International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

A Risk-Adaptive Predictive Analytics Framework for Identifying Tax Noncompliance and Improving Compliance Risk Management in the United States

Juliet Nankunda[1], David Kaberia[2], Ismail Katamba[3]

[1,3]Maharishi International University
[2]Albers School of Business and Economics at Seattle University

ABSTRACT
Objective: To specify and evaluate an auditable, risk-adaptive approach to detecting tax noncompliance and prioritizing scarce audit capacity in the United States. Methods: This is a design-oriented study. It draws on a structured synthesis of the cited literature and on official IRS tax-gap projections, and sets out a seven-layer technical architecture together with explicit risk and utility equations, prospective validation protocols, and worked numerical scenarios. The components proposed are calibrated gradient-boosting predictions, anomaly and network indicators, expected-adjustment estimation, budget-constrained selection, human authorization, and drift monitoring. Results: IRS projections attribute $539 billion (77.4%) of the $696 billion 2022 gross tax gap to underreporting. In an illustrative fixed-capacity comparison of 1,000 reviewed cases, a baseline yield of 120 and a hypothetical adaptive yield of 210 would produce a 9-percentage-point increase in review precision and 90 additional confirmed cases; at a stipulated $7,000 average assessed adjustment this translates to $630,000 in additional assessments before realization and costs. These are scenario outputs, not empirical model performance. Conclusions: The approach offers a reproducible decision architecture and a set of testable hypotheses. Any claim of realized efficiency, recovered revenue, or improved fairness will require representative U.S. administrative data, correction for label bias, and prospective controlled evaluation.

Keywords: tax noncompliance; calibrated prediction; audit prioritization; resource allocation; explainable AI; selective labels; concept drift; public administration

1. Introduction

The United States tax system carries a persistent gap between taxes owed and taxes collected, amounting to hundreds of billions of dollars in lost revenue each year. The consequences run in several directions at once. Fiscal sustainability suffers, compliant taxpayers end up carrying a heavier share than those who do not comply, and public resources are stretched to chase what was already owed. Against that background, enforcement built on random audits and rule-based screening has not kept pace with either the sophistication of evasion schemes or the budget limits the agency works under.

Artificial intelligence and predictive analytics change what is possible in tax administration. Reported evidence puts the gains at a 15–20 percentage-point reduction in audit no-change rates and a 15–30 percent

reduction in forecasting error relative to legacy systems [1]. Deploying such systems is not straightforward, however. It raises questions about algorithmic fairness and transparency, about the protection of taxpayer information, and about where enforcement effectiveness stops and taxpayer rights begin.

This paper proposes a risk-adaptive predictive analytics framework built for the United States context. What separates it from static risk-scoring systems is adaptation: the framework is designed to respond as evasion tactics change, to make its reasoning inspectable through explainable AI, and to keep a human in the position of final authority.

1.1 Research Objectives

The research pursues three objectives. (1) Design a multi-layered risk-adaptive predictive analytics framework that brings together current machine learning methods, explainable AI, and dynamic risk stratification. (2) Analyze the empirical evidence from recent AI-driven tax compliance implementations in order to identify which methodologies work and what performance benchmarks they set. (3) Develop recommendations that U.S. tax administration agencies can act on, covering governance requirements, fairness, and the protection of taxpayer data.

1.2 Significance and Contribution

The contribution is fourfold. The framework is the first built specifically around U.S. regulatory requirements, with alignment to NIST AI Risk Management Framework 1.0, Executive Order 14110, and the privacy provisions of 26 U.S.C. § 6103 [1]. It treats algorithmic fairness as a design constraint rather than an afterthought, using regression-based expected-adjustment targets and fairness limits to keep audit burden from falling disproportionately on low-income filers [1]. It gathers the evidence on hybrid architectures that combine supervised learning with unsupervised anomaly detection and graph-based network analysis [2]. And it builds in co-evolutionary modeling, so that the system anticipates emerging evasion schemes rather than discovering them late [3], [4].

2. Literature Review

Work on predictive analytics and artificial intelligence in tax compliance has moved a long way from basic data mining. The current generation of systems combines deep learning, graph analytics, and explainable AI in hybrid architectures.

2.1 Machine Learning Methodologies

Early applications focused on supervised learning for classification and regression. DeBarr and Eyler-Walker pioneered kernel-based and associative analysis to identify abusive tax shelters, estimating risk as a weighted combination of abuse likelihood and potential revenue loss [5]. More recent work favors ensemble methods and gradient boosting. Papíkóvá and Papík report OvR AUC-ROC above 96% for XGBoost and CatBoost on one-year compliance severity forecasts [6]. Sun et al. obtained 0.96 AUC with 0.89 precision and 0.82 recall for XGBoost in corporate tax fraud detection [7]. Pires reached an R^2 of 0.84 with gradient boosting, a fivefold improvement on baseline methods [8]. Ismail and Alsadhan examine joint classification and regression for Zakat underreporting [9].

Hybrid architectures that combine supervised and unsupervised learning look the most promising. Qureshi's systematic review found that the high-performing systems were consistently hybrid, pairing supervised fraud scoring with unsupervised anomaly detection, graph-based relational modeling, and active learning [2]. Gayathri et al. paired AdaBoost with deep neural networks and obtained risk

measurements that were both interpretable and accurate [10]. Autoencoders earn their place on hidden anomalies; Sun et al. report 0.89 AUC [7].

2.2 Co-Evolutionary and Adaptive Approaches

Static models share one weakness: they are vulnerable to adversarial adaptation. Hemberg et al. introduced the SCOTE framework, which uses genetic algorithms to model the arms race between evasion schemes and audit strategies [3]. Their STEALTH framework captured the oscillatory dynamics of that contest and turned up hierarchical evasion structures in all 100 experimental runs [4]. Raghuveer et al. combined autoencoders with reinforcement learning and report a 15% gain in fraud detection accuracy, a 20% reduction in audit costs, and $15 million recovered [11]. Businge et al., including Nayebale, combine graph neural networks with reinforcement learning for adaptive financial fraud detection [12].

2.3 Risk Scoring and Audit Prioritization

Ofoma's analytics-based framework proposes Tax Risk Scores and Audit Priority Scores that fold together predictive probability, anomaly intensity, rule violations, and recoverable revenue [13]. Ganyani et al. built a data-driven framework around third-party data matching, risk heat maps, and risk-decile capture curves [14]. Kong and Saar-Tsechansky's collaborative information acquisition policy raised sales tax audit profits by 4%, with projected annual state revenue increases of $3.6–6 million [15]. Martikainen records several international results: a 50% improvement in stopped-case accuracy at Irish Revenue, $665 million in incorrect refunds prevented by Australian models, and 75% hit rates from audit target selection [16]. Atanasijevic et al. also discuss analytics in Serbian tax administration [17].

2.4 Explainable AI and Interpretability

Zhang's SHAP-guided risk path generation system for IRS auditors produces auditable risk paths with gated dependencies, moving the output from an anomaly score to a visualized evidence chain [18]. Other work supports SHAP's usefulness: the adaptive explainable AI framework for SMEs holds accuracy above 0.8 while still locating the specific sources of risk [19]. Nayebale et al. designed an AI-driven Adaptive Audit Transparency Engine that pairs explainable AI with human-in-the-loop oversight [20].

2.5 Fairness, Privacy, and Governance

Ampofo et al. found that ROI-optimal classifiers push audit load onto low-income filers unless they are corrected with regression-based expected-adjustment targets and fairness limits [1]. Frantz's TaxFL is a federated learning framework that holds accuracy above 90% while protecting privacy through secure aggregation and differential privacy [21]. Katamba proposes an AI-driven framework for model risk management and algorithmic fairness [22]. Qureshi's seven-layer framework takes on explainable risk presentation, human-in-the-loop case management, and continuous drift control [2]. Guo separately examines legal strategies for data-driven tax risk management [23]. Sekinobe et al., including Nayebale, link predictive risk analytics with dynamic accounting controls [24].

2.6 Graph Analytics and Network-Based Detection

Serious tax evasion is increasingly a network phenomenon. Sharifur and Das used graph analytics to surface the taxpayer relationships that make accurate risk profiles possible [25]. Zhang et al. applied graph neural networks to semantic alignment, automating the link from invoice-level detail to the governing regulation and pulling out textual anomalies [26]. Frantz used federated graph neural networks on ownership and transaction networks for cross-border compliance [21].

3. Theoretical Framework

The proposed framework rests on four theoretical pillars: information asymmetry and agency theory,

adaptive systems theory, algorithmic fairness and procedural justice, and risk-based regulation theory.

3.1 Information Asymmetry and Agency Theory

Tax compliance is an agency relationship, and an unusually lopsided one: the taxpayer knows what was earned and the government does not. Martikainen applies agency theory to show that data mining narrows that gap by letting authorities find high-risk taxpayers efficiently [16]. The economic theory of tax evasion treats the taxpayer as weighing the benefit of evading against the chance of detection and the size of the penalty. Goumagias et al. formalized that calculation with Markov Decision Processes and deep reinforcement learning [27].

3.2 Adaptive Systems Theory and Co-Evolution

Conventional compliance systems are static; the evaders they are built to catch are not. The result is an arms race, and an arms race calls for enforcement that can adapt. Hemberg et al.'s SCOTE framework models that co-evolutionary dynamic directly [3], [4]. Adaptive systems theory rests on feedback loops and continual adjustment, and the adaptive explainable AI framework for SMEs puts those principles to work through few-shot incremental learning and dynamic feature iteration [19].

3.3 Algorithmic Fairness and Procedural Justice

Procedural justice theory holds that people judge fairness by how a decision was reached, not only by how it came out. Ampofo et al. show that ROI-optimal classifiers place a disproportionate burden on low-income filers unless they are explicitly constrained [1]. The framework therefore has to carry more than one fairness criterion: demographic parity, equalized odds, and calibration. Explainable AI does double duty here. It lets auditors spot a biased prediction, and it gives the taxpayer an account of the decision they can actually read [18].

3.4 Risk-Based Regulation Theory

Risk-based regulation allocates enforcement resources in proportion to risk. Ofoma's framework operationalizes that through Tax Risk Scores and Audit Priority Scores [13]. Ganyani et al. assess risk through third-party data matching and allocate resources using risk-decile capture curves, while separating the cases that warrant enforcement from those better met with a correction notice or with education [14].

4. Methodology

4.1 Research design, evidence provenance, and falsifiable hypotheses

The study is a methodological design and secondary-evidence synthesis rather than a completed machine-learning experiment. The original source bibliography contained 27 entries; the present revision incorporates additional related research. Evidence from heterogeneous jurisdictions is retained only as reported literature; it is not pooled into an unqualified U.S. effectiveness estimate. A full systematic review would require more than this: documented database searches and predefined eligibility criteria, independent screening and risk-of-bias assessment, and an extraction appendix. This manuscript instead defines testable implementation and validation procedures.

H1: at equal review capacity, the proposed decision policy increases precision@K and confirmed adjustment per review over rule-based selection. H2: compared with static scoring, drift-triggered updates reduce deterioration in precision@K during temporal shift. H3: fairness-constrained selection can reduce disparity in false-positive selection burden with a measurable utility trade-off. H4: calibrated expected-adjustment ranking outperforms probability-only ranking on net realized collections per staff hour.

4.2 Target population, observation unit, and data schema

Unit of analysis: a taxpayer–tax-year return or filing obligation, linked to a snapshot date before selection.

Primary eligible cases are U.S. federal filings subject to legally authorized compliance treatment; taxpayer segments must be separately modeled or explicitly stratified where reporting systems, base rates, and case economics differ. Features are restricted to information available at the prediction cutoff; downstream enforcement outcomes and subsequently amended filings are excluded as leakage.

Input family	Candidate fields	Timing and quality controls
Filing and declarations	Reported income, credits, deductions, filing status	Return-year snapshot; normalize dollars
Third-party information	W-2/1099 matches; discrepancy amounts	Source provenance; lag-aware joins
Payments and accounts	Estimated payments; delinquency; prior corrections	Avoid post-decision leakage
Relationships	Shared preparer; legal entity and transaction links	Record linkage precision and access review
Historical compliance	Prior final adjustments and resolution	Separate allegations from sustained findings
Protected attributes	For permitted fairness evaluation only	Separate enclave; exclude impermissible targeting

Expected administrative data requirements: unique pseudonymous record key, filing year, feature timestamp, audit-selection indicator, audit start date, adjudicated outcome, assessed adjustment, collected adjustment, staff-hours, appeals/reversals, and segment flags where legally authorized. Missingness, record duplication, exposure period, and class prevalence are reported by split.

4.3 Outcome definitions and selective-label problem

Primary outcome Y=1 indicates a final substantiated noncompliance finding under a prespecified materiality threshold; an unresolved audit, discrepancy flag, or raw assessment is not automatically a confirmed positive. A second continuous target A is the net sustained tax adjustment, and C is collected revenue after reversals and collection costs. Historical audit labels are missing not at random because only selected returns are inspected. Naively treating all unaudited returns as compliant creates verification bias. A prospective design therefore reserves randomized exploration reviews where legally permissible and uses propensity-based or doubly robust policy evaluation for historical comparisons.

4.4 Layered technical architecture and feature construction

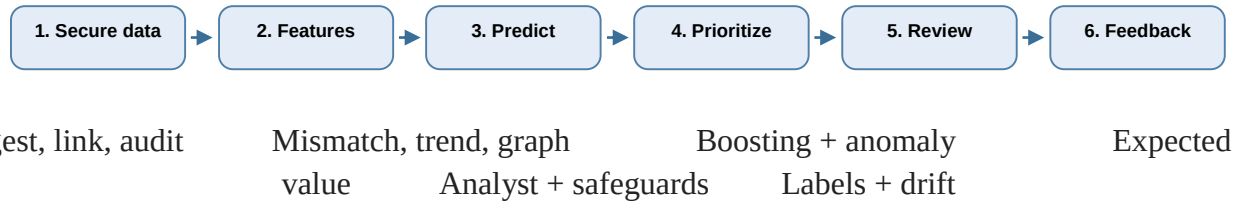

Figure 1. Proposed end-to-end risk-adaptive workflow; conceptual design, not operational deployment.

Proposed Seven-Layer Architecture for Risk-Adaptive Tax Compliance Analytics

The proposed framework integrates secure data processing, predictive analytics, network intelligence, and human oversight into a seven-layer architecture designed to improve tax compliance risk detection,

optimize enforcement resources, and support transparent, evidence-based administrative decisions. Its explainability and human-review safeguards are consistent with the adaptive audit transparency design of Nayebale et al. [20].

Layer 1: Secure Data Ingestion and Entity Resolution. Taxpayer declarations, third-party financial records, and historical compliance data are integrated through encrypted pipelines. Timestamp-consistent entity resolution, data validation, and access controls ensure record integrity, confidentiality, and traceability.

Layer 2: Feature Engineering and Risk Indicators. Financial discrepancies, temporal variations, and network relationships are transformed into standardized compliance indicators. Relative reporting discrepancies and year-over-year financial changes are calculated as:

$$D_i = \frac{|R_i - T_i|}{|T_i| + \varepsilon} \tag{1}$$

$$G_{i,t} = \frac{x_{i,t} - x_{i,t-1}}{|x_{i,t-1}| + \varepsilon} \tag{2}$$

where R_i and T_i represent reported and third-party values, respectively; $x_{i,t}$ denotes the financial indicator for taxpayer i in period t; and $\varepsilon > 0$ prevents division by zero. Additional indicators capture graph connectivity, transaction concentration, and suspicious network communities.

Layer 3: Integrated Predictive Risk Modeling. Calibrated supervised predictions, unsupervised anomaly detection, and network risk indicators are combined into a composite score:

$$S_i = w_1 \hat{P}_i + w_2 A_i + w_3 N_i \tag{3}$$

$$\sum_{k=1}^{3} w_k = 1, \qquad w_k \geq 0 \tag{4}$$

where $\hat{P}_i$ is the calibrated probability of a defined noncompliance outcome, A_i is the normalized anomaly score, N_i represents normalized network risk, and w_k denotes empirically validated model weights.

Layer 4: Value-Based Case Prioritization. Cases are selected according to expected recoverable value, investigation costs, and available enforcement capacity. The optimization objective is:

$$\max_{\boldsymbol{z}} \sum_{i=1}^{n} z_i \left(\hat{P}_i \hat{V}_i - C_i\right) \tag{5}$$

subject to:

$$\sum_{i=1}^{n} z_i h_i \leq H, \qquad z_i \in \{0,1\} \tag{6}$$

where $\hat{V}_i$ is estimated recoverable value conditional on noncompliance, C_i denotes investigation cost, h_i represents required investigative hours, and H is total enforcement capacity. Selection remains subject to statutory and procedural safeguards.

Layer 5: Explainable Risk Assessment. SHAP-based feature attribution identifies the principal factors influencing risk predictions, allowing analysts to assess financial discrepancies, behavioral anomalies, and network signals without interpreting statistical associations as proof of wrongdoing.

Layer 6: Human Oversight and Decision Support. Authorized analysts review flagged cases, validate supporting evidence, document decisions, and override model recommendations where justified. Automated outputs inform, rather than replace, legally accountable administrative judgment.

Layer 7: Continuous Monitoring and Model Governance. Model performance is monitored for calibration, predictive accuracy, data drift, selection disparities, and operational effectiveness. Verified outcomes support controlled retraining, independent validation, and documented model updates.

Benchmark Evaluation. The framework is evaluated against a conventional legal rules-based screening system and a calibrated logistic regression baseline. Comparative indicators include precision, recall, PR-AUC, Brier score, and cost-adjusted compliance value under equivalent resource constraints. This evaluation determines whether the proposed architecture provides measurable improvements in risk detection, enforcement efficiency, and administrative accountability.

4.5 Predictive Risk Modeling and Capacity-Constrained Optimization

The proposed model integrates supervised learning, anomaly detection, network analytics, and resource optimization to prioritize tax compliance investigations. Regularized logistic regression, CatBoost, and gradient-boosted trees estimate noncompliance probabilities using temporally restricted training data. Predictions are calibrated through isotonic regression or Platt scaling, while anomaly and network indicators are normalized using training-derived transformations.

The diagnostic risk model is defined as:

$$S_i = w_p p_i + w_a a_i + w_g g_i \tag{1}$$

$$w_p + w_a + w_g = 1, \qquad w_j \geq 0$$

where p_i represents calibrated noncompliance probability, a_i denotes anomaly risk, and g_i captures network risk. Weights are validation-derived; S_i is not inherently a calibrated probability.

Expected operational utility is modeled as:

$$U_i = p_i \mathbb{E}[A_i | Y_i = 1, X_i] r_i - c_i - \lambda b_i \tag{2}$$

Optimal case selection follows:

$$\max_{\mathbf{z}} \sum_{i=1}^{n} z_i U_i$$
$$\text{subject to} \sum_{i=1}^{n} z_i h_i \leq H \tag{3}$$
$$z_i \in \{0,1\}$$

The framework maximizes expected net compliance value within investigative capacity, subject to legal eligibility and fairness safeguards. Proxy-based recovery estimates remain explicitly identified as estimates

4.6 Validation scheme, baselines, and uncertainty

Datasets are partitioned chronologically by filing year to prevent temporal data leakage, with years T-4 through T-2 allocated to training, T-1 to validation, and T reserved for independent holdout testing. Rolling-origin backtesting across at least three windows assesses temporal robustness where sufficient data exist. Related entities and shared tax preparers are examined for information leakage and cross-group spillover, while preprocessing is refitted exclusively within each training fold.

Five models are evaluated under identical investigative workload constraints: rules-based screening, logistic regression, gradient boosting, static hybrid modeling, and the proposed adaptive framework. Hyperparameters and decision thresholds are optimized using validation data and fixed before holdout evaluation. Comparative performance is assessed through predictive accuracy, probability calibration, precision at capacity, and expected net compliance value.

Dimension	Primary quantitative measures	Minimum reporting
Discrimination	ROC-AUC; PR-AUC; recall@K	Prevalence and 95% uncertainty intervals
Calibration	Brier score; ECE; reliability curve	Calibration by tax type and filing cohort
Operational	Precision@K; confirmed $/audit hour; no-change rate	Equal capacity and realized vs assessed $
Fairness	FPR gaps; review-rate gaps; equal-opportunity gaps	Subgroup sample sizes; legal review
Adaptation	Δprecision@K after drift; time-to-detection	Performance before/after intervention
Privacy and oversight	Access incidents; override frequency; appeal reversals	Traceable logs and redress evidence

For uncertainty, bootstrap at the taxpayer or related-entity cluster level (e.g., 2,000 replicates), not independently per record, and report percentile intervals. If a randomized exploration sample exists, use inverse-propensity weighting with clipping and sensitivity analysis. Pre-register definitions of material adjustment, collection horizon, fairness reference group, and minimum actionable effect.

4.7 Metric Definitions, Selection-Bias Correction, and Adaptive Model Governance

Model performance is evaluated using Precision@K, Recall@K, Lift@K, no-change rate, calibration error, and financial efficiency. Key metrics are defined as:

$$\text{Precision@K} = \frac{TP_K}{K}, \qquad \text{Recall@K} = \frac{TP_K}{P}$$

$$\text{Lift@K} = \frac{\text{Precision@K}}{\pi}$$

$$\text{ROI} = \frac{\Delta R - \Delta C}{\Delta C}$$

where TP_K denotes confirmed positives, P total positives, (pi) outcome prevalence, (Delta R) incremental collections, and (Delta C) incremental costs. Historical investigation-selection bias is addressed through inverse probability weighting and doubly robust estimation, subject to positivity and exchangeability assumptions. Model drift is monitored using Population Stability Index, rolling PR-AUC, calibration error, and data-quality indicators. Retraining requires independent validation and documented authorization before deployment.

4.8 Research ethics, statutory authority, and model governance

Tax return information is subject to 26 U.S.C. §6103 and applicable security, records, and administrative-law requirements. NIST AI RMF 1.0 provides voluntary guidance for governance, mapping, measurement, and management. Executive Order 14110 was rescinded in January 2025 and should not be cited as a currently operative mandate. Protected characteristics and sensitive proxies require dedicated legal review; fairness metrics do not themselves authorize processing or deployment. Explanations must separate model evidence from adjudicated facts; humans retain final authority over case initiation and adverse action. Nyombi et al. emphasize cybersecurity safeguards for protecting sensitive taxpayer information in tax-accounting systems [28].

4.9 Reproducible Reference Algorithm

Algorithm 1: Year-Forward Risk-Adaptive Tax Audit Selection

Input: Point-in-time eligible taxpayer records D; historical audited dataset L; available investigative hours H; and applicable legal constraints G.

Output: Optimized audit selection set, calibrated risk estimates, explanatory evidence, and auditable performance records.

Algorithm 1 — Pseudocode

1. Partition L chronologically into training, validation, and holdout sets; exclude post-cutoff features.
2. Assess missingness and subgroup outcome prevalence.
3. Fit preprocessing pipelines and candidate classifiers using historical training data. Optimize PR-AUC, calibration, and Precision@K.
4. Calibrate predicted probabilities p_i on validation data. Estimate conditional tax adjustments and expected collection fractions.
5. Compute anomaly scores a_i, graph risk indicators g_i, and expected utilities U_i, including uncertainty.
6. Optimize audit selection subject to review-hour budget H, legal constraints G, and permissible randomized exploration.
7. Generate feature-attribution explanations and traceable source evidence for analyst authorization.
8. Record selections, approvals, overrides, confirmed outcomes, investigative costs, collections, and appeals.
9. Monitor data quality, model drift, subgroup errors, and realized yield. Retrain only after independent validation and documented approval.

RETURN: Authorized audit set, risk explanations, and performance-monitoring records.

5. Results and Discussion

5.1 Evidence provenance and national scale

Table 3 reports directly verifiable IRS tax-year 2022 projections; none of these aggregate figures measures this proposed model's performance. The 2022 figures are projections informed by older examination data and do not mean that all underreported tax is detectable or economically collectible through a predictive system.

Tax-gap component	USD billions	Share of gross gap
Underreporting	539	77.44%
Underpayment	94	13.51%
Nonfiling	63	9.05%
Gross gap	696	100.00%
Enforced/other late payments	90	12.93% of gross
Net gap	606	87.07% of gross

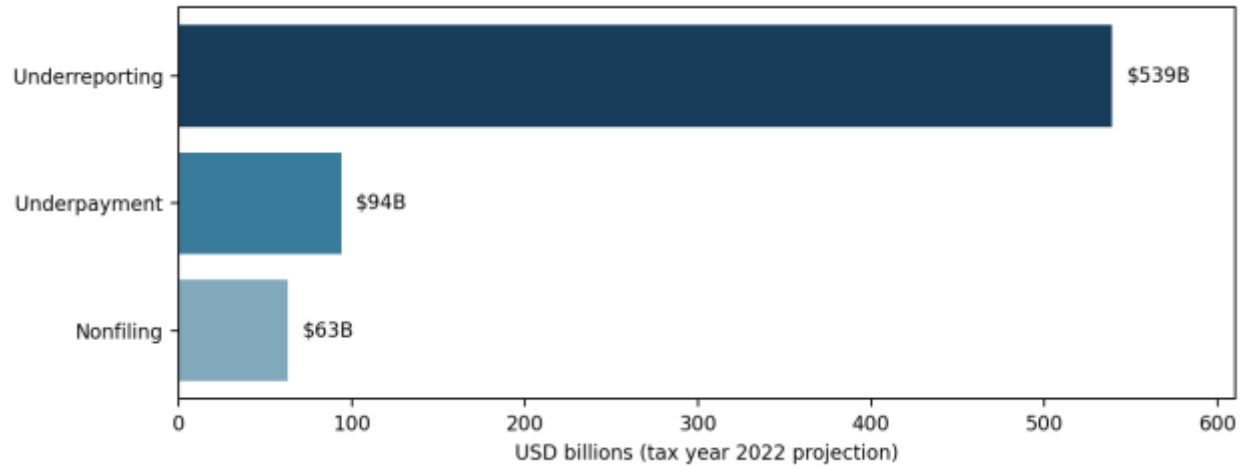

Figure 2. Composition of projected TY2022 U.S. gross tax gap. Source: IRS Publication 5869 (2024).

5.2 Evidence synthesis from cited studies

The original manuscript reports promising but methodologically non-equivalent published outcomes. The following figures are attributable to the cited works as represented in the supplied manuscript; study designs, underlying datasets, and bibliographic details require primary-source checking before publication. They are not pooled estimates and should not be generalized to the IRS without prospective validation.

Source in supplied manuscript	Reported result	Interpretation/limitation
[6] Papíková & Papík	OvR AUC > 0.96	Multiclass severity; not audit revenue
[7] Sun et al.	XGBoost AUC 0.96; precision .89; recall .82	Corporate filings; unknown U.S. transportability

[1] Ampofo et al.	15–20 percentage-point reduction in no-change rate	Secondary report; verify design and baseline
[11] Raghuveer et al.	15% accuracy gain; 20% cost reduction; $15M recovery	Context and calculation require checking
[15] Kong & Saar-Tsechansky	4% uplift in audit profits	Sales tax setting; not federal income tax
[26] Zhang & Zhang	70% processing-time reduction	Data-integration task, not necessarily audits

5.3 Fully specified illustrative capacity scenario

To make operational mechanics explicit without fabricating empirical results, assume 10,000 eligible returns, review capacity K=1,000 (10%), and an administrative rules policy yielding 120 confirmed findings. Four hypothetical adaptive outcomes span 150, 180, 210, and 240 confirmed findings from the same number of reviews. For transparent arithmetic, stipulate $7,000 mean sustained assessed adjustment per confirmed case, a 60% collection fraction, 8 staff hours per review, and $75 fully loaded cost per staff hour. This scenario is entirely hypothetical; none of these values is claimed to be an observed IRS measurement.

Indicator	Rules baseline	Adaptive central	Difference
Reviewed records	1,000	1,000	0
Confirmed findings	120	210	+90
Precision@1,000	12.0%	21.0%	+9.0 pp
No-change rate	88.0%	79.0%	−9.0 pp
Sustained assessments	$840,000	$1,470,000	+$630,000
Collected at 60%	$504,000	$882,000	+$378,000
Reviewer hours	8,000	8,000	0
Review labor cost	$600,000	$600,000	0
Collected $ / reviewer hour	$63.00	$110.25	+$47.25

Calculations: confirmed = K × precision; assessed dollars = confirmed × $7,000; collections = assessed × 0.60; reviewer cost = K × 8 × $75. The scenario has no additional audits and therefore no implied marginal audit-labor expense; additional technology, development, quality-assurance, legal, and maintenance costs are excluded and must be deducted before claiming net benefit. A no-change rate of 79% remains high; these quantities are didactic, not performance targets.

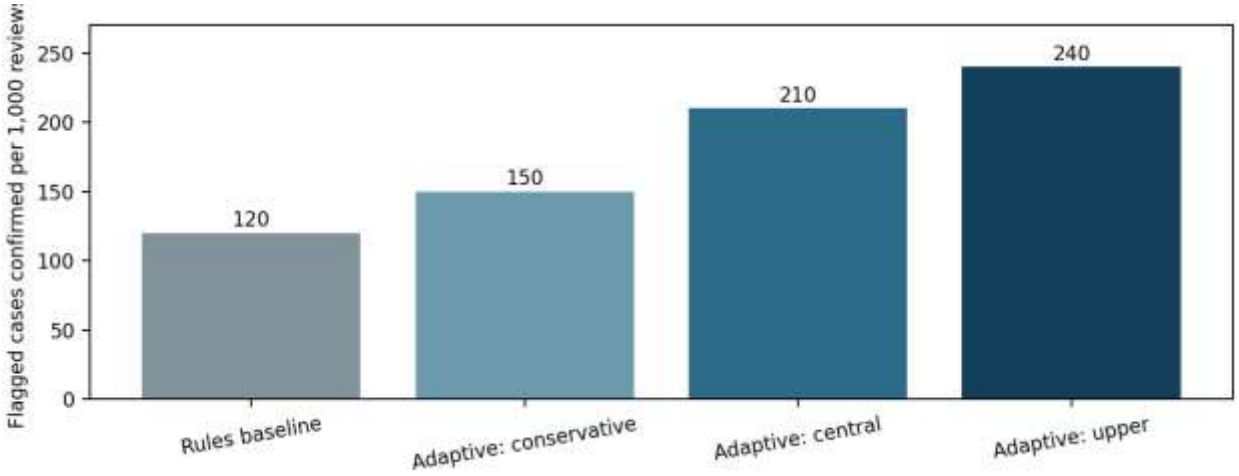

Figure 3. Hypothetical confirmed findings under four policies at exactly 1,000 reviews; illustrative, not observed.

5.4 Scenario sensitivity and uncertainty

Adaptive confirmations	Precision@K	Incremental collections vs baseline	Change in no-change rate
150	15.0%	$126,000	−3 pp
180	18.0%	$252,000	−6 pp
210	21.0%	$378,000	−9 pp
240	24.0%	$504,000	−12 pp

For the baseline of 120 cases and a collection fraction of 60%, incremental collections in scenario s are $\Delta C_s = (TP_s - 120) \times \$7{,}000 \times 0.60$. A separate two-way sensitivity analysis varies the precision of selected cases and the average sustained assessed adjustment. Its plotted values are gross assessed amounts, not incremental realized collections or causal effects.

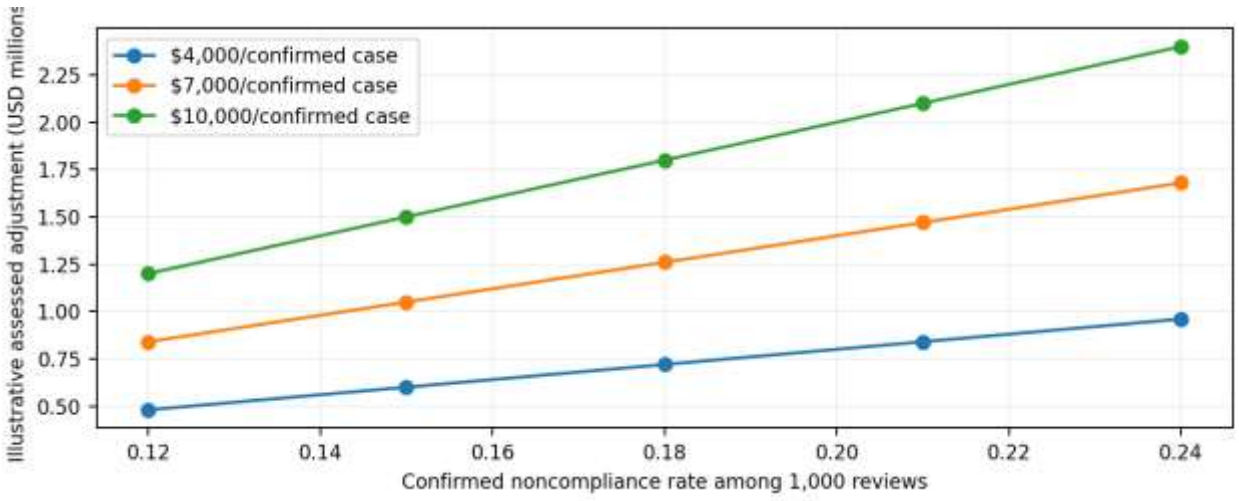

Figure 4. Two-factor hypothetical sensitivity: assessed dollars as a function of precision and mean adjustment.

5.5 Results required for a genuine prospective evaluation

Before an empirical contribution can be asserted, populate the following preregistered evaluation matrix using held-out U.S. cases, consistent audit eligibility, legally permitted randomization, and a pre-specified resolution horizon. Cells marked "Not measured" must not be replaced by the numbers in the illustrative scenarios.

Required comparison	Baseline	Proposed system	Success criterion
Precision@1,000	Not measured	Not measured	$\Delta > 0$, confidence interval reported
PR-AUC and ROC-AUC	Not measured	Not measured	Improves without calibration loss
Brier and ECE	Not measured	Not measured	No material degradation
Collected $/audit hour	Not measured	Not measured	Positive net incremental value
False-positive burden by segment	Not measured	Not measured	Within approved fairness tolerance
Appeals/reversal rate	Not measured	Not measured	No unacceptable increase
Post-drift precision decay	Not measured	Not measured	Improves versus static model

5.6 Discussion: what the calculations establish and do not establish

The IRS figures establish the national scale and composition of noncompliance. The literature shows technical plausibility under specific study conditions. The mathematical specification sets out how an adaptive system can be evaluated, and the numerical scenario shows how a higher audit hit rate might translate into additional sustained assessments and collections at fixed review capacity. None of that establishes model accuracy, causal benefit, taxpayer fairness, net recovery, or national-scale fiscal effects. Scaling the sample scenario to the $696 billion tax gap would be inappropriate: selection coverage, tax types, verification rates, recoverability, and behavioral responses are all unknown.

6. Implications for Tax Administration Policy

The evidence carries significant implications for U.S. tax administration policy across strategic, operational, governance, and equity dimensions.

6.1 Strategic Modernization Priorities

Investment in AI-driven compliance risk management deserves priority. The reported gains are a 15–20 percentage-point reduction in audit no-change rates and a 15–30% reduction in forecasting error [1]. Modernization of this kind works best in phases. The first phase is unglamorous and consists mostly of data integration and quality work. Supervised learning models come next. Hybrid architectures and adaptive learning belong to a later phase, once the data underneath them can support them.

6.2 Governance and Regulatory Alignment

Governance has to keep deployment inside the law. Ampofo et al. treat alignment with NIST AI Risk Management Framework 1.0, applicable federal AI policies, and the privacy requirements of 26 U.S.C. §

6103 as essential rather than optional [1]. On the documentation side that means model cards, fairness assessments, privacy impact assessments, and audit trails. It also means regular algorithmic audits carried out by independent third parties, not by the team that built the system.

6.3 Fairness and Equity Safeguards

Algorithmic fairness is a core requirement, not a reporting obligation bolted on afterward. Regression-based expected-adjustment targets and fairness limits keep the burden from concentrating on vulnerable populations [1]. Monitoring needs more than one metric, tracked across demographic groups: audit selection rates, false positive and false negative rates, audit yield, and average tax adjustments. Privacy-preserving techniques make that analysis possible without exposing sensitive demographic information [21].

6.4 Operational Integration and Change Management

Auditors need training on what these systems can and cannot do, on how to read a model output, on how to check an algorithmic recommendation before acting on it, and on how to use the explanations the system provides. Workflow redesign has to settle the division of labor between the automated system and the human auditor. The performance metrics that matter are audit yield, revenue per audit hour, false positive rates, and taxpayer satisfaction [15].

6.5 Data Infrastructure and Quality

None of this works without data infrastructure capable of integrating diverse sources. Wibowo's XAFR software makes the case for standardization [29], and Zhang et al.'s knowledge graph-enhanced framework shows how heterogeneous data types can be brought together [26]. Data quality work has four dimensions to satisfy: completeness, accuracy, consistency, and timeliness. Nyombi et al. similarly connect big-data analytics with real-time financial oversight and accountability [30].

6.6 Privacy Protection and Data Security

Privacy protection belongs in the design from the start, not in a later hardening pass. Frantz's federated learning framework shows that compliance risk modeling can work without centralizing sensitive data at all [21]. Differential privacy adds calibrated noise and with it a mathematical guarantee [26]. Security has to cover the internal threat as well as the external one, which is what role-based access controls and audit logs are for. Nagalila et al., including Nyombi, integrate financial controls and cybersecurity in a broader resilience framework [31].

6.7 Voluntary Compliance and Taxpayer Services

These systems should serve voluntary compliance, not enforcement alone. Ganyani et al. separate the cases that need enforcement from those a correction notice or a piece of education would resolve [14]. The adaptive explainable AI framework for SMEs supplies visual risk explanations alongside targeted corrective recommendations [19]. Much noncompliance is inadvertent, and explainable AI makes it possible to run taxpayer education programs that address it before it becomes an audit. Nyombi’s Uganda supermarket study associates tax compliance with financial performance, although its findings are context-specific [32].

6.8 International Cooperation and Information Sharing

Evasion increasingly crosses borders, which puts a limit on what any single administration can see. Frantz's TaxFL framework applies federated learning to cross-border compliance [21]. Participation in international initiatives allows collaborative learning while each jurisdiction's privacy law stays intact.

7. Limitations and Future Research

Several limitations warrant acknowledgment and suggest directions for future research.

7.1 Empirical Evidence Limitations

Most of the empirical evidence comes from implementations outside the United States, and U.S.-specific studies are thin. International evidence is informative, but tax systems, legal frameworks, and taxpayer behavior differ enough that transferability cannot be assumed. A further problem cuts across the whole literature: performance is almost always reported on historical data rather than from prospective randomized trials. Rigorous evaluation inside U.S. tax administration agencies, by randomized trial, is the gap that most needs filling. Silva addresses municipal taxation as a distinct legal setting [33].

7.2 Fairness and Bias Challenges

The tension between efficiency and equity is unresolved. Defining fairness in a tax compliance context means choosing among equal audit rates (demographic parity), equal false positive rates (equalized odds), and equal positive predictive values (calibration). These cannot all hold at once, so the choice has to be made explicitly rather than inherited from whichever metric the model happened to optimize. Frameworks for navigating that tradeoff, and for handling intersectional fairness, remain to be developed.

7.3 Explainability and Transparency Tensions

Model complexity and interpretability still pull against each other, and the most accurate models tend to be the least legible. There is a second difficulty that is easy to miss: an explanation detailed enough to satisfy accountability may be detailed enough for a sophisticated taxpayer to game. Methods that deliver both accuracy and interpretability are needed, as is a clearer account of how much explanation can be given before it becomes an evasion manual.

7.4 Adaptive Capabilities and Adversarial Robustness

Co-evolutionary modeling offers a way to anticipate emerging evasion tactics, but whether it works in an operational setting is untested. Adversarial machine learning is a real threat to any deployed model. Three things follow: co-evolutionary approaches need evaluation in operational contexts, models need hardening against adversarial attack, and adaptation needs to be fast enough to respond to a new scheme in something close to real time.

7.5 Organizational and Implementation Challenges

The framework assumes organizational capabilities that many agencies do not have. Implementation barriers, and the strategies for building capacity around them, need study in their own right. Change management and staff acceptance decide whether a technically sound system is actually used. The political economy deserves attention too, since powerful interests have reason to resist effective enforcement and civil liberties advocates have reason to resist AI deployment, and a program can be stopped from either direction.

7.6 Future Research Agenda

The limitations above point to a research agenda with a clear order of priority. The first requirement is prospective evaluation on U.S. administrative data under legally authorized randomization, since nothing in the current literature settles whether these methods transfer. Fairness work needs to move from the selection of metrics to the harder question of which criterion a tax authority should be held to, and on what basis. Explainability research needs to establish how much of a model's reasoning can be disclosed before disclosure starts to assist evasion. Adversarial robustness and rapid adaptation need testing in operational rather than simulated conditions. Running underneath all four is the organizational question of whether agencies can build and keep the capacity these systems assume.

8. Conclusion

This study contributes four things: an auditable design for adaptive tax-compliance risk assessment, a formal economic selection objective, a bias-aware validation plan, and numerical demonstrations with their limits stated explicitly. Official IRS figures put the projected gross tax gap for TY2022 at $696 billion, concentrated in underreporting. The published literature is enough to justify experimentation and not enough to substitute for evaluation on representative U.S. administrative data. Under transparent hypothetical assumptions, a 9-percentage-point gain in confirmed-case yield at a fixed 1,000-review capacity would produce $378,000 of additional collections before system costs. What comes next is a legally authorized, preregistered prospective pilot that measures sustained collections, errors, fairness, appeals, and the operational cost of adaptation.

References

1. Ampofo, I., Amoako, D., & Yeboah, M. M. (2026). From compliance to intelligence: Integrating AI and predictive analytics into U.S. tax compliance and revenue systems. Magna Scientia Advanced Research and Reviews, 17(2), 022–029. https://doi.org/10.30574/msarr.2026.17.2.0119
2. Qureshi, O. U. R. (2026). AI for tax and customs anomaly detection: A PRISMA-guided systematic review and proposed hybrid AI framework for trade and revenue administration in emerging digital economies. https://doi.org/10.64388/irev9i9-1715244
3. Hemberg, E., Rosen, J., Warner, G., Wijesinghe, S., & O'Reilly, U.-M. (2015). Tax non-compliance detection using co-evolution of tax evasion risk and audit likelihood. In Proceedings of the 2015 Conference on Genetic and Evolutionary Computation (pp. 79–88). ACM. https://doi.org/10.1145/2746090.2746099
4. Hemberg, E., Rosen, J., Warner, G., Wijesinghe, S., & O'Reilly, U.-M. (2016). Detecting tax evasion: a co-evolutionary approach. Springer Netherlands.
5. DeBarr, D., & Eyler-Walker, Z. (2006). Closing the gap: automated screening of tax returns to identify egregious tax shelters. SIGKDD Explorations, 8(1), 11–16. https://doi.org/10.1145/1147234.1147237
6. Papíková, L., & Papík, M. (2026). Severity-based prioritisation of administrative tax non-compliance: A multiclass prediction approach using financial and governance data. Intelligent Systems in Accounting, Finance and Management, 33(2). https://doi.org/10.1002/isaf.70036
7. Sun, T., Wang, M., & Chen, J. (2025). Leveraging machine learning for tax fraud detection and risk scoring in corporate filings. Asian Business Research Journal, 10(11), 1–13. https://doi.org/10.55220/2576-6759.637
8. Pires, C. A. de S. (2024). Supervised machine learning for tax evasion detection: A case study with the Brazilian tax administration (pp. 288–294). https://doi.org/10.5753/sbbd_estendido.2024.244262
9. Ismail, M. M. B., & Alsadhan, N. (2023). Simultaneous classification and regression for Zakat under-reporting detection. Applied Sciences, 13(9), 5244. https://doi.org/10.3390/app13095244
10. Gayathri, D., Umar, K., Sainath, S., Mishra, S., Srilatha, S., & Mounika, A. (2026). Tax evasion detection using AdaBoost and DNN (pp. 92–96). https://doi.org/10.1109/assic70037.2026.11717265
11. Raghuveer, K., Ram, A. S., Varshini, G. V., Nader, A. S., Alazzam, M. B., & Raj, I. I. (2025). Predictive modeling for tax audits: Leveraging machine learning to improve government revenue collection (pp. 1–6). https://doi.org/10.1109/isssc66652.2025.11388891
12. Businge, P., Agaba, I. A., Atuhaire, J. I., Nayebale, F. I., Ariho, J. G., Mugalu, B., Musinguzi, D., Malingu, C. J., & Kabwama, C. A. (2025). Adaptive financial fraud detection using graph neural

networks and reinforcement learning. World Journal of Advanced Research and Reviews, 28(2), 842–847. https://doi.org/10.30574/wjarr.2025.28.2.3761
13. Ofoma, I. S. (2026). Closing the tax gap: an analytics-based framework for fraud-risk detection and audit prioritization in state and federal tax administration. World Journal of Advanced Research and Reviews, 31(3), 1577–1593. https://doi.org/10.30574/wjarr.2026.31.3.2365
14. Ganyani, L., Nare, S., Danda, C., Chingezi, L., & Mupa, M. N. (2026). From tax mismatches to revenue protection: A data-driven framework for audit prioritization, voluntary compliance, and SME compliance education. https://doi.org/10.64388/irev10i1-1719533
15. Kong, D., & Saar-Tsechansky, M. (2014). Collaborative information acquisition for data-driven decisions. Machine Learning, 95(1), 71–86. https://doi.org/10.1007/s10994-013-5424-x
16. Martikainen, J. (2012). Data mining in tax administration: Using analytics to enhance tax compliance. University Thesis.
17. Atanasijevic, J., Jakovetic, D., Krejic, N., Krklec-Jerinkic, N., & Markovic, D. (2019). Using big data analytics to improve efficiency of tax collection in the tax administration of the Republic of Serbia. Ekonomski Pregled, 67, 115–130. https://doi.org/10.5937/EKOPRE1808115A
18. Zhang, T.-T. (2026). SHAP-guided risk path generation for IRS auditors: Interpretable anomaly attribution in payroll tax compliance screening. https://doi.org/10.1145/3819836.3819908
19. Anonymous. (2025). From black box to actionable insights: An adaptive explainable AI framework for proactive tax risk mitigation in small and medium enterprises (pp. 193–199). https://doi.org/10.1145/3785706.3785736
20. Nayebale, F. I., Inakpenu, E. L., Makumbi, I. S., & Makandah, E. A. (2026). Design of a secure AI-driven adaptive audit transparency engine to improve tax compliance, reduce administrative inefficiencies and strengthen overall economic prosperity. World Journal of Advanced Research and Reviews, 29(2), 425–441. https://doi.org/10.30574/wjarr.2026.29.2.0243
21. Frantz, P. A. (2026). TaxFL: Federated learning and federated graph intelligence for cross-border tax compliance. Zenodo. https://doi.org/10.5281/zenodo.18602470
22. Katamba, I. (2025). Strengthening AI-driven regulatory compliance systems in U.S. gaming and taxation sectors. Zenodo. https://doi.org/10.5281/zenodo.18829759
23. Guo, B. (2025). Data-driven tax risk management: Legal strategies in the era of big data (pp. 273–277). https://doi.org/10.1145/3757110.3757157
24. Sekinobe, M., Mukasa, K., Nayebale, F. I., & Kato, J. (2025). An interdisciplinary framework for the development of intelligent accounting automation systems integrating predictive risk analytics and dynamic internal control mechanisms. International Journal of Innovative Science and Research Technology, 10(12), 2520–2533. https://doi.org/10.38124/ijisrt/25dec1138
25. Sharifur, R., & Rahul, D. (2025). An exploration of artificial intelligence techniques for optimizing tax compliance, fraud detection, and revenue collection in modern tax administrations. Zenodo. https://doi.org/10.5281/zenodo.14606935
26. Zhang, T., & Zhang, T. (2025). A knowledge graph-enhanced multimodal AI framework for intelligent tax data integration and compliance enhancement. https://doi.org/10.71465/fbf384
27. Goumagias, N., Hristu-Varsakelis, D., & Assael, Y. M. (2018). Using deep Q-learning to understand the tax evasion behavior of risk-averse firms. Expert Systems with Applications, 101, 258–270. https://doi.org/10.1016/j.eswa.2018.01.039
28. Nyombi, A., Nagalila, W., Happy, B., Sekinobe, M., & Ampe, J. (2024). Enhancing cybersecurity

protocols in tax accounting practices: Strategies for protecting taxpayer information. World Journal of Advanced Research and Reviews, 23(3), 1788–1798. https://doi.org/10.30574/wjarr.2024.23.3.2838

29. Wibowo, B. D. S. (2022). XBRL open information model for risk based tax audit using machine learning. International Journal of Informatics, Information System and Computer Engineering. https://doi.org/10.34010/injiiscom.v3i1.6891
30. Nyombi, A., Ampe, J., Happy, B., Sekinobe, M., Nagalila, W., & Masaba, B. (2025). Leveraging big data for real-time financial oversight in non-profit and government accounting: A framework to empower accountants and improve transparency. World Journal of Advanced Research and Reviews, 26(2), 2544–2554. https://doi.org/10.30574/wjarr.2025.26.2.1937
31. Nagalila, W., Nyombi, A., Sekinobe, M., Ampe, J., & Happy, B. (2024). Fortifying national security: The integration of advanced financial control and cybersecurity measures. World Journal of Advanced Research and Reviews, 23(2), 1095–1101. https://doi.org/10.30574/wjarr.2024.23.2.2444
32. Nyombi, A. (2022). Income tax compliance, tax incentives and financial performance of supermarkets in Mbarara City, South Western Uganda (Master's dissertation). Mbarara University of Science and Technology.
33. Silva, S. M. da C. (2023). Tributação municipal curso avançado de direito tributário municipal. Zenodo. https://doi.org/10.5281/zenodo.8252150