



(RESEARCH ARTICLE)



CLOSING THE TAX GAP: AN ANALYTICS-BASED FRAMEWORK FOR FRAUD-RISK DETECTION AND AUDIT PRIORITIZATION IN STATE AND FEDERAL TAX ADMINISTRATION

Izuchukwu Shedrack Ofoma *

Tax Consultant II, Deloitte, Atlanta, Georgia, United States.

* Corresponding Author

World Journal of Advanced Research and Reviews, 2026, 31(03), 1577–1593

Publication history: Received on 02 August 2026; revised on 13 September 2026; accepted on 16 September 2026

Article DOI: <https://doi.org/10.30574/wjarr.2026.31.3.2365>

Copyright © 2026 Author(s) retain the copyright of this article. This article is published under the terms of the Creative Commons Attribution License 4.0

ABSTRACT

The persistence of the tax gap illustrates the limitations of traditional compliance monitoring in the face of large-scale electronic filing, complex business structures, cross-jurisdictional activity and rapidly evolving fraud patterns. This study develops an analytics-based framework for fraud-risk detection and audit prioritisation for state and federal tax administration. A structured evidence synthesis and design-science approach was utilised on recent tax administration reports, peer reviewed tax analytics studies and explainable artificial intelligence literature with a focus on 2024-2026 evidence. The resulting framework leverages authorised federal and state data sources, data-quality controls, rule-based screening, anomaly detection, supervised machine learning, cross-jurisdictional consistency checks, explainable risk scoring, and model governance based on human review and feedback. We propose two decision measures: A Tax Risk Score that combines predictive probability, anomaly intensity, rule violations, compliance history, and cross-source inconsistency; and an Audit Priority Score that combines risk with expected recoverable revenue and administrative cost. The synthesis shows that predictive accuracy is not enough for public tax administration: model performance should be assessed in conjunction with precision–recall behaviour, audit yield, false-positive burden, explainability, fairness, privacy and model drift. Consequently, the framework positions analytics as a decision-support mechanism rather than an autonomous means of determining fraud. It offers a practical architecture for improving case selection while protecting taxpayer rights and directing scarce audit resources toward material noncompliance. Future validation should use de-identified administrative audit data that has been legally authorised across federal and state jurisdictions.

Keywords: Tax gap; Tax fraud detection; Predictive analytics; Machine learning; Tax compliance; Risk-based auditing

1 INTRODUCTION

Tax administration is one of the pillars of public finance, because governments need to collect taxes in a timely and accurate way to support infrastructure, education, health services, public safety and other public functions. So the effectiveness of a tax system is not only a question of the statutory tax rates but also a question of the extent to which legally due liabilities are correctly reported and paid. The Internal Revenue Service (IRS) defines the tax gap as the difference between the total tax liability for a period of time and the amount voluntarily and timely paid for that period of time. The IRS estimated a gross federal tax gap of \$696 billion and a voluntary compliance rate of 85.0 percent for tax

year 2022 [1–4]. These estimates suggest that a very large absolute shortfall in revenue could coexist with a high overall compliance rate in a large economy.

The tax gap is typically considered to be made up of three broad components: nonfiling, underreporting, and underpayment. Underreporting is particularly relevant since it may be due to omitted income, overestimated deductions, incorrect business expenses, misreported credits or discrepancies between the taxpayer declaration and the information provided by the third party [1–3]. But tax-gap figures should not be automatically taken as evidence of fraud. Some of the non-compliance is attributable to errors, ambiguity, liquidity constraints and late payments. Fraud and evasion are going to need stronger indications of intentional or materially deceptive conduct. An analytics framework for public administration should therefore identify risk and evidence requiring examination rather than make an automatic legal determination that a taxpayer has committed fraud.

Traditional tax enforcement relies on combinations of statutory rules, information matching, risk filters, professional judgment and post-filing audits. These mechanisms remain necessary, but several features of contemporary tax systems create pressure for more advanced analytical support: the scale of electronic filing, the proliferation of third-party information, complex entity structures, cross-state activity, refund fraud, rapidly changing schemes and limited audit capacity. The U.S. Government Accountability Office (GAO) has documented the IRS use of automated systems in refundable-credit audit selection and emphasized that selection systems should be assessed not only for revenue outcomes but also for equity and unintended bias [5]. More recent work by the GAO [6, 7] reported that as of June 2025, the IRS had an inventory of 126 active artificial intelligence use cases, many of which were focused on operational efficiency, tax compliance and fraud detection. GAO has also pointed to the potential for AI to improve risk selection but emphasised the importance of governance, data quality, and workforce capability [8].

The international direction is identical. OECD evidence indicates that artificial intelligence has evolved from an emerging capability to a mainstream feature of tax-administration modernisation. In Tax Administration 2025, 69% of the administrations surveyed indicated using AI in 2023, and another 24% indicated plans to use AI in the future [9]. In a separate OECD study on digitalisation, more than 70% of the tax administrations surveyed were already using AI, with the most common use case being the detection of tax evasion and fraud, followed by risk assessment and virtual-assistant functions [10]. Similarly, the wider AI in the public sector literature emphasises the significance of governance, accountability, institutional capacity and responsible deployment in high-stakes administrative decisions [11, 12].

The state dimension of the problem is no less important. Annual Survey of State Government Tax Collections from the United States Census Bureau provides state-level data across broad and detailed tax categories and the Quarterly Summary of State and Local Tax Revenue provides higher frequency revenue data [13, 14]. State tax agencies also maintain data on returns, payments, withholding, sales, payroll and business registrations that can provide useful consistency signals when lawfully linked to federal or third-party information. A taxpayer might be internally consistent within a data set, but show materially inconsistent patterns across jurisdictions or information sources. This creates a strong analytical case for cross-source validation, but also creates increased privacy, security and legal-governance obligations.

Tax-risk analytics has evolved beyond deterministic rules. Recent studies have looked into machine learning, ensemble models, anomaly detection, risk classification, artificial intelligence in VAT reporting, explainable artificial intelligence and institutional implications of algorithmic auditing [20–30]. Belahouaoui and Alm [20] point out that the detection of tax frauds through AI is a growing research field. The potential of ensemble approaches is demonstrated by Alrasheedi, Ijaz, Alrashdi and Lee [21], Yang [22] compares machine-learning methods for tax compliance risk, Battaglini, Guiso, Lacava, Miller and Patacchini [23] show that machine learning can improve audit selection, and Górski and colleagues [25] directly investigate explanation techniques in the tax domain. These developments suggest that the research question has moved from can tax administrations use advanced analytical tools to how they can use it effectively, transparently and fairly.

Fraud data are also class imbalanced, selectively labelled and institutionally sensitive [31–39]. Audit labels are not a random sample of all taxpayers because the historical audit process determines which cases are audited. Naive models trained on such data can replicate selection effects, and resampling procedures can affect predictive performance and interpretability [23]. These problems are particularly important in public administration, where false positives have compliance costs, explanations may be needed for review or appeal and misapplication of personal data can damage public trust. Current XAI and financial-fraud research thus supports a combination of high-performing models with interpretable models, post-hoc explanations, explanation-quality checks and human supervision [31–39].

Therefore, the research gap is not related to the lack of AI applications in taxation. Instead, there is a need for a practical architecture that integrates state and federal tax information, rule-based controls, anomaly detection, supervised prediction, explainable risk scoring, audit capacity optimisation, and governance into a single operational workflow.

Existing research is often focused on model accuracy, specific tax types, general digital transformation or tax-specific AI applications not fully linking these layers to case prioritisation and public-sector safeguards [20–30].

This study is thus intended to develop an analytics-based framework to identify and rank potential fraud and material noncompliance in state and federal tax administration. Specific objectives are to identify major fraud-risk indicators; define a data architecture suitable for authorised state-federal risk assessment; integrate rule-based, anomaly-detection and supervised-learning approaches; propose transparent Tax Risk Score and Audit Priority Score formulations; define metrics for performance that reflect both predictive quality and audit efficiency; and embed explainability, fairness, privacy, human review and model governance into the decision process.

1.1 Research questions

- RQ1. Which financial, behavioural, reporting and cross-jurisdictional indicators are most useful for identifying potential tax fraud and material noncompliance?
- RQ2. How can authorized state and federal information be combined in a layered analytics architecture without treating algorithmic outputs as automatic fraud determinations?
- RQ3. Which combination of rule-based screening, anomaly detection and supervised learning is most appropriate for tax-risk prioritization?
- RQ4. Which metrics should be used to evaluate tax-fraud models when fraud is rare and audit capacity is constrained?
- RQ5. How should explainability, fairness, privacy, security and human review be embedded in analytics-driven tax administration?

2 MATERIAL AND METHODS

2.1 Research design

The study employed a design-science and structured evidence-synthesis approach. The aim was not to estimate the prevalence of taxpayer level fraud, nor to make performance claims based on non-accessible confidential administrative microdata, but to translate the existing tax-administration evidence and the recent analytical research into an operational framework that can be put into practice and empirically validated by authorised agencies. Design-science is appropriate because the target product is an artefact (a decision-support architecture) that is traceable to identified administrative requirements and analytical methods.

The evidence base included 50 sources. The tax-administration core focused on IRS, GAO, OECD, U.S. Census Bureau, and peer-reviewed tax-analytics publications for 2024–2026 [1–39]. Older sources were kept only if they are fundamental descriptions of methods used within the framework, such as SHAP, LIME, XGBoost, Isolation Forest, Random Forest, SMOTE and standard discrimination metrics [40–49]. OECD Tax Administration 3.0 was retained as a previous strategic reference because its natural-systems and digital-transformation logic is still directly relevant to integrated tax data architecture [50]. No confidential client, employer or taxpayer data were used in the development of this manuscript.

2.2 Evidence selection and synthesis

Sources were purposively selected for direct relevance to at least one of six framework domains: tax-gap measurement; public tax-administration digitalisation; audit selection; tax-specific machine learning; financial fraud analytics and anomaly detection; or explainability, fairness and AI governance. Official primary institutional sources and peer reviewed publications with verifiable bibliographic details were preferred. The source set comprised heterogeneous administrative reports, conceptual studies, reviews, legal instruments and empirical machine-learning applications, thus evidence was synthesised thematically rather than pooled statistically.

The synthesis also looked into what each source implies about operational tax-risk detection, what failure mode it addresses, and where it belongs in an end-to-end decision workflow.

Table 1: Evidence domains used in framework development

Evidence domain	Principal sources	Role in framework
Federal tax gap and audit administration	IRS; GAO [1–8]	Defines the compliance problem, audit-selection context and current IRS AI governance needs.
Digital tax administration	OECD [9–12, 50]	Establishes the move toward data-rich, risk-based and digitally integrated tax administration.
State-level tax information	U.S. Census Bureau [13, 14]	Supports the state dimension and availability of cross-jurisdictional tax aggregates.
Responsible AI governance	NIST, EU and U.S. Treasury [15–19]	Provides governance principles for privacy, security, accountability and risk management.
Tax-specific analytics	Peer-reviewed tax studies [20–30]	Supports model families, tax-risk applications, explainability and institutional implications.
Fraud/XAI methodology	Recent and foundational methods [31–49]	Supports imbalance handling, anomaly detection, model evaluation and explanations.

2.2.1 Inclusion and exclusion criteria

Sources were included if they satisfied at least one of the following criteria: (i) official tax-administration publications addressing the tax gap, audit selection, digitalisation or artificial intelligence; (ii) peer-reviewed studies directly addressing tax fraud, tax-risk prediction, tax auditing, fraud analytics, anomaly detection or explainable artificial intelligence; or (iii) foundational methodological studies necessary to define the analytical tools incorporated into the framework. Preference was given to publications from 2024–2026, except where an older source represented the original or widely accepted methodological reference.

Sources were excluded if they were unrelated to tax administration or fraud-risk analytics, lacked verifiable bibliographic information, duplicated previously selected evidence, consisted only of opinion without analytical relevance, or could not be linked to a specific component of the proposed framework.

2.2.2 Evidence identification

Evidence was identified from official institutional repositories and publisher-indexed academic sources. Institutional sources included the Internal Revenue Service, U.S. Government Accountability Office, Organisation for Economic Co-operation and Development, U.S. Census Bureau, National Institute of Standards and Technology, U.S. Department of the Treasury and European Union legal sources. Peer-reviewed literature was identified using combinations of terms including “tax fraud detection”, “tax compliance machine learning”, “tax audit analytics”, “tax risk prediction”, “explainable AI tax”, “financial fraud detection”, “anomaly detection”, “class imbalance” and “audit selection”. Sources were screened for relevance to the framework domains before final inclusion.

2.3 Proposed data architecture

The framework assumes that a tax authority uses only data that it is legally allowed to collect, receive, match and process. The architecture thus separates analytical usefulness from legal availability: a variable may be theoretically informative but excluded if there is no lawful authority, adequate data-quality basis or defined retention purpose. At implementation, the data catalogue has to have the source, legal basis, refresh frequency, owner, quality checks, permitted uses and controls on access for every field.

Table 2: Candidate data domains and fraud-risk indicators

Data domain	Illustrative variables	Potential analytical signal
Taxpayer/entity profile	Entity type; industry; filing status; account age; business registrations	Profile inconsistency or unusual change
Income and receipts	Wages; self-employment income; gross receipts; investment income	Underreporting or unexplained divergence
Deductions and expenses	Business expenses; charitable deductions; travel; vehicle; home-office items	Outlier ratios or unsupported claims
Credits and refunds	Credit claims; refund amount; refund frequency; refund destination	Abnormal refund intensity or repeated patterns
Payments and filing behaviour	Estimated payments; withholding; late filing; amendments; balances due	Persistent or abrupt compliance instability
Third-party information	W-2/1099-type information; payment reports; withholding records	Taxpayer–third-party mismatch
State information	State income; sales/use tax; payroll-related records; business registrations	Federal–state or state–state inconsistency
Audit/compliance history	Prior adjustments; penalties; no-change outcomes; correspondence outcomes	Recurrence, selection effects or historical risk
Relationship/network data	Shared address; bank destination; tax preparer; common ownership	Coordinated or networked anomalies

2.4 Data preparation and feature engineering

Before training models, a production pipeline should perform identity resolution, duplicate detection, missing-data checks, date harmonisation, inflation or scale adjustments where necessary, source provenance tracking and validation of categorical codes. It is important to control for leakage: fields created after the audit outcome should not be used to predict the audit outcome. Temporal validation is preferred to random-only validation when the objective is to deploy on future filing periods, because fraud patterns, policy rules and economic conditions change over time.

Feature engineering makes sense of tax records by developing interpretable risk indicators. The principal indicators proposed in this framework are the deduction-to-income ratio (DIR), expense-to-revenue ratio (ERR), effective tax rate (ETR), refund intensity (RI), income growth rate (IGR), and income mismatch ratio (IMR), as expressed in Equations (1)–(6).

$$\text{DIR} = \frac{\text{Total Deductions}}{\text{Total Income}} \quad (1)$$

where **DIR** is the Deduction-to-Income Ratio.

$$\text{ERR} = \frac{\text{Business Expenses}}{\text{Business Revenue}} \quad (2)$$

where **ERR** is the Expense-to-Revenue Ratio.

$$\text{ETR} = \frac{\text{Tax Liability}}{\text{Taxable Income}} \quad (3)$$

where **ETR** is the Effective Tax Rate.

$$\text{RI} = \frac{\text{Refund Claimed}}{\text{Reported Income}} \quad (4)$$

where **RI** is the Refund Intensity.

$$\text{IGR} = \frac{\text{Income}_t - \text{Income}_{t-1}}{\text{Income}_{t-1}} \quad (5)$$

where **IGR** is the Income Growth Rate, I_t is income in the current period, and I_{t-1} is income in the preceding period.

$$\text{IMR} = \frac{\text{Third-Party Income} - \text{Reported Income}}{\text{Third-Party Income} + E} \quad (6)$$

where **IMR** is the Income Mismatch Ratio and is a small positive constant introduced to prevent division by zero.

The small constant prevents division by zero. Cross-jurisdictional features should also be expressed as scale-adjusted differences so that large entities are not flagged merely because their absolute transaction values are high. Accordingly, the cross-jurisdictional difference may be expressed as shown in Equation (7), where the absolute difference between corresponding federal and state values is normalized by the larger reported value.

$$\text{CJD} = \frac{|X_F - X_S|}{\max(|X_F|, |X_S|) + E} \times 100 \quad (7)$$

where **CJD** is the cross-jurisdictional difference, X_F is the corresponding federal value, X_S is the corresponding state value, and E is a small positive constant used to prevent division by zero.

For implementation, zero and near-zero denominators should be handled explicitly using predefined business rules, appropriate stabilizing constants or separate indicator variables rather than unrestricted ratio calculation. Extreme ratio values should also be examined for robust transformation, truncation or winsorization during model development.

2.5 Layered analytical strategy

The proposed analytical strategy combines four complementary layers of detection. The deterministic compliance rules first detect known violations or logically impossible combinations. Second, unsupervised anomaly detection identifies observations that are outliers relative to peer observations in the absence of confirmed fraud labels. Third, supervised models estimate the probability of material noncompliance in cases where reliable historical audit labels are available. Fourth, cross-source and network analytics test if entities that appear normal within one return look suspicious when linked with third-party, state or relationship data. This layered approach is consistent with the evidence that no one algorithm is sufficient to address the heterogeneous forms of tax noncompliance [20–27, 31–39].

Table 3: Analytical layers, strengths and limitations

Layer	Illustrative technique	Primary value	Main limitation/control
Rule-based screening	Deterministic rules and thresholds	Transparent detection of known conditions	Rules age quickly; version and review them.
Anomaly detection	Isolation Forest; peer-group outlier scoring	Finds unusual records without complete labels	Unusual does not equal fraudulent; human validation required.
Supervised prediction	Logistic regression; Random Forest; XGBoost	Estimates probability of verified noncompliance	Historical labels may be selective or biased.
Cross-source consistency	State-federal and third-party reconciliation	Finds contradictions across information sources	Requires lawful linkage and data-quality controls.
Network analytics	Shared accounts, preparers, addresses or ownership	Detects coordinated patterns	Relationship alone is not evidence of wrongdoing.
Explainability	Interpretable models; SHAP/LIME-style explanations	Produces reason codes for review	Explanation fidelity must be evaluated, not assumed.

2.6 Candidate predictive models

Logistic regression is retained as a transparent statistical baseline because it is relatively easy to explain the direction and magnitude of coefficients. Decision tree models offer explicit decision paths but Random Forest can model non-linear interactions and is robust to mixed tabular features [44]. This includes gradient boosting such as XGBoost, which is known to perform well on structured data, as well as model complex interactions [42]. The framework does not

prescribe a model that is best for all possible situations; instead model selection should be guided by out-of-time performance, calibration, operational cost and explanation requirements, not just headline accuracy.

The main unsupervised component is the Isolation Forest, which isolates anomalous observations by means of random partitioning and provides a continuous anomaly score [43]. Anomaly scores should not be converted directly into fraud labels. They should instead enter the overall risk score and be assessed in conjunction with rule violations, supervised probability, taxpayer history and cross source discrepancies.

2.7 Class imbalance and selective labels

Generally, compliant observations will be far more frequent than confirmed fraud or material audit adjustments. Accuracy can therefore be misleading: a model may achieve apparently high accuracy by predicting the majority class. The framework requires stratified evaluation and prioritises precision, recall, F1, area under the precision-recall curve and rank-based audit yield [31, 33, 37, 38, 45, 46]. Resampling methods such as SMOTE can be considered during model training [45], but the effect on the stability of prediction and explanation should be evaluated as recent work demonstrated that class-imbalance countermeasures can change interpretability [31, 38].

Selective labels further complicate the problem. Historically audited taxpayers are not a random sample of all filed returns, so a supervised model may partly learn prior audit policy instead of underlying noncompliance. The proposed framework addresses this challenge through retention of unsupervised anomaly detection, explicit recording of selection policy in model documentation, performance comparisons across taxpayer strata and periodic randomised or quasi-random validation samples (where legally and operationally feasible). The goal is to decouple the detection of noncompliance from the reproduction of historical patterns of selection.

2.8 Tax Risk Score

The Tax Risk Score (TRS) is the principal decision-support measure, which is normalized to a scale of 0–100. For taxpayer i , the TRS is calculated as shown in Equation (8), subject to the weighting constraint in Equation (9). The score combines calibrated supervised probability, anomaly intensity, rule-based risk, historical compliance risk, and cross-source inconsistency. In this conceptual study the weights are not fixed. They need to be estimated or calibrated with administrative objectives, validation data and sensitivity analysis. If a jurisdiction has no lawful cross-source data, it can set $w_5 = 0$ and re-estimate the other components.

$$TRS_i = 100 (w_1 P_i + w_2 A_i + w_3 R_i + w_4 H_i + w_5 C_i) \quad (8)$$

subject to:

$$w_1 + w_2 + w_3 + w_4 + w_5 = 1 \quad [9]$$

$$0 \leq P_i, A_i, H_i, C_i \leq 1$$

$$W_j \geq 0$$

These constraints ensure that the weighted combination remains bounded and that multiplication by 100 produces a score on the intended 0–100 scale.

Where:

TRS_i = Tax Risk Score for taxpayer i , normalized to a 0–100 scale

P_i = calibrated supervised probability of material noncompliance

A_i = normalized anomaly intensity

R_i = normalized rule-based risk score

H_i = historical compliance risk

C_i = cross-source inconsistency score

w_1, w_2, w_3, w_4, w_5 = weights assigned to the respective risk components

Table 4: Tax Risk Score components

Component	Notation	Interpretation	Implementation note
Predictive probability	P	Estimated probability of material noncompliance	Calibrate on held-out/out-of-time audit outcomes.
Anomaly intensity	A	Degree of unusual behaviour relative to relevant peers	Use peer groups and temporal context.
Rule score	R	Severity/number of deterministic compliance flags	Maintain rule versioning and legal ownership.
Compliance history	H	Prior verified adjustments, repeated filing/payment issues	Avoid treating historical selection as proof of current fraud.
Cross-source inconsistency	C	Mismatch across authorized federal, state or third-party sources	Require provenance and reconciliation controls.

2.9 Audit Priority Score

Risk probability alone does not resolve the audit-allocation problem. Tax administrations face finite examiner hours and should consider materiality. An Audit Priority Score (APS) is therefore proposed as a second-stage measure.

$$APS_i = \frac{TRS_i \times ER_i}{AC_i + R} \quad (10)$$

where ER_i is the expected recoverable revenue conditional on material noncompliance being confirmed, AC_i is the estimated audit cost, and R is a positive stabilizing constant.

Defining ER_i conditionally avoids double-counting noncompliance probability, since the TRS already incorporates the supervised probability component P_i .

This formulation prevents the system from ranking only by probability and encourages a transparent distinction between “likely noncompliance” and “high expected public-value examination”. Expected recovery should itself be conservatively modelled and reviewed because aggressive revenue estimates can distort selection incentives.

2.10 Model validation and operational metrics

Predictive models should be validated using a stratified hold-out sample and, if feasible, out-of-time validation. Model selection can be performed in the development sample using five-fold cross-validation. The standard predictive performance measures include precision, recall, specificity and F1 score, as expressed in Equations (11)–(14). Tax-specific operational measures are subsequently defined in Equations (15)–(18).

$$Precision = \frac{TP}{TP + FP} \quad (11)$$

$$Recall = \frac{TP}{TP + FN} \quad (12)$$

$$Specificity = \frac{TN}{TN + FP} \quad (13)$$

$$F1 = 2 \left(\frac{Precision \times Recall}{Precision + Recall} \right) \quad (14)$$

where TP is true positive, FP is false positive, TN is true negative, and is FN false negative.

ROC-AUC is useful for ranking performance but should be accompanied by precision–recall analysis in highly imbalanced settings [46, 47].

Operational measures are also needed for tax administration. Audit hit rate is defined as audits yielding a material adjustment divided by total audits.

$$AHR = \frac{\text{Audits Yielding a Material Adjustment}}{\text{Total Audits Conducted}} \quad (15)$$

Revenue yield per audit is additional verified tax divided by audits conducted.

$$RY A = \frac{\text{Additional Verified Tax}}{\text{Total Audits Conducted}} \quad (16)$$

Revenue capture at $k\%$ measures the share of verified noncompliant tax captured when only the highest-risk $k\%$ of returns can be examined.

$$RC@k = \frac{\text{Verified Noncompliant Tax Captured in Top } k\%}{\text{Total Verified Noncompliant Tax}} \quad (17)$$

Lift@ compares the hit rate in the highest-risk with the overall base rate.

$$Lift@k = \frac{\text{Hit Rate in Top } k\%}{\text{Overall Base Hit Rate}} \quad (18)$$

False-positive burden should be reported as both a rate and an absolute count because even a modest false-positive rate can impose substantial taxpayer burden at national scale.

2.11 Explainability, fairness and governance

Tax analytics operates in a high-stakes administrative environment. The framework thus takes a human-in-the-loop approach in which the analytical system suggests cases, points to principal reasons and measures uncertainty, but an authorised official decides whether further verification, correspondence, examination or no action is appropriate. This approach is supported by recent tax and finance literature where model opacity, class imbalance and post-hoc explanation limitations can undermine otherwise strong predictive systems [25, 29, 31, 32, 36–39]. SHAP-style feature attribution and LIME-style local explanation can be useful, but they should be supplements rather than substitutes for model documentation and evidentiary review [40, 41].

Public-sector [15–19] governance controls were mapped to AI and financial-sector risk guidance. Controls required include data minimisation, role-based access, encryption, provenance, model inventories, change management, independent validation, drift monitoring, bias testing, logging, incident response and an appeal-compatible record of the factors that materially contributed to a case-selection decision. The EU AI Act is not a U.S. tax rule, but it is relevant comparative evidence of the global trend toward risk-based AI governance [17].

3 RESULTS AND DISCUSSION

3.1 Framework synthesis

This study produces an eleven stage operational architecture through the evidence synthesis. It starts with collecting authorised data and ends with feedback about the verified results of the cases. The framework is designed to separate detection from adjudication: the analytical components generate risk evidence, legal and administrative decisions are retained under human control. This is in line with GAO concerns on equity and performance measurement in automated audit selection [5–8] and with OECD evidence that AI is increasingly used to support compliance-management functions, rather than to replace them [9–12].

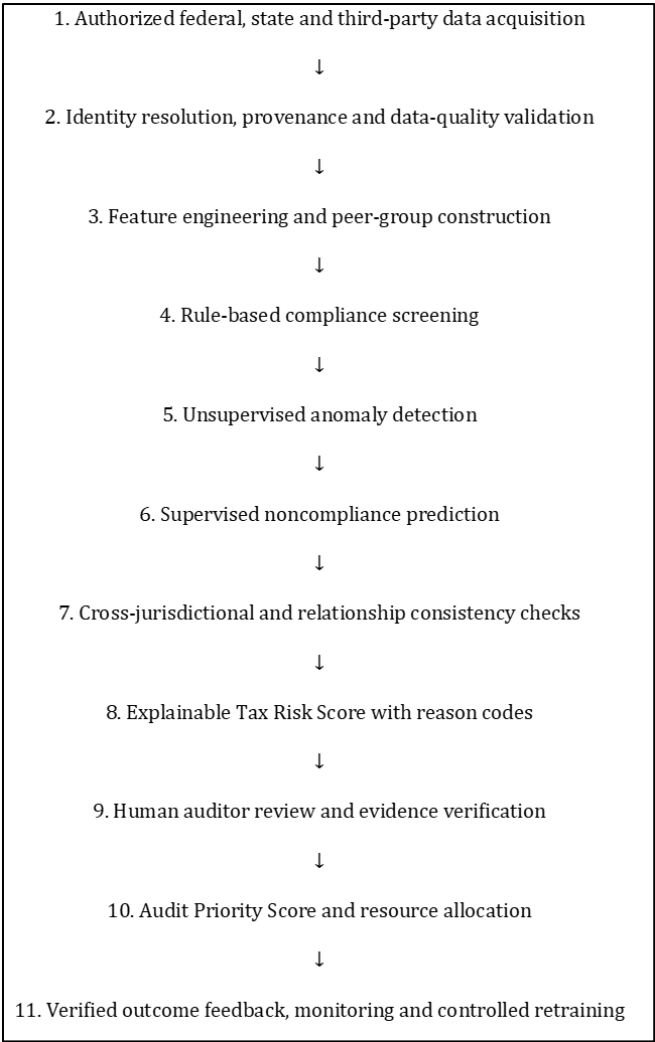


Figure 1: Proposed analytics-based tax fraud detection and audit-prioritization framework

3.2 Why a hybrid model is preferable to a single algorithm

The literature does not support tax fraud detection as a competition of classifiers. Rule systems are interpretable and legally aligned, but weak to novel schemes. Unsupervised models can uncover new patterns, but they are unable to assess intention. Supervised models use ground-truth outcomes, but suffer from label scarcity and prior selection effects. Cross-source reconciliation is effective for factual inconsistency, but relies on legal and high-quality linkage. Hybrid systems employ each method for the failure mode it solves [20–30, 35–39].

This synthesis also explains why a transparent baseline should stay in the model suite. For specific case types where explanation is the priority, a simple logistic or tree model may be preferred, but a more complex boosted model might provide a better ranking. The right comparison is not just which model has the highest ROC-AUC, but which model or ensemble offers the best combination of calibrated risk, capture of material noncompliance, stable explanation, and acceptable taxpayer burden. Similarly, recent work in finance and auditing treats interpretability as a real performance dimension, rather than a cosmetic add-on [31, 32, 36, 37].

3.3 Cross-jurisdictional analytics as a distinctive contribution

One of the key contributions of the proposed framework is the state-federal layer. Public tax data show that tax information is naturally spread across government levels [13, 14]. At the operational level risk can be enhanced by testing whether independent records provide a coherent economic story. Examples include the following: federal income reported materially differs from state taxable activity; payroll indicators inconsistent with the scale of business reported; large swings in federal income that do not correspond with state sales activity; or recurring refund patterns linked through a common account or preparer. These are signs of risk, not automatic fraud findings.

Relative and contextual measures should be used to do cross-jurisdictional analytics. A dollar difference can unfairly benefit large taxpayers. Ratios, industry-normalized residuals and time-series breaks can disentangle size from

abnormality. The system also needs to keep track of source provenance so that an auditor can determine whether an inconsistency is a timing issue, an accounting basis issue, a difference in tax definitions, a data error, or actual noncompliance. That's important because federal and state tax bases don't always align.

3.4 Risk scoring and audit-capacity optimization

The two-score architecture addresses a critical operational difference. TRS Answers: What is the strength of the evidence indicating that this return or taxpayer has material compliance risk? APS answers: what is the expected value of applying scarce audit resources to the case in light of that risk? This keeps a low-dollar, high-probability anomaly from automatically outweighing a moderately probable situation with a much larger potential fiscal impact. It also creates a transparent place to put examination cost, instead of hiding resource constraints inside an opaque classifier.

Table 5: Illustrative risk bands for operational triage

TRS range	Risk band	Default administrative response
0–29	Low	Routine processing; retain for population monitoring.
30–49	Moderate	Automated validation or low-burden consistency check.
50–69	Elevated	Request additional documentation or route to specialist review.
70–84	High	Human auditor review for examination suitability.
85–100	Critical	Priority review, subject to evidence, materiality and governance checks.

Note: Bands are proposed design thresholds only; they are not IRS or state-agency thresholds and require empirical calibration before use.

3.5 Evaluation under limited audit capacity

Tax authorities usually do not have the resources to audit all returns over a statistical threshold. This means that the evaluation must be based on the real capacity constraints. The authority should rank cases by risk for each candidate model and report the percentage of verified noncompliance and additional tax found in the top 1%, 5%, 10%, and 20% of the ranked population. It is a direct measure of how useful rankings are in practice when resources are limited. Battaglini et al. [23] provide a strong motivation to consider audit selection as a policy-allocation problem rather than a generic classification exercise.

Table 6: Recommended model evaluation dashboard

Metric	Why it matters in tax administration
Precision	Share of selected cases that are genuinely noncompliant; directly related to avoidable taxpayer burden.
Recall	Share of known noncompliance captured; reflects missed-risk exposure.
F1 score	Balances precision and recall where both false positives and false negatives matter.
PR-AUC	More informative than accuracy when the positive class is rare.
ROC-AUC	General rank-discrimination measure; useful but insufficient alone.
Calibration	Checks whether predicted probabilities correspond to observed outcome frequencies.
Hit rate	Share of audits resulting in a material verified adjustment.
Revenue yield per audit	Links case selection to fiscal productivity.
Capture@k / Lift@k	Shows how much risk is captured at realistic audit-capacity levels.
False-positive burden	Quantifies compliant taxpayers drawn into unnecessary review.
Explanation stability	Tests whether similar cases receive stable reason patterns.
Stratified performance	Detects material differences across relevant taxpayer/entity groups.

3.6 Explainability and reason codes

The XAI literature in the tax domain clearly states that explanation is not equivalent to model transparency [25, 29, 31, 32]. Even if the classifier is doing well, a post-hoc explanation can be unstable or misleading. This framework thus requires two levels of explanation. The first is model level: training population, target definition, key features, calibration, limitations and validation results. The second is case level: the individual factors that materially increased or decreased the risk score, expressed as auditor-readable reason codes. Examples include “third-party income materially exceeds reported income,” “refund intensity is extreme relative to peer group,” or “state–federal business activity is inconsistent after normalization.”

SHAP can give additive feature contributions and LIME can give local surrogate explanations [40, 41]. However, where the evidentiary or legal context requires the highest transparency, intrinsically interpretable models may be preferred. Hence the framework does not need a black-box model, it needs a demonstrably fit for purpose model whose explanations are tested for stability, relevance and non-discrimination. Transparency in AI in auditing is also emphasised by Zhong and Goel [36], and interpretability under imbalance has been shown to be important in recent finance literature [37, 38].

3.7 Fairness and taxpayer protection

Proxy variables, historical selection patterns and differential response behaviour may still cause disparities. The GAO’s audit selection analysis for refundable credits exemplifies the interplay of automated processes, outcome definitions, and nonresponse with demographic equity [5]. The proposed framework thus requires false positive rates, false negative rates, hit rates, and burden measures that are stratified by legally permitted categories of taxpayers and by geographical or organisational segments, while observing limitations on the collection and use of sensitive personal information.

Human review is also a fairness check. The auditor should be able to override a model recommendation if the rationale is weak, the data is unreliable, a potential mismatch has a legitimate accounting reason, or the case is outside the model’s validated domain. Watch out for override rates. If the override rate for a model or taxpayer segment is persistently high, that is evidence of model or data failure, not an invitation to reduce auditor discretion.

3.8 Privacy, security and institutional governance

The value of integrated tax information increases the consequence of unauthorised access or misuse of information. Thus, governance is part of model performance. NIST guidance emphasises lifecycle risk management and documented controls. Financial-sector AI guidance and emerging AI regulation emphasise security, accountability and traceability [15–19]. For tax administration, the minimum controls should include data minimisation, encryption, access segmentation, immutable logging, separation of development and production environments, third-party risk controls, model inventories, periodic recertification, and incident response.

Table 7: Governance risks and required controls

Risk	Required control
Poor data quality	Source-level validation, provenance, reconciliation and quality thresholds before scoring.
Privacy breach	Data minimization, encryption, role-based access and purpose limitation.
Unauthorized linkage	Documented legal authority and data-use agreements before cross-source matching.
Historical bias	Selection-bias analysis, stratified validation and periodic independent review.
False positives	Human verification, calibrated thresholds and burden monitoring.
Model drift	Scheduled and trigger-based drift tests; retraining only under controlled change management.
Opaque decisions	Case-level reason codes, model cards and auditable decision logs.
Overreliance on AI	Human decision authority and documented override mechanism.
Cyber compromise	Secure development lifecycle, monitoring, credential controls and incident response.
Incentive distortion	Separate risk probability from expected revenue and review both for unintended targeting effects.

3.9 Relationship to recent literature

The framework builds on previous literature in four ways. First, it considers state-federal consistency as a separate analytical tier rather than discussing tax data at only one administrative level. Second, it combines known-rule detection, unsupervised anomaly detection and supervised prediction, instead of considering one family of models as the complete solution. Third, it separates a Tax Risk Score from an Audit Priority Score explicitly linking statistical risk to materiality and cost of resources. Fourth, it embeds explainability and governance in the workflow, not only after model development. This differentiates the proposed architecture from recent tax-AI reviews and model-specific studies [20–30].

The architecture also aligns with wider evidence in fraud detection. Tree ensembles, imbalance treatments and explainability are still used for financial fraud detection in recent reviews [31–39]. Foundational methods are used to explain the technical building blocks: SHAP for additive feature attribution [40], LIME for local surrogate explanation [41], XGBoost for scalable gradient boosting [42], Isolation Forest for unsupervised anomaly detection [43], Random Forest for ensemble classification [44] and SMOTE for synthetic minority oversampling [45]. Evaluation choices are supported by precision-recall and ROC literature [46, 47]. Interpretability needs disciplined evaluation, as opposed to intuitive visualisations [48, 49].

3.10 Implementation roadmap

It has to be implemented stepwise. Phase 1 should include data governance, common taxpayer/entity key, quality controls, understandable baseline models. Phase 2 should contain anomaly detection and cross-source consistency functionalities. Phase 3: Add supervised models with higher capacity, formal explanation testing and audit-capacity optimisation. Phase 4 should ensure controlled retraining, cross-jurisdictional governance and ongoing monitoring. This sequencing reflects the logic of the OECD digital transformation, with analytics being dependent on data infrastructure, digital identity, system integration and organisational capability rather than algorithms alone [9,10,50].

The pilot should be for a narrow compliance risk for which the agency has good historical results and clear legal authority. Success criteria such as hit rate, yield, false positive burden, explanation coverage, stratified performance and examiner override should be agreed before deployment. The model should only expand its scope once it has been functioning acceptably. This reduces the risk of a great prototype becoming a black box production system with unstable data, governance, or institutional ownership.

3.11 Limitations and future validation

This study does not claim to be empirically accurate at the taxpayer level, but proposes an operational framework. There is no publicly available unconstrained national microdataset with full fraud labels, detailed return variables, and similar state-federal audit results, based on IRS and state data. The framework should be validated by an authorised tax authority on de-identified administrative data or by researchers on a legally available benchmark data set with clearly stated limitations. Synthetic data may be useful for software testing and sensitivity analysis, but they should not be presented as evidence of the prevalence of real-world fraud or the yield of enforcement.

Future research should test the framework on out-of-time audit data, measure the incremental value of state-federal features relative to federal-only or state-only models, examine network indicators for preparer or entity relationships, compare interpretable and black-box models under the same capacity constraints, and evaluate whether explanations are stable under class-imbalance treatments. Federated or privacy-preserving learning may also be explored where jurisdictions cannot pool raw taxpayer records, but such techniques still require legal authority, common data definitions and governance [34].

4 CONCLUSION

Closing the tax gap is about more than just increasing audits. Need to better identify cases where the evidence of material noncompliance is strongest, to better use the information that already exists in authorised data sources, and to implement a decision process that efficiently uses limited administrative resources. The study developed an analytics-based framework that integrates data governance, rule-based controls, anomaly detection, supervised prediction, state-federal consistency checks, explainable risk scoring, and human review and audit prioritisation into a single operational architecture.

The main contribution is to separate detection from decision. The Tax Risk Score aggregates different kinds of risk evidence. The Audit Priority Score includes expected recoverable revenue and administrative cost. They give organised evidence to authorised officials who are still responsible for verification and enforcement decisions. This design is especially important in taxation, where false positives generate real costs, historical audit flags may be arbitrary, and complex models can create opacity if explanation and governance are an afterthought.

The framework also suggests that consistency is a matter of state-federal integration rather than indiscriminate data aggregation. Cross-jurisdictional information should only be used where lawfully authorised and with controls over source provenance, normalisation and reconciliation to address differences in tax bases and timing. Analytics can identify contradictions worth paying attention to, not assume every mismatch is evasion.

For the tax administrations, controlled implementation is a practical way forward: starting from a well-defined compliance risk and a transparent baseline, testing under real-world audit capacity constraints, monitoring taxpayer burden and stratified performance, deploying advanced models only where they provide quantifiable incremental value, and maintaining a human-in-the-loop decision architecture. With the right governance, analytics can help with both revenue assurance and fairness in case selection. The social benefit is a more efficient and accountable tax administration which preserves public revenue and reduces unwarranted compliance costs on taxpayers.

STATEMENTS ON COMPLIANCE WITH ETHICAL STANDARDS

Acknowledgments

The author received no financial support for the preparation of this study. The manuscript was based on publicly available institutional information and published academic literature, and no proprietary client or taxpayer records were used.

Disclosure of conflict of interest

The author has no conflict of interest to declare regarding this manuscript. The views expressed are those of the author and do not necessarily represent the views of Deloitte or any client, tax authority or other institution. No confidential employer, client or taxpayer data was employed.

REFERENCES

- [1] Internal Revenue Service, Research, Applied Analytics & Statistics. **Tax Gap Projections for Tax Year 2022**. Publication 5869 (Rev. 10-2024). Washington, DC: Internal Revenue Service; October 2024. Available from: <https://www.irs.gov/pub/irs-pdf/p5869.pdf>
- [2] Internal Revenue Service, Research, Applied Analytics & Statistics. **Tax Gap Projections Methodology**. Publication 6031 (Rev. 10-2024). Washington, DC: Internal Revenue Service; October 2024. Available from: <https://www.irs.gov/pub/irs-pdf/p6031.pdf>
- [3] Internal Revenue Service, Research, Applied Analytics & Statistics. **Tax Gap Projections for Tax Year 2022 [Map]**. Publication 5870 (Rev. 10-2024). Washington, DC: Internal Revenue Service; October 2024. Available from: <https://www.irs.gov/pub/irs-pdf/p5870.pdf>
- [4] Internal Revenue Service. **The Tax Gap** [Internet]. Washington, DC: Internal Revenue Service [cited 2026 Aug 21]. Available from: <https://www.irs.gov/statistics/irs-the-tax-gap>
- [5] United States Government Accountability Office. **Tax Enforcement: IRS Audit Selection Processes for Returns Claiming Refundable Credits Could Better Address Equity**. GAO-24-106126. Washington, DC: U.S. Government Accountability Office; 2024. Available from: <https://www.gao.gov/products/gao-24-106126>
- [6] United States Government Accountability Office. **Artificial Intelligence: IRS Actions Needed to Address Skills Gaps, Information Quality, and Strategic Management**. GAO-26-107522. Washington, DC: U.S. Government Accountability Office; 2026. Available from: <https://www.gao.gov/products/gao-26-107522>
- [7] United States Government Accountability Office. **Supplemental Material for GAO-26-107522: IRS Artificial Intelligence Use Case Inventory as of June 2025**. GAO-26-108418. Washington, DC: U.S. Government Accountability Office; 2026. Available from: <https://www.gao.gov/products/gao-26-108418>
- [8] United States Government Accountability Office. **Artificial Intelligence May Help IRS Close the Tax Gap** [Internet]. Washington, DC: U.S. Government Accountability Office; 2024 Jun 6 [cited 2026 Aug 21]. Available from: <https://www.gao.gov/blog/artificial-intelligence-may-help-irs-close-tax-gap>
- [9] Organisation for Economic Co-operation and Development. **Tax Administration 2025: Comparative Information on OECD and Other Advanced and Emerging Economies**. Paris: OECD Publishing; 2025. doi:10.1787/cc015ce8-en.
- [10] Organisation for Economic Co-operation and Development. **Tax Administration Digitalisation and Digital Transformation Initiatives**. Paris: OECD Publishing; 2025. doi:10.1787/c076d776-en.

- [11] Organisation for Economic Co-operation and Development. **Governing with Artificial Intelligence: The State of Play and Way Forward in Core Government Functions**. Paris: OECD Publishing; 2025. doi:10.1787/795de142-en.
- [12] Stenbøg S. **Artificial Intelligence in Tax Administration: From Early Innovation to Modern Transformation** [Internet]. Paris: Organisation for Economic Co-operation and Development; 2026 Apr 29 [cited 2026 Aug 21]. Available from: <https://www.oecd.org/en/blogs/2026/04/artificial-intelligence-in-tax-administration-from-early-innovation-to-modern-transformation.html>
- [13] United States Census Bureau. **Annual Survey of State Government Tax Collections (STC)** [Internet]. Suitland, MD: U.S. Census Bureau [cited 2026 Aug 21]. Available from: <https://www.census.gov/programs-surveys/stc.html>
- [14] United States Census Bureau. **Quarterly Summary of State & Local Tax Revenue (QTAX)** [Internet]. Suitland, MD: U.S. Census Bureau [cited 2026 Aug 21]. Available from: <https://www.census.gov/programs-surveys/qtax.html>
- [15] National Institute of Standards and Technology. **Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile**. NIST AI 600-1. Gaithersburg, MD: National Institute of Standards and Technology; July 2024. doi:10.6028/NIST.AI.600-1.
- [16] National Institute of Standards and Technology. **Artificial Intelligence Risk Management Framework (AI RMF 1.0)**. NIST AI 100-1. Gaithersburg, MD: National Institute of Standards and Technology; 2023. doi:10.6028/NIST.AI.100-1.
- [17] European Parliament, Council of the European Union. **Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying Down Harmonised Rules on Artificial Intelligence and Amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act)**. Official Journal of the European Union. 2024;L 2024/1689. Available from: <https://eur-lex.europa.eu/eli/reg/2024/1689/oj>
- [18] United States Department of the Treasury. **Managing Artificial Intelligence-Specific Cybersecurity Risks in the Financial Services Sector**. Washington, DC: U.S. Department of the Treasury; March 2024. Available from: <https://home.treasury.gov/system/files/136/Managing-Artificial-Intelligence-Specific-Cybersecurity-Risks-In-The-Financial-Services-Sector.pdf>
- [19] United States Department of the Treasury. **Artificial Intelligence in Financial Services: Report on the Uses, Opportunities, and Risks of Artificial Intelligence in the Financial Services Sector**. Washington, DC: U.S. Department of the Treasury; December 2024. Available from: <https://home.treasury.gov/system/files/136/Artificial-Intelligence-in-Financial-Services.pdf>
- [20] Belahouaoui R, Alm J. Tax fraud detection using artificial intelligence-based technologies: Trends and implications. **Journal of Risk and Financial Management**. 2025;18(9):502. doi:10.3390/jrfm18090502.
- [21] Alrasheedi MA, Ijaz S, Alrashdi AM, Lee SW. Advanced tax fraud detection: A soft-voting ensemble based on GAN and encoder architecture. **Mathematics**. 2025;13(4):642. doi:10.3390/math13040642.
- [22] Yang L. Predictive modeling of tax compliance risks: A comparative study of machine learning approaches. **PLOS ONE**. 2025;20(9):e0331715. doi:10.1371/journal.pone.0331715.
- [23] Battaglini M, Guiso L, Lacava C, Miller DL, Patacchini E. Refining public policies with machine learning: The case of tax auditing. **Journal of Econometrics**. 2025;249:105847. doi:10.1016/j.jeconom.2024.105847.
- [24] Abad-Segura E, Castillo-Díaz FJ, Hernández-Pérez N, Belmonte-Ureña LJ. Technological integration and tax compliance: A quantitative evaluation of fraud detection efficiency in judicial tax proceedings in EU member States. **Empirica**. 2025;52(3):493–522. doi:10.1007/s10663-025-09654-2.
- [25] Górski Ł, Kuźniacki B, Almada M, Tyliński K, Calvo M, Asnaghi PM, Almada L, Iñiguez H, Rubianes F, Pera O, Nigrelli JI. Exploring explainable AI in the tax domain. **Artificial Intelligence and Law**. 2025;33:551–579. doi:10.1007/s10506-024-09395-w.
- [26] Al Najjar M, Mahboub R, Nakhal B, Gaber Ghanem M. Exploring the role of AI in improving VAT reporting quality: Experimental study in emerging markets. **Journal of Risk and Financial Management**. 2024;17(11):477. doi:10.3390/jrfm17110477.
- [27] Zheng Q, Xu Y, Liu H, Shi B, Wang J, Dong B. A survey of tax risk detection using data mining techniques. **Engineering**. 2024;34:43–59. doi:10.1016/j.eng.2023.07.014.

- [28] Angelov A. Artificial intelligence in European Union tax administrations: A comparative assessment. **Journal of Risk and Financial Management**. 2026;19(4):295. doi:10.3390/jrfm19040295.
- [29] Azenzoul A, Mahouat N, Vandapuye S, Nait Slimane S, Jbene M, Mokhlis K. From predictive accuracy to algorithmic justice: Mapping the multidimensional impact of AI in tax auditing. **Journal of Risk and Financial Management**. 2026;19(5):354. doi:10.3390/jrfm19050354.
- [30] Zaim H, Sahbani S. Artificial intelligence in tax compliance and evasion mitigation: Trends, mechanisms, and institutional implications. **Journal of Risk and Financial Management**. 2026;19(7):513. doi:10.3390/jrfm19070513.
- [31] Zafar U, Wu F. Methodological challenges in explainable AI for fraud detection: A systematic literature review. **Artificial Intelligence Review**. 2026;59:115. doi:10.1007/s10462-026-11516-7.
- [32] Černevičienė J, Kabašinskas A. Explainable artificial intelligence (XAI) in finance: A systematic literature review. **Artificial Intelligence Review**. 2024;57:216. doi:10.1007/s10462-024-10854-8.
- [33] Hernandez Aros L, Bustamante Molano LX, Gutierrez-Portela F, Moreno Hernandez JJ, Rodríguez Barrero MS. Financial fraud detection through the application of machine learning techniques: A literature review. **Humanities and Social Sciences Communications**. 2024;11:1130. doi:10.1057/s41599-024-03606-0.
- [34] Awosika T, Shukla RM, Pranggono B. Transparency and privacy: The role of explainable AI and federated learning in financial fraud detection. **IEEE Access**. 2024;12:64551–64560. doi:10.1109/ACCESS.2024.3394528.
- [35] Li Z, Zhu Y, van Leeuwen M. A survey on explainable anomaly detection. **ACM Transactions on Knowledge Discovery from Data**. 2023;18(1):Article 23, 1–54. doi:10.1145/3609333.
- [36] Zhong C, Goel S. Transparent AI in auditing through explainable AI. **Current Issues in Auditing**. 2024;18(2):A1–A14. doi:10.2308/CIIA-2023-009.
- [37] Chen Y, Calabrese R, Martin-Barragan B. Interpretable machine learning for imbalanced credit scoring datasets. **European Journal of Operational Research**. 2024;312(1):357–372. doi:10.1016/j.ejor.2023.06.036.
- [38] Cemernek D, Siddiqi S, Kern R. Effects of class imbalance countermeasures on interpretability. **IEEE Access**. 2024;12:45342–45358. doi:10.1109/ACCESS.2024.3381536.
- [39] Hasan M, Rahman MS, Janicke H, Sarker IH. Detecting anomalies in blockchain transactions using machine learning classifiers and explainability analysis. **Blockchain: Research and Applications**. 2024;5(3):100207. doi:10.1016/j.bcra.2024.100207.
- [40] Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Guyon I, von Luxburg U, Bengio S, Wallach H, Fergus R, Vishwanathan SVN, Garnett R, editors. **Advances in Neural Information Processing Systems 30**. Red Hook, NY: Curran Associates, Inc.; 2017. p. 4765–4774.
- [41] Ribeiro MT, Singh S, Guestrin C. Why should I trust you?: Explaining the predictions of any classifier. In: **Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. New York, NY: Association for Computing Machinery; 2016. p. 1135–1144. doi:10.1145/2939672.2939778.
- [42] Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: **Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining**. New York, NY: Association for Computing Machinery; 2016. p. 785–794. doi:10.1145/2939672.2939785.
- [43] Liu FT, Ting KM, Zhou ZH. Isolation Forest. In: **2008 Eighth IEEE International Conference on Data Mining**. Piscataway, NJ: IEEE; 2008. p. 413–422. doi:10.1109/ICDM.2008.17.
- [44] Breiman L. Random Forests. **Machine Learning**. 2001;45:5–32. doi:10.1023/A:1010933404324.
- [45] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. **Journal of Artificial Intelligence Research**. 2002;16:321–357. doi:10.1613/jair.953.
- [46] Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. **PLOS ONE**. 2015;10(3):e0118432. doi:10.1371/journal.pone.0118432.
- [47] Fawcett T. An introduction to ROC analysis. **Pattern Recognition Letters**. 2006;27(8):861–874. doi:10.1016/j.patrec.2005.10.010.
- [48] Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. **arXiv [Preprint]**. 2017. arXiv:1702.08608. Available from: <https://arxiv.org/abs/1702.08608>

- [49] Molnar C. **Interpretable Machine Learning: A Guide for Making Black Box Models Explainable**. 2nd ed. Munich, Germany: Christoph Molnar; 2022. Available from: <https://christophm.github.io/interpretable-ml-book/>
- [50] Organisation for Economic Co-operation and Development. **Tax Administration 3.0: The Digital Transformation of Tax Administration**. Paris: OECD Publishing; 2020. doi:10.1787/ca274cc5-en.

AUTHOR'S SHORT BIOGRAPHY

Izuchukwu Shedrack Ofoma is a tax and accounting professional whose work spans state and local tax compliance, financial analysis, audit, internal control and fraud-risk assessment. He holds a Master of Science in Finance from the University of Utah, a Bachelor of Science in Accountancy from the University of Nigeria, Nsukka, and a Higher National Diploma in Accountancy from Federal Polytechnic Nekede, Owerri. His professional credentials include Certified Public Accountant (CPA), Certified Management Accountant (CMA) and Associate Chartered Accountant (ACA). His research interests include tax analytics, financial fraud detection, internal controls, financial technology and data-driven decision support.