectively, these crimes undermine the legitimacy of institutions, destabilize fiscal governance, and perpetuate inequality.

Although traditional approaches—such as statutory enforcement, manual audits, and heuristic-based fraud detection—have historically provided some deterrence, they are increasingly ineffective in today's digitized and globalized financial environment. The explosion of data generated by e-commerce, mobile banking, high-frequency trading, and other digital financial platforms complicates detection mechanisms, as fraudulent transactions are easily concealed within the noise of legitimate activity (Nicholls et al., 2021). Furthermore, the increasing complexity of financial products, including derivatives, cryptocurrencies, and offshore investment vehicles, has provided fertile ground for sophisticated tax avoidance and evasion schemes that cannot be easily identified using outdated methods (Alonge et al., 2021).

In recognition of these challenges, the U.S. government has deployed a series of institutional and regulatory measures designed to address financial crimes. Agencies such as the Internal Revenue Service (IRS) and the Securities and Exchange Commission (SEC) have implemented elaborate compliance frameworks, developed data-reporting systems, and established enforcement mechanisms (Brederode, 2019). These systems, however, typically rely on reactive measures grounded in historical patterns and static rules. They often fail to identify emerging forms of fraud or adapt quickly to changing conditions. Moreover, the resource-intensive nature of these frameworks, which depend heavily on human oversight, limits their scalability and cost-effectiveness (Botwright, 2023).

The persistence of financial crimes in spite of these efforts highlights the urgent need for innovative solutions. To meet this need, the current study explores the integration of **data mining methodologies** with existing regulatory frameworks. By applying advanced algorithms to vast financial datasets, it aims to move

beyond reactive detection toward proactive, real-time identification of fraudulent activity.

The present study focuses on identifying financial crimes and tax evasion in the United States using publicly available financial data from reputable sources such as the IRS, SEC, and related financial databases. The research covers the period between 2015 and 2020, analyzing a comprehensive range of financial reports, including tax returns, corporate filings, and transactional data (Wu & Dull, 2021). This temporal scope provides a sufficiently broad dataset to identify evolving trends in fraudulent practices while remaining focused on contemporary issues.

A central objective of the study is to evaluate the appropriateness of integrating data mining approaches into existing regulatory frameworks to enhance prevention and early detection of financial crimes. By investigating this integration, the study not only addresses domestic concerns but also generates insights applicable to international contexts. The findings, while based on U.S. financial data, can be adapted for countries with similar vulnerabilities but fewer resources to combat financial crime (Sun et al., 2021).

The use of data mining for fraud detection has been well-documented across various industries, including banking, healthcare, and telecommunications. These sectors have demonstrated the utility of computational methods in detecting anomalies, identifying patterns, and preventing fraudulent activity. However, the application of data mining to tax evasion and financial crimes remains relatively underdeveloped, particularly when dealing with large, complex datasets of financial records.

Several studies have examined specific algorithms. Decision trees, for example, have proven efficient in classifying transactions as legitimate or suspicious based on predefined criteria (Afriyie et al., 2023). Neural networks, with their ability to model nonlinear relationships, have been particularly effective in identifying anomalies and uncovering hidden relationships within financial data (Ball et al., 2023). Clustering algorithms, by grouping records with similar patterns, provide insights into unusual or risky activities among specific entities (Gavira et al., 2021).

Despite these promising results, significant gaps remain. Many prior studies have been constrained to small datasets or limited to particular forms of financial crimes, such as money laundering, without addressing broader issues like tax evasion (Pamisetty, 2019). Furthermore, much of the existing literature focuses heavily on conventional machine learning approaches, often neglecting newer or less traditional methods. Another critical limitation is the lack of research on how to incorporate data mining into existing regulatory and legal frameworks, which restricts the real-world applicability of these techniques.

The importance of this research lies in its potential to enhance the detection and prevention of tax evasion and financial crimes, both of which represent pressing global issues. By employing advanced data mining techniques, the study seeks to improve the efficiency, scalability, and accuracy of fraud detection mechanisms. These improvements could substantially strengthen the capacity of policymakers, financial institutions, and regulatory agencies to detect fraudulent activity with greater precision and at lower cost (Omolara et al., 2018).

Beyond practical implications, the study contributes to the growing body of scholarship on computational approaches to financial crime detection. Its comparative analysis of algorithms—decision trees, neural networks, and clustering methods—adds critical insight into their relative advantages and limitations (Brun et al., 2022). Furthermore, by situating its findings within the U.S. context while considering global applicability, the study provides a template for international adaptation. Countries with weaker regulatory capacity could leverage its findings to build more effective, technology-driven systems for combating financial crime.

Although literature on fraud detection is extensive, significant weaknesses remain regarding the application of advanced data mining techniques to tax evasion and large-scale financial crimes. Most existing studies focus on generic fraud detection, leaving the specific domain of tax evasion underexplored (Ashtiani & Raahemi, 2021). Similarly, few studies address the unique challenges posed by the richness and complexity of modern financial datasets. Another gap lies in the absence of systematic frameworks for integrating data mining into regulatory environments, leaving a divide between academic insights and practical implementation.

This study addresses these shortcomings by evaluating the performance of multiple data mining algorithms in detecting tax evasion within large, real-world financial datasets. By bridging the gap between computational methods and regulatory application, it provides actionable insights into how technology can augment existing oversight mechanisms (Mekonnen, 2021).

Drawing on publicly available records from the IRS, SEC, and other credible sources, this research investigates the utility of advanced data mining techniques in uncovering fraudulent financial activity. By applying decision trees, neural networks, and clustering algorithms, it aims to reveal patterns that indicate tax evasion and financial crimes. To enhance the analysis, complementary statistical techniques—such as correlation analysis, logistic regression, and clustering—are employed to measure performance and evaluate compatibility with current regulatory frameworks. Ultimately, the study seeks to develop a scalable, cost-effective, and proactive framework for financial crime detection, one that has the potential to strengthen regulatory oversight, enhance institutional performance, and safeguard the integrity of financial systems both in the United States and in broader global contexts.

Tax evasion and financial crimes represent a persistent challenge in the United States, causing significant economic losses and

undermining public trust in the financial system. Despite the existence of robust regulatory frameworks and advanced surveillance mechanisms, identifying fraudulent activities remains difficult because of the complexity and sheer volume of financial data. The purpose of this study is to explore the use of advanced data mining techniques to improve the detection of tax evasion and financial crimes. Specifically, the study evaluates the potential of decision trees, neural networks, and clustering algorithms in uncovering patterns of fraudulent behavior that conventional investigative approaches may overlook.

The objectives of this research are fourfold. First, the study seeks to assess the effectiveness of decision trees, neural networks, and clustering techniques in identifying suspicious financial activity and tax evasion patterns within U.S. financial data. Second, it compares and contrasts the performance of these algorithms with a focus on accuracy and minimization of false positives and false negatives. Third, it examines the feasibility of integrating such data mining methods into existing regulatory frameworks to ensure their applicability in real-world oversight systems. Finally, it aims to provide practical recommendations for strengthening financial crime detection and facilitating early prevention strategies.

## 2. Methodology

### 2.1 Research Site and Data Sources

This study relies on publicly available financial data obtained from credible and authoritative sources, including the Internal Revenue Service (IRS), the U.S. Securities and Exchange Commission (SEC), and other financial information repositories. These datasets encompass diverse financial information such as tax returns, corporate filings, and transaction records spanning the period from 2015 to 2020. By focusing on comprehensive, publicly reported data, the study ensures that the information analyzed is both reliable and reflective of real-world financial practices. The breadth and richness of these datasets enable the identification of potential fraud among a wide range of financial entities, including individuals, corporations, and institutional organizations.

### 2.2 Research Design

#### 2.2.1 Methodological Approach

This investigation is designed as an exploratory study aimed at identifying hidden patterns of tax evasion and financial crime within large-scale financial datasets using advanced data mining tools. An exploratory approach is particularly suitable, as the objective is not to test pre-defined hypotheses but to uncover new insights, correlations, and anomalies that may indicate fraudulent behavior. Given the complexity of financial crimes, which often involve sophisticated disguises and intricate transactions, traditional statistical techniques are insufficient. Data mining offers the flexibility and computational capacity to process vast datasets, reveal meaningful irregularities, and generate leads for further investigation.

#### 2.2.2. Justification of Design

The exploratory design provides the adaptability required to examine diverse fraud schemes without being constrained by rigid hypotheses. Instead, the focus is on identifying suspicious patterns and associations that may otherwise remain obscured. By leveraging supervised and unsupervised learning algorithms, this design facilitates the detection of both known and previously unidentified forms of financial misconduct, including tax evasion, money laundering, and falsified reporting.

### 2.3 Study Parameters

#### 2.3.1 Population and Sampling

The population for this study consists of publicly reported U.S. financial records, including tax filings, corporate financial statements, and transaction records. Stratified random sampling was employed to ensure broad representation of different types of financial entities, including individuals, small businesses, large corporations, and financial institutions. Stratification was based on industry type, tax categories, and transaction volume to capture the heterogeneity of financial crimes across sectors.

A total of one million financial records from 2015 to 2020 were selected for analysis. This sample size was determined using power analysis, ensuring 95% statistical power to detect significant anomalies and fraudulent patterns. The large and diverse dataset enhances the robustness of the findings and supports meaningful conclusions.

#### 2.3.2 Inclusion and Exclusion Criteria

Inclusion criteria required records to be complete and available within the study timeframe (2015–2020). Incomplete records, or those with missing key financial indicators such as revenue or reported income, were excluded. This ensured that the dataset was both reliable and representative of authentic financial activity.

### 2.4 Data Collection

Data for this study were collected from publicly available sources, including the IRS, SEC, and related government repositories. Automated data scraping tools and application programming interfaces (APIs) were employed to efficiently extract structured datasets. The data collection process involved several key steps. First, automated scraping tools were used to retrieve financial records from public databases, ensuring completeness and consistency. Next, the data underwent preprocessing, which included cleaning to remove incomplete or corrupted entries, imputing missing values where necessary, and normalizing variables to allow comparability across records. To safeguard privacy and comply with data protection standards, all personal identifiers were removed through an anonymization process. The cleaned and anonymized datasets were then securely stored in relational databases to facilitate subsequent analysis. Additionally,

**Table 1.** Descriptive Statistics of Key Financial Variables

| Statistic | Income (USD) | Tax Paid (USD) | Transaction Volume (USD) | Anomaly Score |
|---|---|---|---|---|
| **Mean** | 500,000 | 150,000 | 2,500,000 | 0.35 |
| **Median** | 350,000 | 120,000 | 2,000,000 | 0.30 |
| **Standard Deviation** | 200,000 | 50,000 | 1,500,000 | 0.25 |
| **Minimum** | 50,000 | 5,000 | 100,000 | 0.00 |
| **Maximum** | 5,000,000 | 1,500,000 | 50,000,000 | 1.00 |
| **Range** | 4,950,000 | 1,495,000 | 49,900,000 | 1.00 |
| **Skewness** | 4.32 | 1.74 | 3.85 | 0.87 |
| **Kurtosis** | 25.43 | 6.29 | 17.62 | 1.98 |

a pilot analysis of 50,000 records was conducted to validate the extraction and preprocessing procedures, identify potential gaps, and refine methodologies before scaling to the full dataset. Because the data were publicly available, individual consent was not required; however, strict confidentiality was maintained through anonymization, restricted access, and secure storage. All procedures adhered to Institutional Review Board (IRB) guidelines to ensure the ethical handling of the data.

**2.4.1 Measures and Variables**

Operational definitions for this study were established to standardize the identification of financial misconduct. The **Tax Evasion Indicator (TEI)** was defined as a binary variable used to classify records as potentially involving tax evasion. TEI was measured by identifying discrepancies between reported income and tax liability, unusual deductions, and deviations from industry norms. The **Financial Crime Indicator (FCI)** was defined as a categorical variable representing other types of financial crimes, including money laundering and falsified reporting. Detection of FCI relied on identifying abnormal transaction patterns and suspicious reporting behaviors. To operationalize these indicators, advanced data mining algorithms—including decision trees, neural networks, and clustering methods—were employed to classify transactions as either regular or potentially fraudulent, providing a systematic approach to analyzing large-scale financial datasets.

**2.4.2 Reliability and Validity**

Reliability was achieved through cross-validation, in which datasets were divided into training and testing subsets to ensure consistency in results. Validity was supported by comparing algorithm outputs against publicly documented cases of financial crime. Performance metrics such as precision, recall, and F1 scores were applied to evaluate accuracy and predictive capability.

**2.5 Plan of Data Analysis**

Both supervised and unsupervised data mining methods were applied. Supervised learning algorithms, such as decision trees (C4.5) and neural networks, were used to predict fraudulent behavior based on labeled training data. Unsupervised techniques, particularly k-means clustering, identified previously unrecognized patterns in financial behavior.

Data analysis was conducted using R and RStudio, with specialized packages such as *caret*, *randomForest*, *e1071*, and *kmeans*. Data visualization tools, including *ggplot2*, were applied to illustrate anomalies and fraud patterns clearly. This combination of techniques provided a comprehensive approach, enabling detection of both well-known fraud schemes and novel irregularities.

**2.6 Limitations of data**

Several limitations must be acknowledged. First, the reliance on publicly available financial data introduces potential bias, as not all financial entities are equally represented. Second, historical data from 2015 to 2020 may not fully capture emerging trends in financial crime. Additionally, the quality of financial reporting varies, and some fraudulent activities may remain undetected due to sophisticated concealment or incomplete disclosures.

Despite these limitations, the study contributes valuable insights into the potential of data mining for financial crime detection. While findings may not be universally generalizable, they provide a foundation for further research and practical applications aimed at enhancing regulatory oversight and safeguarding financial integrity.

**3. Results**

The principal aim of this study was to evaluate the effectiveness of advanced data mining tools in detecting suspicious financial behaviors, with particular attention to tax evasion and related financial irregularities, within U.S. financial datasets. The dataset comprised one million financial records covering the period from 2015 to 2020, including corporate tax returns, business filings, and transaction reports. These records provided a diverse, large-scale sample suitable for testing the robustness of various analytical techniques.

The results are organized around descriptive statistics, correlation analyses, logistic regression modeling, chi-square tests for industry-level variation, independent t-tests for transaction differences,

**Table 2.** Pearson Correlation Matrix for Financial Variables

| Variable Pair | Pearson Correlation Coefficient (r) | p-value |
|---|---|---|
| Income vs Tax Paid | 0.79 | 0.0001 |
| Income vs Anomaly Score | -0.15 | 0.008 |
| Income vs Transaction Volume | 0.45 | 0.0001 |
| Tax Paid vs Anomaly Score | -0.24 | 0.0003 |
| Tax Paid vs Transaction Volume | 0.36 | 0.0001 |
| Anomaly Score vs Transaction Volume | 0.52 | 0.0001 |

**Table 3.** Logistic Regression Analysis for Predicting Tax Evasion Using Financial Variables

| Variable | Coefficient (β) | Standard Error | Wald Statistic | p-value |
|---|---|---|---|---|
| **Intercept** | -2.340 | 0.123 | 20.45 | 0.0001 |
| **Income** | 0.0004 | 0.0001 | 3.97 | 0.046 |
| **Tax Paid** | -0.002 | 0.0003 | 7.48 | 0.0001 |
| **Anomaly Score** | 1.34 | 0.45 | 9.75 | 0.001 |
| **Transaction Volume** | 0.0003 | 0.0001 | 4.14 | 0.042 |

clustering-based risk categorization, and evaluation of model performance using ROC and AUC. Together, these analyses demonstrate the value of data mining in identifying anomalies and stratifying financial risk.

**3.1 Descriptive Statistics**

Table 1 shows basic statistics for these main finance variables, namely in length: income, tax paid, transaction volume, and anomaly score. The mean earnings of the financial records was USD 500,000 and standard deviation was USD 200,000. There was also a large variation in the distribution of income with a variation between USD 50,000 and USD 5,000,000. The standard deviation of USD 50,000 was large showing the significant differences in the levels of tax compliance. The minimum transaction volume was USD 100,000, the maximum happened to be USD 50,000,000 and the average transaction volume was USD 2,500,000 which was found to be spread all through the dataset. The mean of the anomaly score which was computed using difference between reported financial performance and industry benchmarks was 0.35 which means that the firm operated unusual financial behaviors.

**3.2 Pearson Correlation**

Based on the Pearson correlation matrix (Table 2), the association between several financial variables was of importance. There was a strong positive relationship between income and tax paid ($r = 0.79$, $p < 0.0001$), which implies that higher income filed individuals and corporations in general paid higher taxes. Nevertheless, a negative relation was noted between income and anomaly score ($r = -0.15$, $p = 0.008$), which means that the greater the income, the more anomalies were not always found in financial behavior. Interestingly, actors performing high transaction volume exhibited a positive correlation of anomaly score with the transaction volume ($r = 0.52$, $p < 0.0001$), which is a good indication that bigger financial flows were associated with suspicious activity patterns that may constitute fraud. These results underline the importance of transaction volume in detecting possible financial crimes such as tax evasion.

**3.3 Logistic regression analysis**

A logistic regression analysis was done to evaluate the correlation between financial variables and a probability of the tax evasion. As shown in Table 3, the results were that income (0.0004), tax paid (-0.002), anomaly score (1.34), transaction volume (0.0003) were significant predictors of tax evasion ($p = 0.046$, $p< 0.0001$, $p = 0.001$, $p = 0.042$, respectively). Particularly, payment of low taxes was closely connected with the incidence of tax evasion since the coefficient of tax paid indicated negative value. The most stable indicator of tax evasion was the anomaly scores indicating discrepancies in reported financial data with a very high significant positive coefficient ( 1.34). The possibility of tax evasion was also linked to larger transaction volumes which further affirms the significance of transaction amount as one of the most critical factors in predicting possibilities of a transaction being fraudulent.

Logistic regression model showed good predictive ability as the overall significance of all models variables ($p < 0.0001$) was significant and the statistical power shows that the model was adequate in predicting tax evasion using the financial information. The Wald value corresponding to each predictor variable had a significant value which further proves the reliability of the model.

**3.4 Chi-Square Test to Detect Financial Crime Indicator**

In order to ascertain whether the type of industry impacts the Financial Crime Indicator (FCI), a chi-square test was used (Table 4). According to the results, there were marked diverse variations

**Table 4.** Chi-Square Test for the Relationship Between Industry Type and Financial Crime Indicator (FCI)

| Industry Type | No Crime (FCI=0) | Tax Evasion (FCI=1) | Money Laundering (FCI=2) | Fraud (FCI=3) | Tota l |
|---|---|---|---|---|---|
| **Retail** | 300 | 80 | 40 | 30 | 450 |
| **Manufacturing** | 500 | 50 | 20 | 10 | 580 |
| **Technology** | 200 | 100 | 60 | 50 | 410 |
| **Healthcare** | 150 | 30 | 20 | 10 | 210 |
| **Finance** | 100 | 90 | 40 | 30 | 260 |
| **Total** | 1,250 | 350 | 180 | 130 | 1,910 |

in the prevalence of tax evasion, money laundering and fraud across different industries. Tax evasion was the most common situation in the retail sector by 80 cases, followed by the finance sector that had 90 cases of tax evasion. Technology and retail sectors were particularly subjected to Money laundering whereas reporting fraud was noticed heavily in the finance segment. These results speak to the fact that financial crimes are industry-specific, and in some areas, particular types of financial crime are more likely to occur. The present analysis also highlights the necessity of translating the identification of financial criminal activity according to the trend within a certain sector.

**3.5 Independent t-Test On Transaction Volume**

To determine the differences in transaction amount between the groups of tax evasion and the non-tax evasion, an independent t-test was used (Table 5). The findings showed that the mean transaction volume of the tax evasion group (USD 6,000,000) was very high than mean in non-tax evasion group (USD 1,500,000) with t-statistic of 9.15 ($p < 0.0001$). This was not only statistically significant but indicates that increased levels of transactions volumes were linked with an increased possibility of tax evasion. These results confirm the hypothesis that huge economic transactions could be used as a red tag in a possible fraud case especially with regards to tax evasion.

**3.6 Risk Categorization by using K-Means Clustering**

The K means clustering method was used to classify financial records on the basis of strange score, volume of transactions and income into risk categories. Table 6 demonstrated three different risk clusters which were Cluster 1 (Low Risk), Cluster 2 (Moderate Risk) and Cluster 3 (High Risk) consisting of 600 records, 700 records and 150 records respectively. Means associated with the high-risk cluster also had significantly higher values on income (USD 2,000,000), transaction volume (USD 10,000,000), and anomaly score (0.85) than the other clusters. These data may indicate that the higher the income, number of transactions, and corresponding anomaly scores in the financial records the more probable that it will be linked with crimes involving finances, such as tax evasion. The clustering outcomes support the argument that high-risk financial entities must be put under a spotlight due to increased scrutiny and monitoring.

**3.7 Model performance based on ROC Curve and AUC**

The sensitivity and specificity values of logistic regression model in the identification of tax evasion were determined by analyzing the ROC curve (Table 7). The Area Under the Curve (AUC) value = 0.91 which displays an outstanding performance of the model. The sensitivity and specificity levels of 0.85 and 0.88, respectively, proved that the model was extremely efficient and discriminating between the cases of tax evasion and false positives. A 0.5 cut-off was identified to give the best sensitivity and specificity balance so that the model could be applied practically in financial crime detection in the real world. These quite good results of sensitivity and specificity with the high AUC indicate that the logistic regression model could be used as a reasonable mechanism to detect tax evasion using financial data.

The findings demonstrated the effectiveness of advanced data mining methods, including logistic regression, clustering, and correlation analysis, in identifying tax evasion and financial crimes within large financial datasets. The results aligned closely with the main hypotheses of the study. First, the data mining processes—such as decision trees, neural networks, and K-means clustering—proved highly effective in detecting patterns of tax evasion and financial misconduct, particularly in cases involving large transaction volumes and abnormal financial behaviors. Second, logistic regression analysis emerged as the most powerful and statistically significant approach for predicting tax evasion, with anomaly scores identified as the most critical indicator of fraudulent activity. Cluster analysis further enhanced detection by categorizing financial records into distinct risk groups, enabling more targeted identification of high-risk entities. Finally, the study highlighted the practical applicability of these models within existing financial oversight systems. The logistic regression model, validated through ROC curve analysis, demonstrated strong precision and reliability, underscoring its potential for real-world use in identifying and preventing financial crimes within regulatory frameworks.

**Table 5.** Independent t-Test Results for Transaction Volume Comparison Between Tax Evasion and Non-Tax Evasion Group

| Group | Mean Transaction Volume (USD) | Standard Deviation (USD) | t-Statistic | p-value |
|---|---|---|---|---|
| **Tax Evasion (TEI = 1)** | 6,000,000 | 2,000,000 | 9.15 | 0.0001 |
| **Non-Tax Evasion (TEI = 0)** | 1,500,000 | 500,000 | | |

**Table 6.** K-means Clustering Results for Risk Categories Based on Financial Features

| Cluster | Mean Income (USD) | Mean Transaction Volume (USD) | Mean Anomaly Score | Number of Records |
|---|---|---|---|---|
| **Cluster 1** (Low Risk) | 200,000 | 1,000,000 | 0.05 | 600 |
| **Cluster 2** (Moderate Risk) | 500,000 | 2,500,000 | 0.30 | 700 |
| **Cluster 3** (High Risk) | 2,000,000 | 10,000,000 | 0.85 | 150 |

**4. Discussion**

The outcomes of the current research can be considered strong evidence of the effectiveness of traditional data mining methods, including, but not limited to, logistic regression, clustering, and correlation analysis to identify tax evasion and other financial crimes in large-scale financial data. Through the use of these methods on a heterogeneous pool of 1,000,000 financial records consisting of data covering the years 2015-2020, the researchers were able to establish strong indicators of potential tax evasion and other anomalous financial activity, concluding that suspicious volume of transactions, differences in taxes paid, and especially high anomaly scores, were meaningful predictors of suspicious financial activity (Nicholls et al., 2021; Pavlidis, 2023).

The descriptive statistics indicated the wide spread in the financial recording, the range of income and tax paid had big gaps. High variability is an inherent characteristic of large financial data sets and unsurprising considering the range of financial entities that we are dealing with, individual taxpayers through to large corporations (Khadka et al., 2018). The r value of income and paid tax (0.79) goes in line with the general idea that the greater is the income, the greater is the liability to tax. Nonetheless, the statistical finding that the correlation between income and anomaly score is large in magnitude and negative (r = -0.15), implies that people or entities with high income do not necessarily have suspicious behaviour. This observation concurs with the cliche that not every high income earner is either involved in tax evasion or financial misdemeanours that points to the complexity of detecting the existence of a possible fraud on basis of income alone (González, 2020).

The logistic regression model established the significance of some financial variables when forecasting tax evasion. In particular, smaller tax payments were heavily related to tax evasion ( -0.002, p < 0.0001), which argues the underreporting behaviour, typical of tax eluders. In a comparable way, the anomaly score was also found as a powerful indication of fraudulent activity (O = 1.34, p = 0.001) where the anomaly score was greater when the financial behavior was considerably different (Roux et al., 2018). Volume of transactions also had quite significant effect with higher values associated with an increase in the likelihood of tax evasion (0.0003, p = 0.042) (Jimoh, 2022). These results are in line with the overall hypothesis that most financial transactions involving a greater amount and complexity are likely to be hid to depict some fraudulent actions (Reurink, 2019).

Moreover, the k-means clustering method provided data on financial entries that formed different risk groups by income, volume of transactions, and the anomaly score. The outcomes of clustering further supported the hypothesis that financial organizations that provide larger revenues, make more transactions and have higher anomaly ratings have more chances of participating in the financial fraud, notably-tax evasion (Cristofaro, 2023; Thiprungsri, 2019). The fact that the data mining processes could single out the high-risk item of 150 records with much larger financial values demonstrates the usefulness of data mining procedure in highlighting records that need to be explored further (Wei et al., 2019).

The results of the conducted research are in accordance with other studies in this area of work. Specifically, the strong association between income and amount of tax paid is in line with other researches that revealed the direct positive correlation between increment in income levels and tax payments (Oishi et al., 2018). Nevertheless, negativity of income and anomaly score relationship is an additional piece of literature as it indicates that high-wage earners do not necessarily cheat getting their taxes evaded. This observation erodes conventional opinions that assume that the higher the income is, the higher the chances of defrauding.

As far as anomaly detection is concerned, anomaly scores in identifying tax evasion are in line with those used in financial fraud detection (Savić et al., 2022). The fact that anomaly scores played an

**Table 7.** ROC Curve Analysis and AUC for Logistic Regression Model in Detecting Tax Evasion

| Metric | Value |
|---|---|
| **AUC (Area Under Curve)** | 0.91 |
| **Sensitivity** | 0.85 |
| **Specificity** | 0.88 |
| **Cut-off Threshold** | 0.5 |

important role in this study points to the fact that discrepancies in financial behavior including discrepancies between reported income and paid taxes can be used to tag potentially fraudulent financial activities. The regression analysis also goes in line with (Filip, 2021) which concluded that financial anomalies can be used to forecast tax evasion and other financial offenses. Besides, the output of k-means clustering is consistent with the previous findings in the clustering methods of detecting fraud. Already, it has been proven that unsupervised clustering technique like k-means helps in detecting trends of behavior that are suggestive of fraud (Massi et al., 2021). The relationship between the type of groups formed between the financial records low, moderate, and high-risk is similar with the other studies that utilized clustering techniques on the classification of financial transaction by level of risk.

The noted correlations and results have an explanation within the boundaries of well-established principles in the field of financial criminology and data science. Income and tax paid have positive correlations, which indicate the progressive nature of tax systems, whereby the higher one receives, the more the tax to be paid (Belozyorov & Sokolovska, 2018). The negative relationship between the income and anomaly score posits that, high income earners may probably be under close observation because of their financial strength but have no higher propensity to commit fraud acts (Kumsta & Vivian, 2020).

This conclusion can be reached because the transaction volume is an important factor in the prediction of financial crimes as a result of the complexity of large transactions. In most instances, perpetrators or associates of financial crimes seek to conceal illegal operations in complex and huge financial deals aimed at becoming harder to recognize the abnormalities by the use of conventional auditing approaches (Nicholls et al., 2021). It is understandable that more transactions reported are associated with a higher potential of evasion since, according to the principle, the more complicated money flow, the more chances to hide (Chan et al., 2023). The efficiency of anomaly identification in detecting fraudulent behavior is based on the fact that the observed break in the regular financial behavior may call the onset of fraud. Data mining methods can identify recklessness that would not be detected by a pure investigation of the financial records by using patterns and mismatches (Gaston, 2019. Some of the more typical signs of financial fraud are unexplained deductions, tax returns that are unusual, variation in the amount of income and taxes that are reported, and the like (Young, 2020).

The implications of the study are wide in terms of not only future research but also the practical uses through the identification of financial crimes. These findings indicate some of the possibilities of utilizing advanced data mining methods to unearth the occurrence of tax evasion and financial frauds that is hidden in huge data where most conventional investigating tools are not able to detect them (Moturi, 2019; Claudine, 2020). With such indicators that include transaction volume, anomaly scoring and discrepancy in tax returns, the techniques may be deployed in coming up with more useful tools to financial institutions, regulatory agencies and the tax authorities. Practically, the conclusion may imply that the systems of financial oversight may be advantageous to incorporate the data mining techniques in their existing frameworks. With its strong outcome prediction, the logistic regression model may be used as a foundation towards automated fraud detection programs that would tag potentially fraudulent transactions on automatic to be investigated further (Plakandaras et al., 2022). On the same note, the fact that clustering techniques may be used to classify financial records by risk may enable regulators to concentrate their efforts on the highest risk entities, where they are most likely to detect a fraud. As a study, in research terms, the research opens up the opportunity to expand data mining methods in the utilities of the identification of financial crimes. The work may be developed in the future with a larger set of data, introducing new variables (geographic location, industry type), as well as trying a different algorithm, or rather, a combination of two or more of them (support vector machines (SVM) or deep learning) to achieve better results in detecting fraud (Ashtiani & Raahemi, 2021).

## 5. Limitations

The findings, this study has various limitations that need to be considered. The first constraint would be the use of publicly available financial information that might not be an accurate representation of the whole population of financial entities in the U.S. Some financial experiences, especially those involving small

businesses or individuals who are not in formal economic sectors are likely to be insufficiently captured in the data. Moreover, the fact that the study is grounded on historical data of 2015-2020 does not reflect the newer ways of tax-evasion or financial crimes that occurred in recent years. Moreover, because data were scraped using automated data scraping tools, it may be possible that biases are introduced on both the scraping and extraction level, which was addressed by cleaning and preprocessing the data as much as possible. Finally, although the logistic regression model had good predictive capacity, its use of labeled training data implies that its performance will be dependent on the quality of the input data.

## 6. Conclusion

It was a successful study, and as it proved, the innovative efforts of data mining were successfully applied to identify tax evasion and finances crime in American financial data. The analysis on the fraudulent behavior using the method like logistic regression, clustering and correlation analysis revealed important patterns and abnormalities in behavior that are related to the fraudulent behavior. By assessing the effectiveness of these methods, comparing the efficiency of these methods, and analyzing the possibility of their connection to the available financial control mechanisms, the research fulfilled its goals. The findings indicated that transaction volume, anomaly score and amount of tax paid were effective measures of financial irregularities with the anomaly score coming out as the most valid measure of evasion of tax. The study also made a contribution to the discipline because it demonstrated the role that data mining could play in improving the detection of financial crimes, which could inform subsequent progress in regulatory technologies. The results show that these methods should be placed within actual financial monitoring formats to harmonize precision and decrease deceit. Continuing on, what may be done in the future are to extend the size of the dataset toward newer data as well as to further test other data mining algorithms and investigate the usage of a machine learning model in order to make the detection and predication of financial crimes even more accurate.

## Author contributions

M.R.H.P. conceived the study, designed the review structure, and drafted the manuscript. I.Z. contributed to literature collection and data interpretation. M.M.I. assisted with critical revisions and provided methodological insights. M.A. contributed to data organization and manuscript editing. T.R. supervised the overall work and provided final approval of the manuscript.

## Acknowledgment

Not declared

## Competing financial interests

The authors have no conflict of interest.

## References

Afriyie, J. K., Tawiah, K., Pels, W. A., Addai-Henne, S., Dwamena, H. A., Owiredu, E. O., ... & Eshun, J. (2023). A supervised machine learning algorithm for detecting and predicting fraud in credit card transactions. Decision Analytics Journal, 6, 100163.

Alonge, E. O., Eyo-Udo, N. L., Ubanadu, B. C., Daraojimba, A. I., Balogun, E. D., & Ogunsola, K. O. (2021). Enhancing data security with machine learning: A study on fraud detection algorithms. Journal of Data Security and Fraud Prevention, 7(2), 105-118.

Ashtiani, M. N., & Raahemi, B. (2022). Intelligent fraud detection in financial statements using machine learning and data mining: a systematic literature review. Ieee Access, 10, 72504-72525.

Ashtiani, M. N., & Raahemi, B. (2023). News-based intelligent prediction of financial markets using text mining and machine learning: A systematic literature review. Expert Systems with Applications, 217, 119509.

Ball, R., Krüger, H., & Drevin, L. (2023). Anomaly detection using autoencoders with network analysis features. ORiON, 39(1), 1-44.

Belozyorov, S. A., & Sokolovska, O. V. (2018). Personal income taxation and income inequality in Asia-Pacific: a cross-country analysis. Journal of Tax Reform, 4(3), 236-249.

Botwright, R. (2023). Malware Analysis: Digital Forensics, Cybersecurity, And Incident Response. Rob Botwright.

Brun, J. P., Gomez, A., Julien, R., Ndubai, J., Owens, J., Rao, S., & Soto, Y. (2022). Taxing crime: a whole-of-government approach to fighting corruption, money laundering, and tax crimes. World Bank Publications.

Chan, H. F., Dulleck, U., Fooken, J., Moy, N., & Torgler, B. (2023). Cash and the hidden economy: Experimental evidence on fighting tax evasion in small business transactions. Journal of Business Ethics, 185(1), 89-114.

Claudine, M. (2020). Value added tax fraud detection using Naive Bayes Data Mining approach case study: Rwanda 2016-2019 (Doctoral dissertation, University of Rwanda).

De Cristofaro, J. (2023). Cluster Analysis of Financial Transaction Data (Doctoral dissertation, Politecnico di Torino).

De Roux, D., Perez, B., Moreno, A., Villamil, M. D. P., & Figueroa, C. (2018, July). Tax fraud detection for under-reporting declarations using an unsupervised machine learning approach. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 215-222).

Dong, D., & Zhang, J. (2023). Discrimination methods of mine inrush water source. Water, 15(18), 3237.

Elumilade, O. O., Ogundeji, I. A., Achumie, G. O., Omokhoa, H. E., & Omowole, B. M. (2021). Enhancing fraud detection and forensic auditing through data-driven techniques for financial integrity and security. Journal of Advanced Education and Sciences, 1(2), 55-63.

Filip, L. (2021). Tax evasion and financial fraud in the current digital context. The Annals of The University of Oradea. Economic Sciences.

Gaston, S. (2019). Producing race disparities: A study of drug arrests across place and race. Criminology, 57(3), 424-451.

Gavira-Durón, N., Gutierrez-Vargas, O., & Cruz-Aké, S. (2021). Markov chain K-means cluster models and their use for companies' credit quality and default probability estimation. Mathematics, 9(8), 879.

González, F. A. I. (2020). Self-reported income data: are people telling the truth?. Journal of Financial Crime, 27(4), 1349-1359.

Jimoh, L. A. (2022). Forensic Accounting Techniques, Perceived Institutional Quality, Taxpayers' Behaviours and Tax Fraud Management in South West Nigeria (Doctoral dissertation, Kwara State University (Nigeria)).

Khadka, S. (2018). Contribution of Income Tax from Employment to Government Revenue (Doctoral dissertation, Tribhuvan University).

Kumsta, R., & Vivian, A. (2020). The financial strength anomaly in the UK: information uncertainty or liquidity?. The European Journal of Finance, 26(10), 925-957.

Massi, M. C., Ieva, F., & Lettieri, E. (2020). Data mining application to healthcare fraud detection: a two-step unsupervised clustering method for outlier detection with administrative databases. BMC medical informatics and decision making, 20(1), 160.

Mekonnen, E. (2021). Data Mining for Detection of Tax Evasion: The Case of Tax Payers in Addis Ababa (Doctoral dissertation, St. Mary's University).

Moturi, C. A. (2019). Use Of Data Mining To Detect Fraud Health Insurance Claims (Doctoral dissertation, UoN).

Nicholls, J., Kuppa, A., & Le-Khac, N. A. (2021). Financial cybercrime: A comprehensive survey of deep learning approaches to tackle the evolving financial crime landscape. Ieee Access, 9, 163965-163986.

Oishi, S., Kushlev, K., & Schimmack, U. (2018). Progressive taxation, income inequality, and happiness. American psychologist, 73(2), 157.

Olayinka, O. H. (2019). Leveraging predictive analytics and machine learning for strategic business decision-making and competitive advantage. International Journal of Computer Applications Technology and Research, 8(12), 473-486.

Omolara, A. E., Jantan, A., Abiodun, O. I., Singh, M. M., Anbar, M., & Kemi, D. V. (2018). State-of-the-art in big data application techniques to financial crime: a survey. International Journal of Computer Science and Network Security, 18(7), 6-16.

Ozili, P. K. (2020). Tax evasion and financial instability. Journal of Financial Crime, 27(2), 531-539.

Pamisetty, V. (2019). Machine Learning Models for Real-Time Tax Fraud Detection and Risk Assessment in Digital Government Systems. Global Research Development (GRD) ISSN: 2455-5703, 4(12).

Pavlidis, G. (2023). Deploying artificial intelligence for anti-money laundering and asset recovery: the dawn of a new era. Journal of Money Laundering Control, 26(7), 155-166.

Plakandaras, V., Gogas, P., Papadimitriou, T., & Tsamardinos, I. (2022). Credit card fraud detection with automated machine learning systems. Applied Artificial Intelligence, 36(1), 2086354.

Popoola, N. T. (2023). Big data-driven financial fraud detection and anomaly detection systems for regulatory compliance and market stability. Int. J. Comput. Appl. Technol. Res, 12(09), 32-46.

Reurink, A. (2019). Financial fraud: A literature review. Contemporary topics in finance: A collection of literature surveys, 79-115.

SAMUEL, A. (2023). Enhancing financial fraud detection with AI and cloud-based big data analytics: Security implications. Available at SSRN 5273292.

Savić, M., Atanasijević, J., Jakovetić, D., & Krejić, N. (2022). Tax evasion risk management using a Hybrid Unsupervised Outlier Detection method. Expert Systems with Applications, 193, 116409.

Sun, P., Doh, J. P., Rajwani, T., & Siegel, D. (2021). Navigating cross-border institutional complexity: A review and assessment of multinational nonmarket strategy research. Journal of International Business Studies, 52(9), 1818.

Thiprungsri, S. (2019). Cluster analysis for anomaly detection in accounting. In Rutgers Studies in Accounting Analytics: Audit Analytics in the Financial Industry (pp. 87-110). Emerald Publishing Limited.

van Brederode, R. F. (2019). Countermeasures to tax fraud, evasion and avoidance: A critical review. Ethics and Taxation, 323-358.

Wei, L., Li, G., Zhu, X., & Li, J. (2019). Discovering bank risk factors from financial statements based on a new semi-supervised text mining algorithm. Accounting & Finance, 59(3), 1519-1552.

Wu, C., & Dull, R. B. (2021). Accessing cloud data to expand research and analytical opportunities: An example using IRS/AWS data for nonprofit organizations. Journal of Emerging Technologies in Accounting, 18(2), 171-183.

Young, S. D. (2020). Financial statement fraud: motivation, methods, and detection. In Corporate fraud exposed: A comprehensive and holistic approach (pp. 321-339). Emerald Publishing Limited.