

Systematic Review

Artificial Intelligence in Tax Compliance and Evasion Mitigation: Trends, Mechanisms, and Institutional Implications

Houda Zaim *  and Siham Sahbani 

National School of Commerce and Management (ENCG), Sidi Mohamed Ben Abdellah University (USMBA),
Fez 30000, Morocco; siham.sahbani@usmba.ac.ma

* Correspondence: houda.zaim1@usmba.ac.ma

Abstract

This study presents a systematic literature review of 68 peer-reviewed articles (2015–2025) on artificial intelligence in tax compliance and evasion mitigation. Using the PRISMA 2020 protocol and textometric analysis via IRAMUTEQ software, we map publication trends, geographic distribution, and three research paradigms: machine learning and predictive modeling; artificial intelligence, technology and tax compliance; and government, financial development and revenue administration. The CIMO (Context–Intervention–Mechanism–Outcome) framework structures our synthesis of how institutional conditions shape intervention design and why identical technologies produce divergent outcomes across settings. While existing reviews have focused primarily on detection metrics without theorizing institutional boundary conditions, behavioral dynamics without addressing governance capacity, or ethical deficits without a theoretical framework, this study constructs the Adaptive AI Tax Compliance Framework (AAITCF), a context-sensitive implementation roadmap differentiated across three institutional maturity tiers. The results indicate that AI achieves high detection accuracies in digitally mature economies, yet effectiveness is contingent on data quality, governance capacity, and organizational readiness. Developing countries face structural asymmetries, infrastructural deficits, and human capital gaps that constrain algorithmic performance even where technical sophistication is high. The AAITCF treats context as constitutive of intervention effectiveness and identifies under-explored areas regarding causal pathways from AI deployment to long-term institutional change, taxpayer trust, and equitable fiscal governance.

Keywords: artificial intelligence; tax compliance; tax evasion; CIMO framework; systematic review; textometric analysis; machine learning; tax administration



Academic Editor: Thanasis Stengos

Received: 20 May 2026

Revised: 29 June 2026

Accepted: 2 July 2026

Published: 9 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons](#)

[Attribution \(CC BY\)](#) license.

1. Introduction

Worldwide, tax non-compliance and tax evasion reduce governments' fiscal capacity, taking away revenues needed for public investment. The revenue loss from corporate taxes due to offshore jurisdictions is estimated at more than USD 1 trillion per year, and cumulative losses over the next decade are projected to approach USD 5 trillion ([Tax Justice Network, 2024](#); [Alstadsæter et al., 2024](#)). Distributional impacts are uneven: losses in lower-income countries take a significant proportion of public health budgets, whereas in advanced economies, the issues involve complex profit shifting by multinationals ([OECD, 2020](#); [IMF, 2024](#)). This diversity further increases the call for enforcement measures tailored to meet varying institutional contexts.

Artificial intelligence has emerged as one of the most promising areas in tax administration. Tax authorities can now perform pattern recognition, predictive risk assessment, and real-time anomaly detection at unprecedented scales thanks to machine learning algorithms, neural networks, and natural language processing systems. Empirical findings show significant performance improvements, such as those of [Murorunkwere et al. \(2022\)](#), who achieved 92% fraud detection accuracy in Rwanda with neural networks. In Italy, [Battaglini et al. \(2025\)](#) found that machine learning-selected audits substantially improved the detection of evasion (39%) and the recovery of revenue (29%) compared to human-selected audits. For Chinese enterprise data, [Yang \(2025\)](#) achieved similar accuracy with random forest models. In addition to detection, researchers have considered some specific technical challenges: hybrid unsupervised approaches for small amounts of labeled data ([Savić et al., 2022](#)) and the generation of synthetic datasets using generative adversarial networks when tax returns are not directly available ([Visitpanya & Samanchuen, 2023](#)).

However, the progress made varies substantially across contexts. In Estonia, the advanced use of multi-application AI has been made possible by a well-developed digital infrastructure, while in Botswana and Zambia, researchers found that algorithmic tools need to be significantly adjusted to deal with fragmented infrastructure and limited audit data ([Adu-Ababio et al., 2025](#); [Munjeji & Schutte, 2024](#)). In addition to this, algorithmic risk scoring coupled with targeted interventions nearly doubled the compliance effects as compared to untargeted interventions ([Hauser et al., 2025](#)), and governance quality and institutional trust are independent moderators of tax evasion ([Khaltar, 2024](#); [Davidescu et al., 2022](#)).

AI implementation in the tax administration process can be viewed through the lens of operational risk, where its impact is dependent on identifying, assessing, and mitigating context-specific risks like data quality, regulatory fragmentation, and organizational preparedness. This puts the study into the nexus of technological innovation, institutional capacity, and governance in financial systems.

Even with this growing body of evidence, recent systematic reviews remain limited. [Belahouaoui and Attak \(2024\)](#) conducted textometric analysis on 62 studies, but they focused on detection performance metrics rather than theorizing institutional boundary conditions. In a review of 19 studies on generational preferences in the adoption of AI, [Hakam et al. \(2024\)](#) focused on behavioral dynamics but not on the capacity of governance or the readiness of infrastructure. [Mudome and Ndiege \(2025\)](#) collated the findings of 20 studies from developing countries but did not provide a theoretical or operational framework to address these ethical weaknesses. To our knowledge, no existing review applies the CIMO (Context–Intervention–Mechanism–Outcome) analytical approach or offers a differentiated implementation roadmap at different institutional maturity levels.

This research aims to fill these gaps by conducting a systematic review of 68 peer-reviewed articles (2015–2025) based on the PRISMA protocol and combining textometric analysis through the software IRAMUTEQ with the CIMO ([Denyer et al., 2008](#)) framework, which is unique in considering institutional context as an integral factor in the success of interventions.

The main research question is: **Under what institutional, technological, and organizational conditions are AI interventions effective for increasing tax compliance and decreasing tax evasion, and through what mechanisms do these effects happen?**

This study pursues two objectives. First, we map the intellectual structure of the field through textometric analysis, identifying three research paradigms that organize the existing literature: machine learning and predictive modeling, artificial intelligence technology, tax compliance, and government financial development and revenue administration. Second, we synthesize CIMO-based evidence chains to construct the Adaptive AI Tax

Compliance Framework (AAITCF), a theoretically grounded, context-sensitive framework applicable across the full spectrum of institutional maturity, from data-rich advanced administrations to resource-constrained environments with fragmented infrastructure. Unlike previous reviews focused primarily on algorithmic performance in advanced-economy settings, this study synthesizes the institutional conditions under which AI-driven compliance interventions succeed or fail, aligning with risk management principles of context-dependent intervention design and adaptive governance.

The article proceeds as follows. Section 2 details the methodological approach, including database selection, the PRISMA protocol, and textometric procedures. Section 3 presents textometric and descriptive findings on publication trends, geographic distribution, and research paradigms. Section 4 synthesizes the CIMO-based evidence and presents the AAITCF. Section 5 discusses contributions, limitations, and future research directions.

2. Materials and Methods

The literature review was carried out using the PRISMA 2020 statement (Page et al., 2021) because of its clarity and systematic rigor. The goal of PRISMA 2020 is to reduce bias in the reporting process and provide full coverage through three phases: identification, screening, and inclusion (Page et al., 2021). This methodology was chosen because it matched our research objectives and was a way to explore the available literature in this emerging field in detail and objectively. The structured approach from the selection of studies to their inclusion in the review, through adherence to the PRISMA 2020 guidelines, enabled a comprehensive and reliable review process, thereby advancing the understanding of how AI technologies can be applied in tax compliance and tax evasion mitigation in various institutional and regulatory contexts. The PRISMA 2020 checklist is presented as Table S1 in the Supplementary Materials.

2.1. Data Selection

We included data from Scopus and Web of Science (Core Collection) as the primary databases, as they cover a higher number of quality publications, have a wide coverage in the disciplines, advanced searching capabilities, and provide the possibility for reproducible systematic searching (Mengist et al., 2020; Paul et al., 2021). The selection of this strategic combination was made to achieve a balanced and comprehensive search, especially for research at the intersection of economics, management, and artificial intelligence. Combined, they allow for a broad yet targeted search of the literature available, supporting a comprehensive understanding of the topic.

This study adopted a multi-stage approach and was conducted based on the PRISMA 2020 research protocol (Figure 1). We performed systematic searches across two main databases, Scopus and Web of Science, in January 2026, with a broad list of keywords covering a combination of terms associated with the concepts of “artificial intelligence”, “machine learning”, “tax compliance”, and “tax evasion” to identify relevant papers. The search was conducted independently by both authors; any disagreements or ambiguities were resolved by consensus.

The advanced search employed a targeted combination of keywords and codes, as detailed in Table 1. Search strings paired “Artificial Intelligence” with terms related to tax compliance, tax evasion, tax avoidance, and tax administration. Keywords such as “Machine learning”, “Tax compliance”, and “Tax evasion” were used to filter article titles, abstracts, and keywords within the decade spanning 2015–2025.

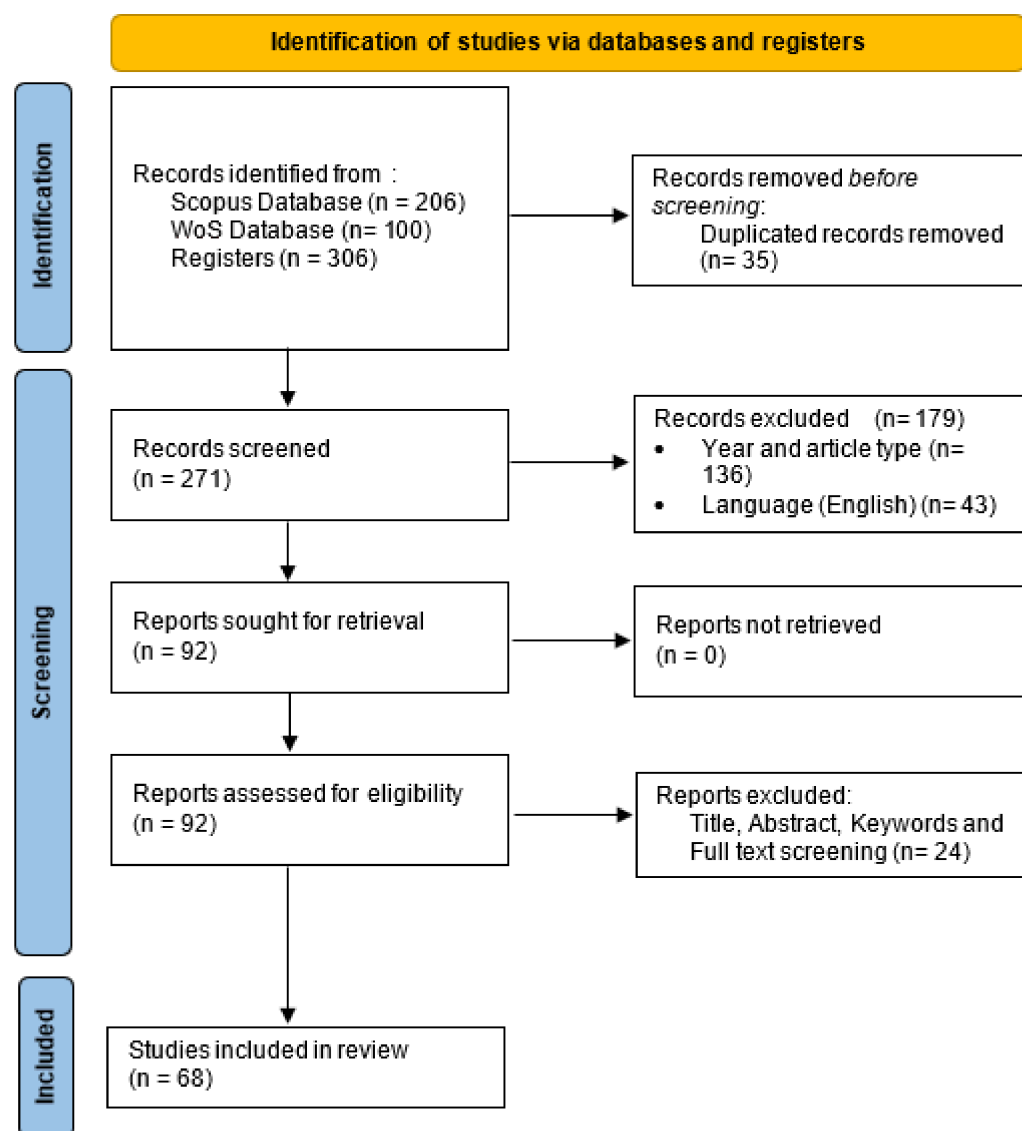


Figure 1. PRISMA 2020 research protocol.

Table 1. Keywords and code used in advanced search.

Database	Scopus	Web of Science
Search strings	TITLE-ABS-KEY ("artificial intelligence" OR "machine learning") AND ("tax compliance" OR "tax evasion")	TS = ("artificial intelligence" OR "machine learning") AND ("tax compliance" OR "tax evasion")
Keywords filter	"Artificial intelligence" "Machine learning" "Tax compliance" "Tax evasion" "Tax avoidance" "Tax administration" "Fraud detection" "Tax systems"	
Inclusion and exclusion criteria	The inclusion criteria retained peer-reviewed articles published between 2015 and 2025, written in English and explicitly addressing AI-based technologies in tax compliance or tax evasion, while the exclusion criteria eliminated non-English or inaccessible papers, studies outside the selected timeframe, and works unrelated to AI applications in tax compliance or lacking academic rigor	
Scanned items	Article title, Abstract, Keywords, and Full Text	
Year range	2015–2025	

Note: The keywords filter terms "tax avoidance", "tax administration", "fraud detection", and "tax systems" were used as screening keywords during title/abstract screening and full-text assessment to capture studies employing alternative terminology but were not part of the initial Boolean search strings in Scopus and Web of Science.

During the identification stage, 306 records were identified through database screening (Scopus: 206; Web of Science: 100). Duplicates were removed using a two-step procedure: first, automatic deduplication was performed in EndNote 20 based on title, author, and DOI matching; second, manual verification was conducted to resolve ambiguous cases (e.g., preprint vs. published versions). A total of 35 duplicate records were removed, leaving 271 unique records. After screening year and language, 179 records were excluded (136 due to year and article criteria; 43 records were excluded because they were not written in English), leaving 92 records for full-text eligibility. All 92 records were successfully retrieved and assessed for eligibility. After assessing titles, abstracts, keywords, and full contents, 24 records were excluded as unrelated to AI applications in tax compliance or lacking sufficient academic rigor. This resulted in 68 studies being identified as eligible, and all 68 studies were included in the systematic literature review.

2.2. Data Analysis

The data analysis begins with a descriptive and textometric review to map the research landscape on artificial intelligence and tax compliance, first examining publication trends over time to characterize the field's evolution and growth.

The trend of publications over time is shown in Figure 2, which reveals a gradual increase from 2015 to 2022, before the number increased significantly in recent years, with 10 publications in 2023, 12 in 2024, and the highest number of 25 in 2025. This trend mirrors the increasing academic interest in the field of using AI for tax compliance and will be explored more in Section 3.

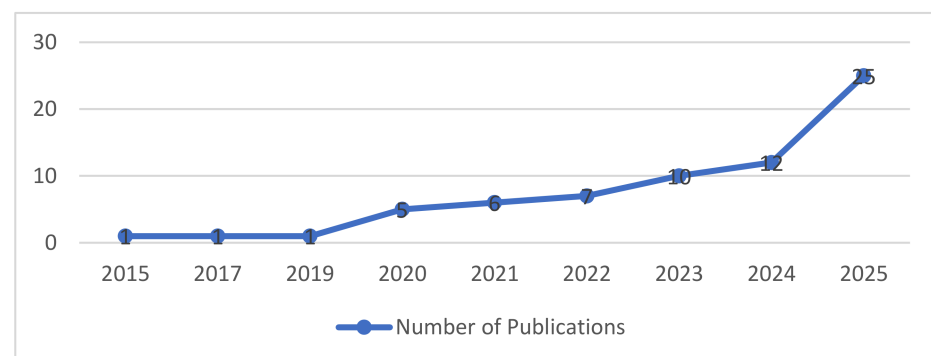


Figure 2. Documents by year. Source: Created by the authors. Note: the figure uses the publication year downloaded from Scopus, which may differ from the actual publication year reported in Table S2.

The analysis then moves to the geographical distribution of publications to determine the top regions contributing to the publications and to see the global reach of the research in this area.

The geographic distribution of publications is shown in Figure 3. Research is widely spread among continents, where the highest share comes from Europe, followed by Asia and North America. Less significant but still notable contributions from other areas, such as Africa, indicate the growing global impact of AI-in-tax-compliance scholarship. These distributional patterns are discussed in more detail in Section 3.

For the textometric analysis, we used IRAMUTEQ Version 0.8 alpha 7, a statistical software used to cluster the corpus into thematic areas (Ratinaud, 2009). The tool is particularly useful for data visualization and the detection of emerging themes in a text corpus. The abstracts and titles of all the selected articles were assembled into a plain text file following IRAMUTEQ conventions, with each article coded as an initial context unit (UC). In the pre-processing stage, the software lemmatizes words to their dictionary forms

and filters out common stop words such as prepositions and conjunctions to extract the active vocabulary.

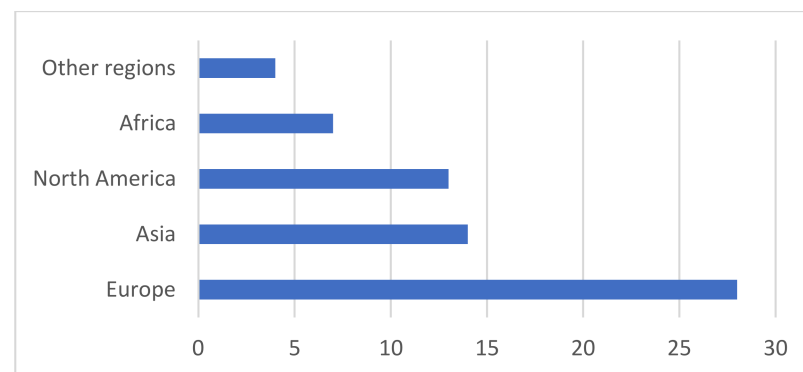


Figure 3. Geographic distribution of publications. Source: Created by the authors.

The software then segments the preprocessed text into elementary context units (UCs) and applies Descending Hierarchical Classification (DHC) and Correspondence Factorial Analysis (CFA) to extract underlying themes. The corpus was divided into text units of 40 occurrences each, yielding 420 segments. DHC was performed using Reinert's algorithm, producing three clusters with a classification rate of 83.81% (352 of 420 segments classified).

Each cluster was assigned a name and interpretation grounded in the literature; this naming process was validated through researcher consensus. The method relies on word co-occurrence patterns, with cluster categorization driven by the classification algorithm rather than subjective interpretation, thereby reducing researcher bias and ensuring the reproducibility of thematic results.

The analysis starts with a general view of the data (Table 2), the list of the 50 most frequent words (Table 3), a cluster dendrogram (Figure 4), and then a detailed analysis of the most representative words per class (Table 4), factorial correspondence analysis (Figure 5), and a proposed framework (Figure 6).

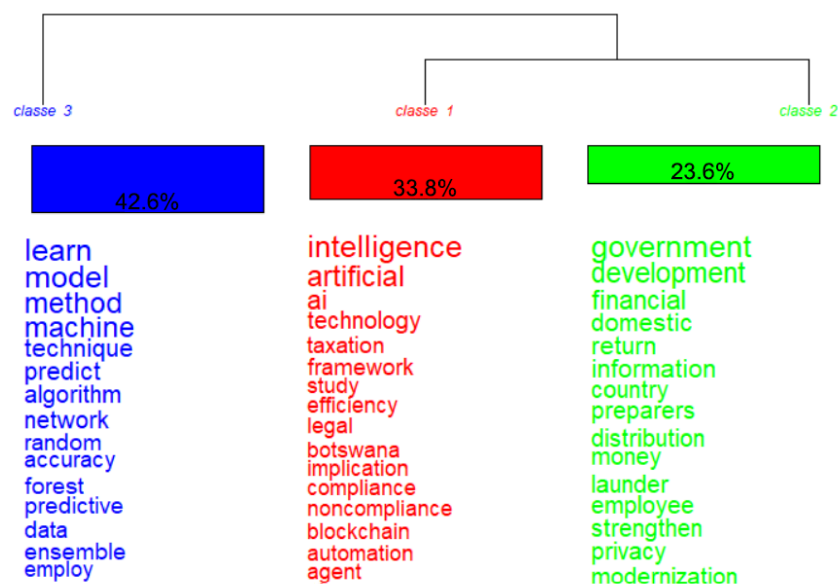
Table 2. Statistical summary of the corpus.

Concept	Total Number
Number of texts	68
Number of text segments	420
Segment size	40 occurrences
Number of occurrences	14,762
Mean of occurrences per text	217.09
Number of lexical forms	2986
Lemmatized forms	2394
Active forms	2030
Supplementary forms	364
Retained active forms	737
Minimum frequency threshold	≥ 3
Number of clusters	3
Classified segments	352/420
Classification rate	83.81%
Number of hapaxes	1090 (45.53% of forms/7.38% of occurrences)

Source: Created by the authors using IRAMUTEQ software.

Table 3. The 50 most frequently used active words.

Order	Active Forms	Freq.	Order	Active Forms	Freq.
01	tax	492	26	digital	35
02	data	94	27	result	34
03	model	93	28	paper	33
04	Ai	86	29	detection	31
05	compliance	85	30	increase	30
06	learn	82	31	policy	30
07	study	81	32	accuracy	29
08	evasion	68	33	identify	28
09	taxpayer	68	34	information	28
10	machine	65	35	business	26
11	artificial	59	36	fraud	26
12	intelligence	57	37	finding	25
13	audit	56	38	enhance	25
14	approach	50	39	include	25
15	administration	49	40	collection	24
16	risk	48	41	application	24
17	revenue	45	42	gap	24
18	technology	43	43	report	24
19	base	41	44	financial	24
20	improve	39	45	government	23
21	system	39	46	taxation	23
22	analysis	38	47	individual	23
23	method	38	48	income	23
24	research	36	49	detect	23
25	algorithm	35	50	framework	22

**Figure 4.** Dendrogram of descending hierarchical classification (DHC) illustrating thematic clusters related to artificial intelligence and tax compliance. Source: Created by the authors using IRAMUTEQ software.

The dendrogram in Figure 4 reveals three thematic clusters. Cluster 3 (42.61%) centers on machine learning and predictive modeling, featuring terms such as *learn*, *model*, *method*, *machine*, *predict*, *algorithm*, *network*, *accuracy*, *forest*, *data*, *ensemble*, *detect*, and *risk*. Cluster 1 (33.81%) addresses artificial intelligence, technology, and tax compliance, with key terms including *intelligence*, *artificial*, *AI*, *technology*, *taxation*, *framework*, *legal*, *compliance*, *noncom-*

pliance, blockchain, automation, and adoption. Cluster 2 (23.58%) focuses on government, financial development, and revenue administration, represented by *government, development, financial, domestic, return, country, money, launder, employee, privacy, GDP, and state*. Together, these clusters reflect complementary research paradigms within the corpus.

Table 4. The most representative words by class based on DHC analysis.

Cluster 3	Freq.	%	Chi²	Cluster 1	Freq.	%	Chi²	Cluster 2	Freq.	%	Chi²
learn	53	81.54	49.394	intelligence	43	91.49	80.651	government	14	73.68	27.981
model	54	79.41	46.669	artificial	42	87.50	71.603	development	11	73.33	21.522
method	33	94.29	42.433	ai	45	75.00	54.846	financial	13	61.90	18.204
machine	44	83.02	41.655	technology	27	79.41	34.978	domestic	5	100.00	16.438
technique	20	100.00	28.556	taxation	16	84.21	22.801	return	6	85.71	15.302
predict	18	100.00	25.546	framework	17	80.95	22.182	information	11	61.11	14.829
algorithm	24	85.71	23.108	study	37	57.81	20.144	country	11	57.89	13.124
network	16	100.00	22.573	efficiency	14	82.35	18.812	preparers	4	100.00	13.113
random	12	100.00	16.730	legal	9	100.00	18.084	distribution	4	100.00	13.113
accuracy	20	80.00	15.382	botswana	8	100.00	16.028	money	4	100.00	13.113
forest	11	100.00	15.291	implication	9	90.00	14.523	launder	4	100.00	13.113
predictive	13	92.86	15.051	compliance	39	52.00	14.097	employee	4	100.00	13.113
data	47	61.84	14.655	noncompliance	7	100.00	13.984	strengthen	4	100.00	13.113
ensemble	12	92.31	13.631	blockchain	7	100.00	13.984	privacy	4	100.00	13.113
employ	11	91.67	12.224	automation	6	100.00	11.952	modernization	4	100.00	13.113
detect	16	80.00	12.120	agent	6	100.00	11.952	lack	4	100.00	13.113
propose	14	82.35	11.535	digitalization	6	100.00	11.952	international	4	100.00	13.113
declaration	10	90.91	10.830	article	9	81.82	11.696	infrastructure	4	100.00	13.113
risk	20	71.43	10.328	tool	7	87.50	10.546	gdp	4	100.00	13.113
paper	20	71.43	10.328	adoption	7	87.50	10.546	state	4	100.00	13.113

Source: Created by the author using IRaMuTeQ software Version 0.8 alpha 7.

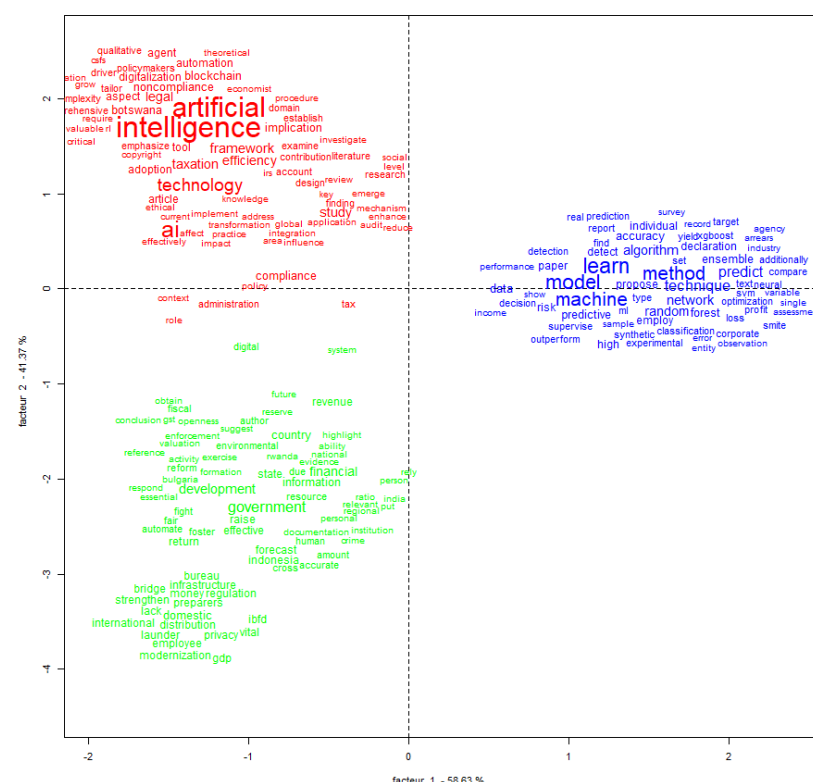


Figure 5. Factorial correspondence analysis of the textual corpus. Source: Created by the authors using IRAMUTEQ software.

The factorial correspondence analysis in Figure 5 visualizes the relationship and correlation between the various keywords used in the literature on AI and tax compliance. In this graph, the closer the words are to each other, the more frequently they are associated within the documents. It is clear from the dense clustering of terms like learn, model, and

predict that there is a strong focus on machine learning and predictive modeling in AI. Another dense cluster includes artificial intelligence, technology, and taxation, highlighting the technological and regulatory convergence in this research area. A third cluster groups government, development, and financial, reflecting the institutional and administrative dimension of tax compliance.

2.3. Assessment of Quality of Included Studies

A formal quality appraisal was conducted for all 68 studies included in this systematic review. Given the methodological heterogeneity of the corpus, spanning supervised machine learning studies, legal-theoretical analyses, qualitative institutional studies, systematic reviews, field experiments, and simulations, quality appraisal criteria were developed a priori, drawing on established mixed-methods quality assessment principles and supplemented by domain-specific criteria for AI-in-tax-compliance research.

The appraisal instrument was adapted from the Mixed Methods Appraisal Tool (MMAT; Hong et al., 2018), selected because the corpus combines randomized and non-randomized quantitative studies, qualitative and mixed-methods studies, and conceptual or legal-theoretical contributions that single-paradigm tools such as the Cochrane Risk of Bias tool (Higgins et al., 2011) or CASP checklists cannot accommodate jointly. The MMAT's generic screening criteria (clarity of the research question, appropriateness of the design, and adequacy of the data to address the question) were retained for all 68 studies, and its method-specific quality criteria (e.g., randomization and allocation procedures for experimental designs; sampling strategy and analytic rigor for qualitative designs; confounding and outcome measurement for non-randomized quantitative designs) were mapped onto the categories reported in Table 5, Panels B–E. Because the MMAT was not designed for AI/ML-specific methodology, it was supplemented with seven domain-specific criteria, reported in Table 5, Panels A and C, covering external validation of ML models, treatment of class imbalance, explainability / feature-importance reporting, comparative algorithm benchmarking, explicit theoretical grounding, reproducibility (code or data availability), and treatment of governance, bias, or taxpayer-rights dimensions. These domain-specific criteria were generated inductively from a pilot coding of ten randomly selected studies spanning the corpus's major methodological types, refined through consensus discussion between the authors, and then applied to the full sample of 68 studies. This two-tier procedure, a validated generic instrument supplemented by literature-specific criteria, follows established practice for quality appraisal in methodologically heterogeneous systematic reviews (Hong et al., 2018; Pluye & Hong, 2014) and is intended to make the basis for each criterion in Table 5 traceable rather than ad hoc.

Appraisal coding was conducted independently by both authors using a standardized protocol, with disagreements resolved by consensus. Rather than individual per-study scores, findings are aggregated as frequencies and percentages across seven appraisal dimensions, providing a cross-corpus summary that directly informs interpretation of the CIMO synthesis.

Implications for the CIMO synthesis. Three appraisal-derived limitations directly shape the synthesis in Section 4. First, the 63.2% single-country scope and 52.9% absence of external validation mean that accuracy benchmarks are context-specific, not transferable, a constraint reflected in the AAITCF's tiered design. Second, the 58.8% of correlation-based studies means that mechanistic claims in Section 4.3 are probabilistic rather than causally established; the only strong causal evidence comes from Hauser et al. (2025) and Battaglini et al. (2025). Third, the 35.3% of studies with insufficient governance and ethics treatment motivate the AAITCF's mandatory governance feedback loop as a structural corrective.

Table 5. Quality Appraisal of Included Studies (N = 68): Aggregated Distribution.

Quality Appraisal Criterion	N (Studies)	% of 68
A. Study Design Strengths		
Uses validated/established ML algorithms (RF, XGBoost, SVM, neural nets, etc.)	28	41.2%
Employs cross-validation or train/test split for model evaluation	25	36.8%
Uses large real-world or administrative datasets (≥ 1000 observations)	23	33.8%
Comparative benchmarking of multiple algorithms	18	26.5%
Provides feature importance or explainability analysis (SHAP, ANOVA, RF importance)	12	17.6%
Addresses class imbalance (SMOTE, GANs, focal loss, positive-unlabeled learning)	9	13.2%
Explicit theoretical grounding (CIMO, TOE, SEM, behavioral economics, legal frameworks)	20	29.4%
All studies: peer-reviewed in Scopus/WoS indexed journals	68	100%
B. Study Design Distribution by Method		
Empirical quantitative with ML/AI methods	32	47.1%
Legal/theoretical/policy analysis	14	20.6%
Mixed methods (survey, social media analytics, content analysis)	7	10.3%
Empirical qualitative (interviews, thematic analysis, case studies)	5	7.4%
Systematic or bibliometric literature review	6	8.8%
Simulation/agent-based/computational modeling	2	2.9%
Randomized field experiment (behavioral + ML targeting)	1	1.5%
Narrative/methodological review	1	1.5%
Total	68	100%
C. Methodological Weaknesses and Gaps		
Limited or no external validation of ML models	36	52.9%
Lack of reproducibility (no code or data sharing reported)	41	60.3%
No causal identification strategy; correlation-based findings only	40	58.8%
Single-country or single-dataset scope	45	66.2%
Data confidentiality constraints limiting dataset quality or size	19	27.9%
Insufficient treatment of governance, ethical, or taxpayer rights dimensions	24	35.3%
Short or cross-sectional timeframe (no longitudinal tracking)	35	51.5%
Theoretical framework absent or implicit in technical studies	18	26.5%
Class imbalance unaddressed in fraud/evasion classification tasks	14	20.6%
D. Risk of Bias		
Selection bias (administrative convenience samples, non-random cases)	38	55.9%
Reporting bias (publishing best-performing model only)	26	38.2%
Temporal bias (train and test from same narrow time window)	17	25.0%
Confirmation bias (results framed to support pre-defined policy narrative)	12	17.6%
No explicit bias mitigation reported	28	41.2%
Partial bias mitigation (cross-validation, holdout sets, SMOTE, etc.)	40	58.8%
E. Generalizability		
Low generalizability, single country or specific tax type/sector	43	63.2%
Moderate, regional or multi-jurisdiction within one continent	10	14.7%
High, multi-country, global scope, or intentionally comparative	12	17.6%
Not applicable, purely theoretical or conceptual	3	4.4%
F. Geographic Distribution of Included Studies		
Europe (Italy, EU-level, Romania, Bulgaria, Ukraine, etc.)	15	22.1%
Asia (China, Indonesia, Malaysia, Thailand, Kazakhstan, Iran, India, Yemen, Taiwan)	17	25.0%
North America (United States)	11	16.2%
Sub-Saharan Africa (Rwanda, Zambia, Botswana, Tanzania, Ethiopia)	7	10.3%
Latin America (Colombia, Brazil)	2	2.9%
Middle East (Iraq)	1	1.5%
Global/Multi-regional/Jurisdiction-unspecific	15	22.1%
Total	68	100%
G. Relevance to Systematic Review on AI in Tax Compliance		
Directly relevant: empirical ML/AI applied to tax fraud, evasion, or compliance detection	35	51.5%
Directly relevant: governance, legal, or taxpayer rights analysis of AI in tax admin	15	22.1%
Directly relevant: behavioral/NLP/sentiment analysis of taxpayer compliance	8	11.8%
Indirectly relevant: blockchain, digital transformation, Tax Administration 3.0	3	4.4%
Indirectly relevant: institutional readiness, policy framework for AI (no direct deployment)	7	10.3%

2.4. Data Extraction and CIMO Coding

A structured data extraction matrix operationalized CIMO into four coding dimensions: (1) Context, geographic setting, institutional maturity, tax type, and structural barriers; (2) Intervention, AI technology type, specific algorithms, and data sources; (3) Mechanism, technical pathways, institutional processes, and behavioral mediators; (4) Outcome, technical performance metrics, fiscal impacts, and institutional effects. Both authors coded all 68 studies iteratively, resolving disagreements by consensus; formal inter-rater reliability statistics were not employed, a limitation acknowledged in Section 5.3. Operational definitions and illustrative examples were established before extraction, and audit trails were maintained throughout. The complete CIMO extraction matrix is provided in Supplementary Materials Table S2.

3. Results (1): Descriptive and Textometric Analysis

3.1. Cluster Structure and Spatial Distribution

It is worth noting that the distribution of the clusters returned by the Descending Hierarchical Classification was as follows: Cluster 3 (Machine Learning and Predictive Modeling) at 42.61% (150 out of 352 classified segments), Cluster 1 (Artificial Intelligence, Technology and Tax Compliance) at 33.81% (119 out of 352 segments) and Cluster 2 (Government, Financial Development and Revenue Administration) at 23.58% (83 out of 352 segments).

The three thematic clusters and lexical proximities can be visualized in the dendrogram in Figure 4. Cluster 3 separates first as the most distinct thematic cluster, followed by the separation between Cluster 1 and Cluster 2. The factorial correspondence analysis shown in Figure 5 confirms the spatial separation of the three paradigms: Cluster 3 terms (learn, model, predict, algorithm, accuracy) occupy one pole, Cluster 1 terms (intelligence, technology, compliance, blockchain, automation) occupy a second pole, while Cluster 2 terms (government, development, financial, country, GDP) occupy a third pole. The spatial proximity between Cluster 1 and Cluster 2 indicates convergence between governance and developmental perspectives, while the distance of Cluster 3 reflects the continued methodological distinctiveness of technical performance studies.

3.2. Thematic Content of the Three Clusters

3.2.1. Cluster 3—Machine Learning and Predictive Modelling (42.61%)

In Cluster 3 (42.61%) the core focus is on algorithmic performance and methodological innovation. Battaglini et al. (2025) challenge traditional experiential audits by finding that ML-selected audits in Italy detected 39% more tax evasion than human-selected audits. On Chinese enterprise data, Yang (2025) uses random forest models to reach an accuracy of more than 92%, and to determine sector-specific risk indicators (tax burden rate, profit fluctuation). Malashin et al. (2025) optimize XGBoost using a multi-objective optimization technique (NSGA-II) with focal loss and achieve an ROC AUC of 0.9417. For the context of limited resources, Kamoun et al. (2025) compare supervised learning algorithms for predicting tax non-compliance in Tunisia, and they find that the random forest algorithm achieves an accuracy of 83%. Xavier et al. (2022) illustrate that, in Brazil, an open data collection and a neural network can correctly detect tax evasion at more than 98% accuracy, which means that public repositories can substitute for proprietary administrative information.

Advanced architectures address data constraints through synthetic data generation (GAN + SMOTE achieving ~0.95 accuracy on Thai evasion tasks; Visitpanya & Samanchuen, 2023), ensemble methods combining GANs and multiple classifiers (Alrasheedi et al., 2025), metaheuristic-optimized ML (IPSO + MLP + SVM, 93.68%, Iran; Mojahedi et al., 2022), and

multi-label classification (MLARAM, F1 = 95.5%, Indonesia; [Prasetyo et al., 2024](#)). [Hauser et al. \(2025\)](#) demonstrate that ML-targeted behavioral nudges nearly double compliance effects versus untargeted approaches, and [Coita et al. \(2023\)](#) show that negative taxpayer linguistic patterns from NLP analysis predict future non-compliance. Additional contributions document double/debiased ML for causal inference ([Chang, 2020](#)), predictive analytics for income tax evasion ([Babu & Vasavi, 2017](#)), undeclared work drivers across EU member states ([Franic, 2022](#)), NAICS code prediction via supervised methods ([Oehlert et al., 2022](#)), and a comprehensive review of AI fraud detection trends and implications ([Belahouaoui & Alm, 2025](#)).

This group establishes technical feasibility in a variety of settings. But accuracy rates (83–98%) depend on data availability, infrastructure development, and the regulatory framework. Technical performance is context-specific and not generally transferable.

3.2.2. Cluster 1—Artificial Intelligence, Technology and Tax Compliance (33.81%)

Cluster 1 (33.81%) focuses on the institutional, governance, and technological aspects of AI use. The key lexical patterns (*intelligence, artificial, AI, technology, taxation, framework, legal, compliance, noncompliance, blockchain, automation, adoption*) highlight issues related to systemic integration, regulatory frameworks, and organizational conditions that can enable or constrain deployment.

This cluster is underpinned by governance and ethical frameworks. [Bogucki \(2025\)](#) discusses the ELSA/RRI aspects of AI in EU tax administration, whereas [Belahouaoui and Attak \(2024\)](#) validate the Tax Administration 3.0 trends and point to regulatory gaps in emerging economies, and [Mudome and Ndiege \(2025\)](#) emphasize ethical and transparency concerns in AI deployment in developing economies. [Calderón Carrero and Ribeiro \(2020\)](#) and [Bevacqua \(2020\)](#) caution that the lack of legal protection in automated enforcement in both civil- and common-law jurisdictions may compromise the rights of taxpayers. Adoption readiness studies show TOE-based success factors in Botswana ([Munjeyi & Schutte, 2024, 2025](#)), regulatory and human-resource gaps in Indonesia ([Saragih et al., 2023](#)), and that long-term investment in e-government is central in Estonia ([Künnapas et al., 2025](#)). In contexts with limited data capacity, the main constraint is lack of data rather than algorithmic sophistication: estimated tax gaps in Zambia range from 47% to 55% due to data limitations and evasion cycles ([Adu-Ababio et al., 2025](#)), while the VAT gap in Tanzania reaches 62% ([Ebrahim et al., 2024](#)), and even partial digitization in Rwanda is already said to allow for ML-based tax-gap forecasting ([Uwizeye, 2025](#)).

Complementary studies examine the impact of AI-driven economic models on tax law concepts ([Fidelangeli & Galli, 2021](#)) and the general effect of digital economic transformation on tax governance ([Suwarni et al., 2020](#)). The combined effect of this cluster highlights a methodological transition from technocentric conceptualizations to socio-technical integration, and the importance of governance and institutional aspects is increasingly recognized as central to the effectiveness of AI.

3.2.3. Cluster 2—Government, Financial Development and Revenue Administration (23.58%)

Cluster 2 (23.58% of the corpus) covers macroeconomic, institutional, and cross-border dimensions of fiscal governance. The key lexical elements (*government, development, financial, domestic, GDP, money laundering, privacy, state*) illustrate themes related to the capacity to mobilize revenues, the sophistication of the regulatory framework and the policy design. Representative studies include unsupervised ML for tax avoidance by overseas Americans ([Miller, 2019](#)), money laundering risk indicators in Bulgaria ([Kostova & Zhelev, 2024](#)), social media analysis of taxpayer attitudes in Kazakhstan ([Kuandykov et al., 2021](#)), AI-driven taxation policy design using deep multi-agent reinforcement learning ([Zheng et al., 2022](#)),

challenges of asset valuation for tax compliance (Soled & Thomas, 2023), due-process risks from Italy's Ve.R.A. algorithm (Saponaro, 2025), and a bibliometric and engineering overview of the use of AI and ML in accounting and auditing (Güney et al., 2025). The cluster shows that the effectiveness of algorithms is tightly coupled to structural elements of fiscal governance, revenue mobilization capacity, regulatory sophistication, and cross-border fiscal coordination. The relatively small proportion of this component (23.58%) indicates that the macroeconomic and developmental aspects of AI-tax compliance are not yet fully explored compared to technical and governance aspects, warranting further empirical study.

3.3. Cross-Cluster Synthesis and Temporal Dynamics

The three clusters represent complementary and temporally layered research trajectories. Cluster 3 established proof-of-concept for algorithmic detection accuracy; Cluster 1 emerged as a corrective emphasizing that governance frameworks are constitutive of intervention effectiveness, not peripheral to it; and Cluster 2 provides the macroeconomic and developmental context that preconditions what is institutionally feasible. The factorial correspondence analysis spatially confirms these three poles, with Cluster 1 and Cluster 2 in closer proximity, reflecting the growing integration of governance and developmental perspectives, and Cluster 3 more distant, reflecting the continued methodological distinctiveness of technical performance studies. Three propositions emerge as theoretical patterns (not causal claims): a methodological hierarchy in which socio-technical perspectives gain weight over time; a geographical imbalance in which technical studies are dominated by advanced-economy data while governance and adoption studies are increasingly present in developing-economy research; and a temporal evolution from single-cluster studies (2015–2019) toward more integrated, cross-cluster designs (2022–2025). These patterns serve as the analytical foundation for the CIMO-based synthesis in Section 4.

4. Results (2): CIMO-Based Synthesis

This section offers a systematic overview of the 68 studies, structured according to the four analytical areas of the CIMO framework: Context, Intervention, Mechanism and Outcome. Instead of analyzing individual studies in isolation, the analysis aims to detect patterns, boundary conditions, and contradictions between studies that help to explain how similar AI technologies can lead to different levels of compliance in different institutional contexts. The synthesis process ends with the Adaptive AI Tax Compliance Framework (AAITCF), which transforms these analytical patterns into differentiated implementation paths, calibrated to institutional maturity level. The CIMO framework of the studies used for analysis is provided in Table S2 in the Supplementary Materials.

4.1. Context (C): Institutional and Structural Conditions Shaping AI Deployment

In CIMO logic, context is not residual but constitutive. It determines which interventions are possible, how mechanisms work, and what outcomes can be expected. Three dimensions of context, all related to structure, are found consistently to be active factors in the effectiveness of interventions across the 68 studies.

4.1.1. Geographic and Institutional Settings

The 68 studies span seven geographic clusters: Asia ($n = 17$; 25.0%), Europe ($n = 15$; 22.1%), Global or multi-regional ($n = 15$; 22.1%), North America ($n = 11$; 16.2%), Sub-Saharan Africa ($n = 7$; 10.3%), Latin America ($n = 2$; 2.9%), and the Middle East ($n = 1$; 1.5%). This distribution is analytically interesting: European and Asian studies dominate the literature on ML performance, while Sub-Saharan African studies disproportionately

document institutional barriers and readiness gaps. The AAITCF is designed in tiers, and these asymmetries directly inform the tiering of the design.

In advanced economies (Italy, Estonia, US), research sets the upper bound for AI-based compliance: In Italy, ML-selected audits detected 39% more cases of evasion and recovered 29% more revenue than human-selected alternatives (Battaglini et al., 2025); in Estonia, multi-application AI deployment (language models, fraud-detection networks, customs imaging) reflects decades of investment in e-government (Künnapas et al., 2025); and in the US, a randomized field experiment demonstrates that ML-targeted behavioral nudges nearly double compliance effects compared to untargeted nudges (Hauser et al., 2025). In contrast, developing and emerging economy studies (Sub-Saharan Africa) show the opposite side: tax gaps of 47–62% (Adu-Ababio et al., 2025; Ebrahim et al., 2024), the non-existence or underdeveloped nature of AI governance structures (Munjeji & Schutte, 2024), and infrastructure gaps that restrict algorithmic training data, even if partial digitization in Rwanda has already enabled the use of ML for revenue forecasting (Uwizeye, 2025). Asian studies (China, Indonesia, Malaysia, Thailand, Iran, Russia, Kazakhstan) fall in the middle range—between 87% and 94% accuracy, using standard supervised methods (Yang, 2025; Prasetyo et al., 2024; Mojahedi et al., 2022); data confidentiality is addressed by synthetic generation (Visitpanya & Samanchuen, 2023); and the readiness gaps are mostly concentrated in regulation and human capital, not in transactional data (Saragih et al., 2023). European and Latin American studies emphasize behavioral and structural determinants: religiosity and institutional trust influence tax morale independently of technology (Davidescu et al., 2022); sentiment in taxpayer language predicts future non-compliance (Coita et al., 2023); and fiscal incentive design causally shapes evasion behavior (Saavedra & Romero, 2021).

4.1.2. Thematic and Regulatory Framing

Studies are distributed across four thematic contexts. The largest group targets tax fraud and evasion directly: VAT underreporting (Tuyishimire & Murorunkwere, 2024; Ebrahim et al., 2024), income tax fraud (Mojahedi et al., 2022; Malashin et al., 2025; Siimon & Lukason, 2021), corporate tax avoidance and tax haven use (Borrotti et al., 2023), self-employment underreporting (Battaglini et al., 2025; Battiston et al., 2024), and satellite-identified illegal mining evasion (Saavedra & Romero, 2021). A second cluster concerns the digital transformation of tax administration and Tax Administration 3.0 (Belahouaoui & Attak, 2024; Karataş Durmu & Arıtı Erdem, 2023; Pylypenko et al., 2022). A third context addresses behavioral and sociological determinants of compliance, including taxpayer trust, morale, religiosity, and sentiment (Davidescu et al., 2022; Coita et al., 2023; Puklavec et al., 2023; Mu et al., 2023). A fourth context covers governance, legal accountability, and taxpayer rights in AI-driven enforcement (Bogucki, 2025; Bevacqua, 2020; Calderón Carrero & Ribeiro, 2020; Saponaro, 2025; Al-zi'aylee et al., 2025).

4.1.3. Structural Barriers and Enabling Conditions

Five structural barriers recur consistently. Data scarcity and confidentiality constraints affect 19 studies (27.9%), limiting labeled training datasets, and are handled by using synthetic data generated with GANs in Thailand (Visitpanya & Samanchuen, 2023) and a Correlational GAN ensemble for financial statement data (Alrasheedi et al., 2025). Class imbalance affects 14 studies (20.6%), where non-compliant taxpayers are a small minority, and is addressed by SMOTE (Chou et al., 2025), focal loss (Malashin et al., 2025), and positive-unlabeled learning (Gao et al., 2023). All 7 Sub-Saharan African studies document fragmented or non-digitized administrative infrastructure as a key feasibility constraint. Ten legal and governance studies highlight regulatory uncertainty, such as the lack of

clarity in data protection laws, unclear accountability requirements for algorithms, and the uncertain application of the EU AI Act to tax administration. In 8 studies across Botswana, Zambia, Ethiopia, and Indonesia, human capital gaps are reported, such as inadequate training of tax staff in data science.

Conversely, the most consistent enablers of effective AI deployment are large administrative datasets (23 studies; 33.8%), digital invoicing infrastructure, and institutional openness to technological innovation. The long-term investment in e-government infrastructure in Estonia (Künnapas et al., 2025; Siimon & Lukason, 2021) is the classic positive example, while Saragih et al. (2023) illustrate that, even without comprehensive regulation, existing transactional data in Indonesia is sufficient for AI use, and Uwizeye (2025) demonstrates how partial digitization in Rwanda can already support ML-based revenue forecasting.

4.2. Intervention (I): AI Architectures Applied Across Tax Compliance Contexts

In the 68 studies, four main categories of interventions can be observed: supervised ML for fraud detection and audit selection; ensemble methods, data augmentation, and network-based architecture; NLP, behavioral interventions, and policy-oriented AI; and legal-governance frameworks as non-algorithmic interventions.

4.2.1. Supervised Machine Learning for Fraud Detection and Audit Selection

Supervised ML is the most widely used intervention (32 of 68 studies; 47.1%), with Random Forest, XGBoost, and Gradient Boosting being the most commonly used algorithms. These algorithms have been found to achieve between 83% and 98% accuracy with sufficient administrative data across various national datasets, with performance differences being dependent on data volume and infrastructure, not algorithm choice (Kamoun et al., 2025; Yang, 2025; Wahab & Abu Bakar, 2021; Xavier et al., 2022; Siimon & Lukason, 2021).

Multi-label and multi-objective classification represent a new methodological step. Prasetyo et al. (2024) apply MLARAM, MLkNN, binary relevance, label powerset, and RAKELd to Indonesian tax risk data, with MLARAM achieving $F1 = 95.5\%$, demonstrating that multi-objective risk profiling captures more informative compliance dimensions than binary classification. Malashin et al. (2025) address the accuracy-interpretability trade-off directly: their NSGA-II multi-objective optimization simultaneously maximizes ROC AUC and minimizes decision tree depth, producing a Pareto-optimal model that is both more accurate (0.9417 vs. 0.9161) and more interpretable than the non-optimized baseline.

Audit selection using ML is documented in two Italian studies with direct fiscal consequences. Battaglini et al. (2025) show that replacing the 10% least promising human-selected audit cases with ML-selected cases increases detected evasion by 39% and recovered revenue by 29% without increasing audit volume. Battiston et al. (2024) report parallel gains for self-employed taxpayers and sole proprietors using regional Random Forest models. Rabasco and Battiston (2023) complement these findings by showing that ML-targeted audits identify taxpayers who increase declared income by an average of €14,461, versus an average decrease of €205 for all audited taxpayers. Borrotti et al. (2023) extend audit targeting to the corporate tax avoidance domain: using publicly available accounting data for 4031 EU group ultimate owners, they predict aggressive tax haven location decisions ex ante, providing a basis for preventive administrative approaches.

4.2.2. Ensemble Methods, Data Augmentation, and Network-Based Interventions

Ensemble learning and synthetic data generation address the twin constraints of model instability and dataset scarcity. Alrasheedi et al. (2025) propose a soft voting ensemble combining a Correlational GAN, an encoder, and multiple classifiers for financial statement fraud detection, outperforming all non-ensemble configurations. Chou et al. (2025) combine SMOTE with a Forensic-Based Investigation (FBI) metaheuristic algorithm

and ensemble ML to detect fraudulent pollution fee declarations in Taiwan, extending AI-driven tax compliance into environmental fiscal governance. [Visitpanya and Samanchuen \(2023\)](#) generate synthetic financial data using GAN+SMOTE (quality score 0.87), training models that achieve ~0.95 accuracy and 0.91 F1 on real Thai tax evasion tasks, directly circumventing the confidentiality barrier preventing direct access to tax returns.

Network-based interventions analyze relational structures among taxpayers and entities. [Didimo et al. \(2020\)](#) present MALDIVE, combining social network analysis with ML risk scoring, information diffusion modeling, and interactive human validation for Italy's Revenue Agency, as the dataset's most complete integration of technical and human judgment. [Gao et al. \(2023\)](#) develop FBNE-PU, which combines taxpayer attributes and graph embeddings from PnCGCN with positive-unlabeled learning to handle the labeling scarcity problem intrinsic to fraud detection. [Lee et al. \(2024\)](#) combine panel data modelling and agent-based simulation to demonstrate that blockchain technology integration can address non-cooperative taxpayer behavior and reduce the US tax gap when paired with improvements in IRS audit efficiency and increased penalties. [Zheng et al. \(2022\)](#) push simulation further with the AI Economist: a two-level deep multi-agent reinforcement learning system that optimizes tax policy for social welfare under inequality and productivity constraints. [Warner et al. \(2015\)](#) apply an evolutionary algorithm to generate novel tax shelter configurations, providing auditors with a tool for proactive scenario-planning beyond known patterns. [Shujaaddeen et al. \(2024\)](#) deploy a hybrid supervised and unsupervised neural network for tax fraud detection in Yemen, one of the few studies from a conflict-affected, low-capacity administrative context.

4.2.3. NLP, Behavioral Interventions, and Policy-Oriented AI

Natural language processing interventions appear in five studies. [Puklavek et al. \(2023\)](#) apply structural topic modeling to 1.16 million tax-related tweets (2010–2020), finding that enforcement content generates 38–40% more retweets than information-service content and carries significantly more negative sentiment. [Coita et al. \(2023\)](#) establish via NLP sentiment analysis that negative word frequency in taxpayer language predicts future non-compliant behavior. [Kuandykov et al. \(2021\)](#) deploy social media analytics in Kazakhstan to map taxpayer attitudes; [Miller \(2019\)](#) applies text mining to overseas American FATCA/FBAR compliance discourse; and [Antinozzi and Cooper \(2025\)](#) find that ChatGPT 3.5 and 4.0 answer only 39–47% of common US tax questions correctly, concluding that generative AI chatbots require substantial human oversight before deployment as taxpayer guidance tools.

Behavioral interventions with algorithmic targeting provide the most causally robust evidence in the dataset. [Hauser et al. \(2025\)](#) design a randomized field experiment in which ML identifies unemployment benefit claimants at high risk of fraudulent underreporting; targeted behavioral nudges produce near-zero average treatment effects without ML targeting but compliance effects nearly double with ML identification, establishing the marginal causal value of algorithmic selection within behavioral interventions. [Mu et al. \(2023\)](#) identify psychological egoism as a mediator between AI-based invoicing technology and tax revenues in Ethiopia, suggesting that individual utility calculus moderates the compliance effect of digitization through an attitudinal pathway. [Axtell and Farmer \(2025\)](#) review agent-based modeling methods for behavioral simulation in economic and tax contexts, providing a methodological foundation for future behavioral AI tax research.

4.2.4. Legal-Governance Frameworks as Institutional Interventions

Fifteen studies (22.1%) treat legal, governance, and institutional frameworks as interventions per se. [Bogucki \(2025\)](#) proposes ELSA/RRI-based governance standards for EU

tax AI, including DPIA requirements, AI-use inventories, human-in-the-loop protocols, and GDPR/EU AI Act compliance. Bevacqua (2020) and Calderón Carrero and Ribeiro (2020) document complementary accountability gaps between common-law and civil-law jurisdictions, noting that taxpayer rights to reasons, challenge, and appeal are structurally undermined without algorithmic accountability frameworks embedded in tax procedure law. The tension in practice is illustrated by Saponaro (2025): Italy's Ve.R.A. requires human sign-off for enforcement, but its learning logic is not disclosed, which is an unresolved due-process gap. Based on 62 papers, Belahouaoui and Attak (2024) find that Tax Administration 3.0 is moving forward all over the world, although governance and capacity gaps remain in emerging economies.

4.3. Mechanism (M): Pathways from Algorithmic Deployment to Compliance Outcomes

Mechanisms explain how interventions turn into outcomes. There are four types of mechanisms that can be distinguished: statistical pattern recognition and risk prioritization; explainability and human-in-the-loop governance; behavioral and socio-psychological mediation; and automation with real-time data integration.

4.3.1. Pattern Recognition, Risk Scoring, and Audit Prioritization

The essence of supervised learning is statistical pattern recognition: models developed from previous compliance information calculate a risk score for each case, which guides audit resources to high-yield cases. The Italian studies (Battaglini et al., 2025; Battiston et al., 2024; Rabasco & Battiston, 2023), which reallocated a fixed audit budget to ML-selected cases, offer the strongest evidence of measurable revenue gains without an increase in enforcement capacity. Feature-importance analyses also support highly sector-specific risk profiles: the features that are important for Chinese manufacturing are distinct from those of real estate and Tunisian taxpayers, meaning that the model needs to be locally tuned for each sector or jurisdiction, which helps explain cross-country accuracy variation (83% Tunisia vs. >98% Brazil). Relational pattern recognition takes detection a step further: MALDIVE (Didimo et al., 2020) spreads risk scores through commercial networks to detect indirect participation in fraud; FBNE-PU (Gao et al., 2023) uses positive-unlabeled learning for graph embeddings to deal with the challenge that unlabeled cases are not necessarily compliant; and Borrotti et al. (2023) use predicted tax-haven probabilities as anticipatory signals to be able to act in advance.

4.3.2. Explainability, Human-in-the-Loop, and Algorithmic Accountability

Twelve studies (17.6%) treat explainability as essential to institutional legitimacy. Maldive (Didimo et al., 2020) implements interactive human validation as an explicit architectural layer; Ve.R.A. (Saponaro, 2025) mandates human sign-off before enforcement despite undisclosed learning logic; and Kamoun et al. (2025) deploy SHAP values to produce interpretable ranked risk attributes for auditors. Bogucki (2025) grounds these requirements in EU law: tax AI systems must offer meaningful explanations under GDPR Article 22, conduct DPIAs under Article 35, and perform fundamental-rights impact assessments under EU AI Act Article 27. However, technical explainability alone may be insufficient; taxpayer rights to reasons, challenge, and appeal require redesigned enforcement processes, yet only 3 of 68 studies explicitly address this legal redesign (Bogucki, 2025; Saponaro, 2025; Al-zi'aylee et al., 2025).

4.3.3. Behavioral and Socio-Psychological Mediation Mechanisms

Several studies have shown that AI does not directly enforce compliance but rather operates through socio-psychological mediating pathways. This is shown experimentally by Hauser et al. (2025): if the behavioral nudge is randomly distributed, it has almost no

impact, but if it is targeted by ML risk scoring, it almost doubles compliance, and the mechanism is not persuasion, but the selection of the subpopulation for which persuasion works. The results of the study conducted by [Mu et al. \(2023\)](#) indicate that psychological egoism plays a mediating role between the adoption of AI-based invoicing and tax revenue in Ethiopia, suggesting that people with self-interest orientations react differently to digitization than those who are prosocially oriented. In Thai e-commerce, [Thanakijkhajonphaisan and Sincharoonsak \(2025\)](#) propose a multi-layered mechanism in which social motivations influence taxpayer compliance intentions, which then influence how taxpayers respond to technological monitoring and burden minimization.

[Davidescu et al. \(2022\)](#) provide evidence at the macro level showing that religiosity and institutional trust significantly affect tax morale on their own and at times more than technological interventions in 79 countries. [Lazovic Pita et al. \(2025\)](#) find that, beyond economic variables, social and institutional determinants continue to have explanatory power in 35 European countries in ML models. The results reveal that the background mechanisms of cultural and institutional embeddedness can either limit or enhance the ability of AI tools to achieve their intended goals, with the same algorithm operating differently in contexts of high trust and high morale versus low trust and conflict settings (cf. [Shujaaddeen et al., 2024](#), Yemen). [Coita et al. \(2023\)](#) find that negative sentiment can be used as an early warning system, as NLP analysis of taxpayer language gives a predictive signal of future non-compliance before observable behavioral deviation.

4.3.4. Automation and Real-Time Data Integration

Automation mechanisms operate at the level of administrative infrastructure. In Rwanda, [Uwizeye \(2025\)](#) finds that the digitization of Electronic Billing Machine (EBM) systems leads to the generation of real-time transaction data that can be used for ML-based revenue forecasting. However, the adoption of EBM itself is not shown to have a direct impact on compliance, as implementation and enforcement gaps still exist. [Neuman and Sheu \(2022\)](#) report a moderated automation mechanism: Big Data analytics in IRS audit selection significantly increases compliance for taxpayers with high digital visibility (online advertisers) but has no significant compliance effect for taxpayers without such visibility, while also reducing perceived fairness for the latter group. In short, automation can exacerbate existing inequalities in tax enforcement.

The analysis of [Topchii et al. \(2025\)](#) shows that digital infrastructure, institutional quality and economic size are more significant determinants of digital tax compliance than the type of AI tools used, highlighting the contextual conditionality of the mechanisms of automation. In Romania, [Cuc et al. \(2025\)](#) show that technology adoption for tax compliance purposes depends on perceived usefulness and trust in institutional digital systems, and not only on the technical features of the tools.

4.4. Outcome (O): Technical Performance, Fiscal Impact, and Institutional Consequences

Outcomes encompass technical performance, fiscal and revenue effects, and institutional, legal, and ethical implications. Methodologically, it is essential to distinguish these dimensions: accuracy rates, tax gap estimates, revenue recovery figures, time savings, and fairness assessments are not equivalent measures, so treating them as equivalent will result in misleading generalizations about the effectiveness of AI.

4.4.1. Technical Performance Outcomes

In the 32 empirical ML studies, the accuracy of the models falls between 83% and over 98%, and four analytical patterns are common. First, high performance is context-dependent and not a transferable benchmark—it is not about the algorithm used, but about the amount of data and the quality of the infrastructure. Second, multi-objective

and multi-label architectures consistently beat single-algorithm binary classifiers (e.g., MLARAM in Indonesia, F1 = 95.5% vs. binary baselines; XGBoost+NSGA-II in Russia, ROC AUC 0.9417 vs. 0.9161 baseline). Third, synthetic data generation using GAN+SMOTE manages to achieve good performance (accuracy of ~0.95 for real Thai evasion tasks) while still addressing confidentiality constraints. Fourth, unsupervised and positive-unlabeled methods (e.g., FBNE-PU) address the basic problem of scarce labels in fraud detection. Under class imbalance, the most useful comparative measure is ROC AUC, and interpretability analysis via SHAP gives the audit-prioritization intelligence needed for administrative deployment.

4.4.2. Fiscal and Revenue Outcomes

Quantified fiscal outcomes are reported in 9 studies and reveal substantial material consequences where institutional conditions enable AI deployment. The most rigorous evidence comes from Italy: [Battaglini et al. \(2025\)](#) find that replacing the 10% least promising human-selected audits with ML-selected cases increases detected evasion by 39% and recovered revenue by 29% without increasing audit volume. [Rabasco and Battiston \(2023\)](#) estimate that ML-targeted taxpayers increase declared income by an average of €14,461, versus an average decrease of €205 for all audited taxpayers, a striking differential demonstrating that algorithmic selection captures heterogeneous deterrence responses that are invisible to non-targeted approaches. [Battiston et al. \(2024\)](#) document additional revenue gains from Random Forest models applied to Italian regional self-employment data.

Tax gap estimation provides another fiscal outcome dimension. [Adu-Ababio et al. \(2025\)](#) estimate Zambia's tax gap at 55% (regression) and 47% (ML), driven primarily by Corporate Income Tax losses in the extractives sector. [Ebrahim et al. \(2024\)](#) estimate a 62% VAT gap for Tanzanian SMEs, with the highest evasion among firms systematically avoiding audit exposure. [Tuyishimire and Murorunkwere \(2024\)](#) identify VAT underreporting risk profiles in Rwanda by sector and taxpayer characteristics, providing actionable prioritization intelligence for the Rwanda Revenue Authority. [Borrotti et al. \(2023\)](#) demonstrate that ML-predicted aggressive tax haven location probabilities, when built into preventive administrative approaches, reduce the welfare costs of anti-avoidance uncertainty. [Galler and Soled \(2024\)](#) argue that AI applied to PTIN-registered preparer data could recover billions annually in the US by identifying derelict practitioners whose returns systematically understate taxable income. [Saavedra and Romero \(2021\)](#) establish that fiscal incentive design directly drives evasion: reducing revenue transfers to mining municipalities in Colombia increases illegal mining by a causal margin, suggesting that AI-detected evasion must be understood within the fiscal contract between jurisdictions.

4.4.3. Institutional, Legal, and Ethical Outcomes

Institutional and ethical outcomes are documented in 14 governance, legal, and cross-cutting studies. Four institutional risk categories recur. First, erosion of taxpayer rights through opaque automated enforcement: both [Calderón Carrero and Ribeiro \(2020\)](#) and [Bevacqua \(2020\)](#) identify the right to reasons, challenge, and appeal as structurally undermined by algorithmic enforcement across civil- and common-law jurisdictions. Second, due-process opacity: Ve.R.A.'s learning logic remains undisclosed despite mandatory human sign-off ([Saponaro, 2025](#)), and ChatGPT answers only 39–47% of common US tax questions correctly ([Antinozzi & Cooper, 2025](#)), suggesting that generative AI may require disclosure requirements before deployment as taxpayer guidance. Third, algorithmic bias risk: [Bogucki \(2025\)](#) identifies selection bias in administrative datasets as a systematic risk producing discriminatory risk scoring if unaudited. Fourth, structural inequality in AI benefits: AI improves outcomes in advanced economies but produces limited or negative

effects in low-infrastructure settings (Mudome & Ndiege, 2025; Belahouaoui & Attak, 2024), meaning fiscal gains are unevenly distributed across and within countries. On the constructive side, AI nudges that are targeted at taxpayers achieve a near-doubling of voluntary compliance without punitive measures (Hauser et al., 2025), Estonia's implementation provides spillovers in service quality and staff capacity (Künnapas et al., 2025), proactive governance under ELSA-RRI can build public trust while maintaining efficiency gains (Bogucki, 2025), AI improves GST implementation outcomes in India (Mathew & Devan, 2025), a proactive and ethically grounded AI strategy is advocated for tax administrations globally (Antón, 2021), and qualitative research with US/Canadian tax professionals reveals that AI is enthusiastically adopted in compliance work but is professionally resisted in advisory services (Radcliffe et al., 2025); that is, the institutional effects are differentiated by function, not uniform across the profession.

4.4.4. Comparative Synthesis: Convergences, Divergences, and Limitations Across the Evidence Base

The 68 studies converge on three points and diverge sharply on two others; each pattern carries a direct implication for the design of the AAITCF (Section 4.5).

Convergence 1: reported accuracy varied more from context to context. Despite spanning Random Forest, XGBoost, neural networks, and ensemble or graph-based methods, reported accuracies cluster within an 83–98% band whenever administrative data are adequate (Kamoun et al., 2025; Yang, 2025; Xavier et al., 2022; Siimon & Lukason, 2021), largely irrespective of which specific algorithm is used. The binding constraint is consistently data volume, quality, and infrastructure (Topchii et al., 2025; Saragih et al., 2023), not algorithmic sophistication. This convergence directly justifies the AAITCF's Step 1 emphasis on contextual and data readiness assessment ahead of technology selection (Section 4.5.1).

Convergence 2: technical performance and institutional legitimacy are reported by almost entirely non-overlapping sets of studies. The 32 empirical ML studies rarely discuss governance, bias, or taxpayer rights (only 24 of 68 studies, 35.3%, address this per Table 5, Panel C), while the 14 governance and legal studies rarely report technical performance metrics. This bifurcation is itself a finding: the corpus exhibits a structural division of labor between technical and governance scholarship that no single study bridges, which is precisely the gap the AAITCF's six-step cycle is designed to close by forcing technical components (Steps 2–3) and governance components (Steps 4 and 6) into a single iterative structure.

Convergence 3: where causal or quasi-causal designs exist, the targeting or selection mechanism, not the intervention content, drives the compliance gain. Hauser et al.'s (2025) randomized experiment and Battaglini et al.'s (2025) quasi-experimental audit reallocation both show that algorithmically selecting who is treated, rather than what they are given, produces the measured gain. This anchors AAITCF Steps 3 and 5.

Divergence 1: identical model classes produce very different outcomes across institutional settings, and the studies disagree on why. Random Forest reaches 83% accuracy in Tunisia (Kamoun et al., 2025) and over 98% in Brazil (Xavier et al., 2022); some authors attribute such gaps to data volume (Siimon & Lukason, 2021), others to institutional trust and tax morale (Davidescu et al., 2022; Lazovic Pita et al., 2025), and others to digital infrastructure and economic size (Topchii et al., 2025). The corpus does not resolve which moderator dominates, and no single study tests them jointly. This unresolved divergence is precisely why the AAITCF treats the institutional tier as a multi-dimensional readiness profile rather than a single composite score (Section 4.5.1), and it is flagged in Section 5.4 as a priority for future multi-moderator empirical testing.

Divergence 2: the studies disagree, implicitly, on whether automation reduces or reproduces inequality. Hauser et al. (2025) and Künnapas et al. (2025) report automation

improving compliance or service quality without apparent distributional cost, whereas Neuman and Sheu (2022) find that Big Data audit selection increases compliance only for taxpayers with high digital visibility while reducing perceived fairness for the remainder, and Mudome and Ndiege (2025) document an analogous asymmetry across, rather than within, countries. Because both directions are evidenced within the corpus rather than resting on a single outlier study, the AAITCF treats equity monitoring as a mandatory, recurring component of Step 6 rather than a one-time ethical checkbox.

A limitation qualifies the whole synthesis: none of these convergences or divergences can be adjudicated causally with the available evidence. Only one randomized experiment (Hauser et al., 2025) and one quasi-experimental design (Battaglini et al., 2025) exist in the entire corpus (Section 2.3), so the comparative patterns above describe theoretically coherent associations rather than established causal mechanisms. This is why the AAITCF (Section 4.5) is presented as a provisional, evidence-grounded roadmap requiring longitudinal validation (Section 5.4), rather than as a causally validated intervention protocol.

4.4.5. From CIMO Evidence to Framework Design

The AAITCF framework was synthesized from the evidence base; each of its six steps was derived from a specific, recurring CIMO pattern documented in Sections 4.1–4.4 and 4.4.4 and is traceable to the studies that motivated it. Table 6 makes this derivation explicit, mapping each step to (a) the CIMO evidence pattern that motivated it, (b) the primary supporting studies, and (c) the quality-appraisal caveat (Section 2.3, Table 5) that qualifies the strength of that evidence. Where the underlying evidence is causally strong (Step 5), the corresponding framework recommendation is directive; where the evidence reflects a documented gap rather than a validated solution (Step 6), the recommendation is corrective and provisional, pending the empirical validation called for in Section 5.4.

Table 6. Derivation of the AAITCF’s Six Steps from CIMO Evidence Patterns.

AAITCF Step	CIMO Evidence Pattern	Primary Supporting Studies	Quality-Appraisal Caveat
1. Contextual Readiness & Data Infrastructure Assessment	Data quality/infrastructure predicts outcomes more than algorithm choice in Tier 3 contexts (Convergence 1, Section 4.4.4)	Adu-Ababio et al. (2025); Uwizeye (2025); Saragih et al. (2023); Munjeyi and Schutte (2024)	63.2% single-country scope limits transferability of readiness thresholds (Table 5, Panel E)
2. Pattern Recognition and Anomaly Detection	RF/XGBoost achieves comparable accuracy across very different data volumes, supporting them as a low-data-requirement baseline (Convergence 1)	Kamoun et al. (2025); Yang (2025); Siimon and Lukason (2021); Didimo et al. (2020)	Accuracy benchmarks are context-specific, not transferable performance standards (Section 2.3 implications)
3. Predictive Risk Assessment and Audit Prioritization	Feature importance is sector-specific, so risk scoring must be calibrated locally rather than imported wholesale (Divergence 1)	Yang (2025); Xu and Kong (2024); Tuyishimire and Murorunkwere (2024); Kamoun et al. (2025)	Sector-specificity itself is the caveat: no cross-sector transfer has been tested
4. Explainability and Human-in-the-Loop Validation	Technical explainability alone is insufficient without redesigned due-process channels; MALDIVE/Ve.R.A. show human-in-the-loop is operationally feasible (Convergence 2)	Bogucki (2025); Bevacqua (2020); Calderón Carrero and Ribeiro (2020); Didimo et al. (2020); Saponaro (2025)	Only 3 of 68 studies explicitly address legal redesign; mandated as structural corrective, not yet causally validated
5. Decision Support and Compliance Outcomes	RCT evidence that ML-targeting, not nudge content, drives compliance gains; NLP sentiment provides an early-warning signal (Convergence 3)	Hauser et al. (2025); Coita et al. (2023); Puklavac et al. (2023)	Strongest causal evidence in the corpus, but from a single setting (one US unemployment agency)
6. Continuous Governance Evolution and Ethical Feedback	35.3% of studies show governance/ethical gaps concentrated in technical studies; feedback loop added as structural corrective (Divergence 2)	Bogucki (2025) (ELSA-RRI); Table 5, Panel C	Reflects an appraisal-derived gap rather than a positive evidentiary success case

This derivation logic clarifies that the AAITCF is best understood as a theoretically saturated synthesis of recurring CIMO patterns rather than a deductively imposed model, and that the directiveness of each recommendation is calibrated to the strength of its underlying evidence.

4.5. The Adaptive AI Tax Compliance Framework (AAITCF)

Based on the structured CIMO synthesis of the 68 studies, the AAITCF proposes a heuristic six-step governance model (Steps 1–6 in Figure 6) explicitly differentiated by three institutional maturity tiers (Tier 1, Advanced: Italy, Estonia, United States; Tier 2, Intermediate: China, Indonesia, Malaysia, Brazil, Tunisia, Thailand, Romania, Kazakhstan; Tier 3, Foundational: Zambia, Botswana, Rwanda, Tanzania, Ethiopia, Yemen). These tiers are heuristic categories intended to guide comparative interpretation rather than fixed classifications of national tax administrations.

4.5.1. Institutional Maturity Tiers: Readiness Criteria and Transition Pathways

The three institutional tiers are not labels assigned post hoc to countries; they are operationalized along four readiness dimensions identified as recurring boundary conditions across the CIMO synthesis (Sections 4.1–4.4): data infrastructure and volume, regulatory/legal capacity, human capital, and governance/oversight maturity. Table 7 specifies indicative readiness criteria for each dimension and tier, drawn from the institutional conditions documented across the corpus.

Table 7. Institutional Maturity Tiers: Readiness Criteria Across Four Dimensions.

Readiness Dimension	Tier 3 (Foundational)	Tier 2 (Intermediate)	Tier 1 (Advanced)
Data infrastructure and volume	Fragmented, small, or non-digitized datasets; limited audit history (Adu-Ababio et al., 2025, Zambia; Ebrahim et al., 2024, Tanzania)	Moderate, structured administrative datasets (~10,000–1 M records), partial digitization (Kamoun et al., 2025: 20,930 obs.; Yang, 2025)	Large-scale, longitudinal administrative microdata (>1 M observations), real-time data feeds (Siimon & Lukason, 2021; Künnapas et al., 2025)
Regulatory/legal framework	Largely silent on AI in tax administration; no formal accountability mechanism (Munjeji & Schutte, 2024, Botswana)	General data-protection law exists; AI-specific tax guidance is emergent (Saragih et al., 2023, Indonesia)	Explicit AI/data-protection legislation with defined algorithmic-accountability obligations (Bogucki, 2025: GDPR, EU AI Act)
Human capital	Documented skills gap; little to no in-house data-science capacity (Munjeji & Schutte, 2024; Mudome & Ndiege, 2025)	Some trained analytical staff, concentrated at headquarters; reliance on external consultants for advanced models	In-house data-science teams; dedicated AI/analytics units within the tax authority
Governance and oversight	No formal oversight mechanism; enforcement decisions not yet algorithmically informed	Ad hoc human review of model outputs; explainability tools (e.g., SHAP) used inconsistently	Established bias-audit, explainability and human-in-the-loop protocols; periodic algorithm review (Bogucki, 2025; Didimo et al., 2020)
Illustrative cases in corpus	Zambia, Botswana, Rwanda, Tanzania, Ethiopia, Yemen	China, Indonesia, Malaysia, Brazil, Tunisia, Thailand, Romania, Kazakhstan	Italy, Estonia, United States

These criteria are diagnostic rather than fixed: an administration may satisfy Tier 1 criteria on data infrastructure while remaining Tier 3 on governance and human capital, a heterogeneity documented in Rwanda, where Uwizeye (2025) shows that EBM-generated transactional data already supports ML-based revenue forecasting even though formal AI governance and in-house analytical capacity remain limited. The AAITCF therefore treats the institutional tier as a dimension-specific profile rather than a single composite score, consistent with the Divergence 1 finding (Section 4.4.4) that no single moderator, data, trust, or infrastructure, dominates uniformly across the corpus.

4.5.2. The Six-Step Implementation Cycle

Step 1, Contextual Readiness and Data Infrastructure Assessment, reflects the finding that 63.2% of studies are single-country in scope and that data quality is more predictive of success than algorithmic sophistication in Tier 3 (Foundational) contexts (Adu-Ababio et al., 2025; Uwizeye, 2025; Saragih et al., 2023). Step 2, Pattern Recognition and Anomaly Detection, recommends Random Forest and XGBoost as the baseline architectures for Tier 2 (Intermediate) and Tier 3 (Foundational) (Kamoun et al., 2025; Yang, 2025), with neural networks, graph models, and ensemble methods reserved for Tier 1 (Advanced). Step 3, Predictive Risk Assessment and Audit Prioritization, implements feature-importance-informed risk scoring calibrated to sector-specific profiles (Yang, 2025; Xu & Kong, 2024; Tuyishimire & Murorunkwere, 2024). Step 4, Explainability and Human-in-the-Loop Validation, mandates SHAP-based interpretability (Kamoun et al., 2025) and interactive expert review (Didimo et al., 2020; Saponaro, 2025) across all three tiers, aligned with legal accountability requirements identified by Bogucki (2025), Bevacqua (2020), and Calderón Carrero and Ribeiro (2020). Step 5, Decision Support and Compliance Outcomes, combines the non-punitive compliance methods of ML-targeted behavioral nudging (Hauser et al., 2025), sentiment-based early warning (Coita et al., 2023), and social media analytics (Puklavec et al., 2023; Kuandykov et al., 2021). Step 6, Continuous Governance Evolution and Ethical Feedback, incorporates bias auditing, legal review and periodic performance evaluation under Bogucki’s (2025) ELSA-RRI recommendations, ensuring that governance frameworks evolve with the changes in AI capabilities and regulatory landscapes.

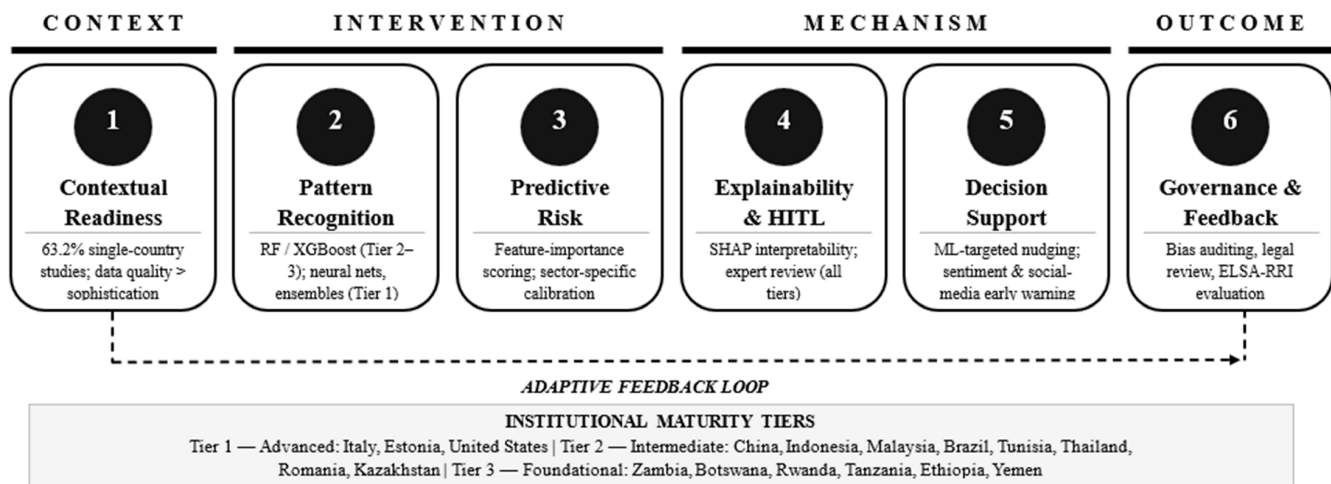


Figure 6. The Adaptive AI Tax Compliance Framework (AAITCF): A CIMO-Grounded, Tiered Implementation Cycle with an Adaptive Feedback Loop.

Figure 6 summarizes the AAITCF by translating the CIMO synthesis into a tiered implementation logic. Rather than proposing a universal model, the framework shows that AI-based tax compliance depends on the interaction between institutional context, technological intervention, operational mechanisms, and measurable outcomes. Its main contribution is to connect the six implementation steps into one adaptive cycle, where technical performance, human validation, legal safeguards, and governance feedback continuously inform each other. The framework is therefore intended as an evidence-grounded and context-sensitive roadmap that can be adjusted to foundational, intermediate, and advanced tax administrations.

5. Conclusions

5.1. Theoretical Implications

This study advances theoretical development in AI-powered tax compliance by clarifying how supervised learning methods interact with behavioral and governance dimensions of tax evasion. It provides theoretical underpinning for context-sensitive approaches aligned with the OECD's Tax Administration 3.0 vision, drawing on data science, legal research, behavioral economics, and institutional theory.

5.2. Practical Implications

5.2.1. Implications for Governments and Tax Administrations

This study provides differentiated recommendations with regard to the degree of technological maturity. Interoperability, ethical governance, and mitigating algorithmic bias should be the focus of advanced economies. The first step is to build digital infrastructure and standardized data in low- and middle-income countries, and then gradually introduce AI starting from low-cost supervised learning techniques. Successful implementation relies on a number of critical success factors, highlighted by Munjevi and Schutte, including capacity building, staff training, and the gradual improvement of regulation.

5.2.2. Implications for Audit Companies and Professionals

AI's influence on audit work is changing the focus from transactional execution to analytical oversight. AI tools help detect irregularities and facilitate risk-based monitoring. However, professionals need to retain their judgment and independence, especially when assessing or challenging unclear automated evaluations. There needs to be continuous training and strong human-in-the-loop systems in place.

5.2.3. Implications for Taxpayers

AI-powered compliance systems may foster trust by enhancing fairness and transparency; however, they also raise legitimate concerns regarding privacy, surveillance, and the risk of algorithmic misclassification. Clear communication of how these systems operate, robust legal safeguards, and accessible mechanisms for contesting decisions are essential to strike an appropriate balance between administrative efficiency and the protection of taxpayer rights.

5.3. Limitations of the Study

Five limitations merit acknowledgment. First, restricting the corpus to Scopus and Web of Science (2015–2025) may introduce representativeness bias by excluding gray literature and non-indexed regional publications. Second, IRAMUTEQ's term-frequency logic may obscure semantic nuances not fully captured by lexical proximity. Third, the geographic asymmetry of the corpus, with Europe and Asia overrepresented relative to Sub-Saharan Africa and Latin America, limits the generalizability of the AAITCF's tiered recommendations. Fourth, this review establishes theoretical propositions grounded in correlational evidence; empirical validation of the AAITCF through longitudinal or quasi-experimental designs remains a necessary next step. Finally, the review did not employ independent dual coding with formal inter-rater reliability statistics; both authors coded with iterative calibration and consensus resolution.

5.4. Recommendations for Future Research

Six avenues merit exploration. First, empirical validation of the AAITCF through longitudinal case studies or quasi-experimental designs. Second, cross-country comparisons of AI adoption between high-income and low/middle-income contexts. Third, exami-

nation of long-term institutional change and taxpayer trust beyond short-term accuracy measures. Fourth, development of implementation guides for different digital maturity stages. Fifth, analysis of ethical and governance dimensions, including transparency, bias, and due process. Sixth, assessment of the AAITCF's role in broader Tax Administration 3.0 governance arrangements.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/jrfm19070513/s1>. Table S1: PRISMA 2020 Checklist. Table S2: Combined Auditable List of Included Studies and CIMO Extraction Matrix (N = 68), providing the master reference list with complete bibliographic information, institutional maturity classification, primary CIMO codes, and database source for every study included in the systematic literature review, is provided as Supplementary Materials. Page et al. (2021) are cited in the Supplementary Materials.

Author Contributions: Conceptualization, H.Z. and S.S.; methodology, H.Z. and S.S.; validation, S.S.; formal analysis, H.Z. and S.S.; Software, H.Z.; investigation, H.Z.; resources, H.Z.; data curation, H.Z.; writing—original draft preparation, H.Z.; writing—review and editing, S.S.; visualization, H.Z.; supervision, S.S.; project administration, H.Z. All authors have read and agreed to the published version of the manuscript.

Funding: The first author receives funding from the National Center for Scientific and Technical Research (CNRST) through the PhD-Associate program, Grant number: 135USMBA2024.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The review is based on bibliographic records retrieved from Scopus and Web of Science. The derived review dataset, including the list of included studies and CIMO extraction matrix, is provided in Supplementary Materials Table S2. Access to the original database records may depend on institutional subscriptions.

Acknowledgments: This work was supported by the National Center for Scientific and Technical Research (CNRST) through the PhD-Associate Program (for the first author, Houda Zaim), grant number 135USMBA2024.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
ANOVA	Analysis of Variance
AAITCF	Adaptive AI Tax Compliance Framework
CIMO	Context–Intervention–Mechanism–Outcome
DHC	Descending Hierarchical Classification
ELSA	Ethical, Legal, and Socioeconomic Aspects
EU	European Union
EVS/WVS	European Values Survey / World Values Survey
FBNE-PU	Feature-Based Network Embedding with Positive-Unlabeled learning
GAN	Generative Adversarial Network
GDPR	General Data Protection Regulation
IPSO	Improved Particle Swarm Optimization
IRAMUTEQ	Interface de R pour les Analyses Multidimensionnelles de Textes et de Questionnaires
IRS	Internal Revenue Service
ML	Machine Learning
MLARAM	Multi-Label Adaptive Resonance Association Map

MLkNN	Multi-Label k-Nearest Neighbors
MLP	Multi-Layer Perceptron
MMAT	Mixed Methods Appraisal Tool
NLP	Natural Language Processing
NSGA-II	Non-dominated Sorting Genetic Algorithm II
OECD	Organisation for Economic Co-operation and Development
PnCGCN	Positive-negative Correlated Graph Convolutional Network
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PU	Positive Unlabeled
RAkELd	RANdom k-labELsets disjoint
ROC AUC	Receiver Operating Characteristic—Area Under the Curve
RPA	Robotic Process Automation
RRI	Responsible Research and Innovation
SHAP	SHapley Additive exPlanations
SLR	Systematic Literature Review
SMEs	Small and Medium-sized Enterprises
SMOTE	Synthetic Minority Over-sampling Technique
SVM	Support Vector Machine
TOE	Technology–Organization–Environment
VAT	Value Added Tax
XGBoost	eXtreme Gradient Boosting

References

- Adu-Ababio, K., Koivisto, A., & Mwale, E. (2025). Estimating audit based tax gaps in Zambia. *International Tax and Public Finance*, 1–51. [\[CrossRef\]](#)
- Alrasheedi, M. A., Ijaz, S., Alrashdi, A. M., & Lee, S.-W. (2025). Advanced tax fraud detection: A soft voting ensemble based on GAN and encoder architecture. *Mathematics*, 13(4), 642. [\[CrossRef\]](#)
- Alstadsæter, A., Godar, S., Nicolaides, P., & Zucman, G. (2024). *Global tax evasion report 2024*. EU Tax Observatory. Available online: <https://taxobservatory.world/publication/global-tax-evasion-report-2024/> (accessed on 17 April 2026).
- Al-zi'aylee, H. M. R., Al-Issawi, M. A. A., Al-Dolamy, D. K. M., & Alajrawi, A. J. S. (2025). The use of digital technologies and artificial intelligence in improving tax collection efficiency in Iraq. *Veredas do Direito*, 22, e223699. [\[CrossRef\]](#)
- Antinozzi, H. S., & Cooper, L. A. (2025). Is ChatGPT an accurate source of information for uninformed taxpayers? *Journal of Emerging Technologies in Accounting*, 22(1), 23–43. [\[CrossRef\]](#)
- Antón, F. S. (2021). Artificial intelligence and tax administration: Strategy, applications and implications, with special reference to the tax inspection procedure. *World Tax Journal*, 13, 575–608. [\[CrossRef\]](#)
- Axtell, R. L., & Farmer, J. D. (2025). Agent-based modeling in economics and finance: Past, present, and future. *Journal of Economic Literature*, 63, 197–287. [\[CrossRef\]](#)
- Babu, S. K., & Vasavi, S. (2017). Predictive analytics as a service on tax evasion using Gaussian regression process. *Helix*, 7, 1988–1993.
- Battaglini, M., Guiso, L., Lacava, C., Miller, D. L., & Patacchini, E. (2025). Refining public policies with machine learning: The case of tax auditing. *Journal of Econometrics*, 249, 105847. [\[CrossRef\]](#)
- Battiston, P., Gamba, S., & Santoro, A. (2024). Machine learning and the optimization of prediction based policies. *Technological Forecasting and Social Change*, 199, 123080. [\[CrossRef\]](#)
- Belahouaoui, R., & Alm, J. (2025). Tax fraud detection using artificial intelligence-based technologies: Trends and implications. *Journal of Risk and Financial Management*, 18, 502. [\[CrossRef\]](#)
- Belahouaoui, R., & Attak, E. H. (2024). Digital taxation, artificial intelligence and Tax Administration 3.0: Improving tax compliance behavior—A systematic literature review using textometry (2016–2023). *Accounting Research Journal*, 37(2), 172–191. [\[CrossRef\]](#)
- Bevacqua, J. (2020). Tax authority immunity in a digital tax administration world. *eJournal of Tax Research*, 18(2), 402–440.
- Bogucki, A. (2025). Ethical, legal, and socioeconomic aspects of implementing artificial intelligence in tax administration. *Acta Universitatis Lodzianis Folia Iuridica*, 110, 19–36. [\[CrossRef\]](#)
- Borrotti, M., Rabasco, M., & Santoro, A. (2023). Using accounting information to predict aggressive tax location decisions by European groups. *Economic Systems*, 47, 101090. [\[CrossRef\]](#)
- Calderón Carrero, J., & Ribeiro, J. (2020). Fighting tax fraud through artificial intelligence tools: Will the fundamental rights of taxpayers survive the digital transformation of tax administrations? *European Taxation*, 60(6), 235–238. [\[CrossRef\]](#)

- Chang, N.-C. (2020). Double/debiased machine learning for difference-in-differences models. *Econometrics Journal*, 23, 177–191. [CrossRef]
- Chou, J.-S., Yeh, P.-C., Liu, C.-Y., & Chen, K.-J. (2025). Improving detection of pollution fee declarations for environmental policy compliance through metaheuristic optimized ensemble learning. *Engineering Construction and Architectural Management*, 33, 5820–5848. [CrossRef]
- Coita, I.-F., Belbe, S., Mare, C., Osterrieder, J., & Hopp, C. (2023). Modelling taxpayers' behavior based on prediction of trust using sentiment analysis. *Finance Research Letters*, 58, 104549. [CrossRef]
- Cuc, L. D., Rad, D., Cilan, T. F., Gomo, B. C., Nicolaescu, C., Almasi, R., Dzitac, S., Isac, F. L., & Pandelica, I. (2025). From AI knowledge to AI usage intention in the managerial accounting profession and the role of personality traits, a decision tree regression approach. *Electronics*, 14, 1107. [CrossRef]
- Davidescu, A. A., Manta, E. M., Stoica Ungureanu, A. T., & Anton (Musat), M. (2022). Could religiosity and religion influence the tax morale of individuals? An empirical analysis based on variable selection methods. *Mathematics*, 10(23), 4497. [CrossRef]
- Denyer, D., Tranfield, D., & van Aken, J. E. (2008). Developing design propositions through research synthesis. *Organization Studies*, 29(3), 393–413. [CrossRef]
- Didimo, W., Grilli, L., Liotta, G., Menconi, L., Montecchiani, F., & Pagliuca, D. (2020). Combining network visualization and data mining for tax risk assessment. *IEEE Access*, 8, 16073–16086. [CrossRef]
- Ebrahim, A., Castillo, S., Leyaro, V., Swema, E., & Haule, O. (2024). Bridging the gap: Unveiling the potential of Tanzania's SMEs through VAT insights. *CESifo Economic Studies*, 70(4), 540–577. [CrossRef]
- Fidelangeli, A., & Galli, F. (2021). Artificial intelligence and tax law: Perspectives and challenges. *Ceridap*, 221(4), 24–58. [CrossRef]
- Franic, J. (2022). What do we really know about the drivers of undeclared work? An evaluation of the current state of affairs using machine learning. *AI and Society*, 39, 597–616. [CrossRef] [PubMed]
- Galler, L., & Soled, J. A. (2024). AI and the regulation of tax return preparers. *Florida Tax Review*, 28(1), 1–50. [CrossRef]
- Gao, Y., Shi, B., Dong, B., Wang, Y., Mi, L., & Zheng, Q. (2023). Tax evasion detection with FBNE-PU algorithm based on PnCGCN and PU learning. *IEEE Transactions on Knowledge and Data Engineering*, 35(1), 931–944. [CrossRef]
- Güney, C., Konyar, M. Z., Akıncı, A., Ala, T., & Alkan, O. (2025). How can we use artificial intelligence and machine learning in accounting and auditing: An engineering overview with the mapping method in the field of business. *International Journal of Contemporary Economics and Administrative Sciences*, 15, 467–510. [CrossRef]
- Hakam, L. I., Kurniawati, S., Akram, F. M., Salim, D. R., & Fuadi, S. S. (2024). Tax transformation through artificial intelligence: A systematic review of generational preference differences and their impact on tax compliance. *Jurnal Pajak Indonesia (Indonesian Tax Review)*, 8(2), 244–266. [CrossRef]
- Hauser, O. P., Greene, M., & DeCelles, K. (2025). Catch me if you can: Using machine learning and behavioral interventions to reduce unethical behavior. *Behavioral Public Policy*, 10, 692–709. [CrossRef]
- Higgins, J. P. T., Altman, D. G., Gøtzsche, P. C., Jüni, P., Moher, D., Oxman, A. D., Savović, J., Schulz, K. F., Weeks, L., Sterne, J. A. C., Cochrane Bias Methods Group & Cochrane Statistical Methods Group. (2011). The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *BMJ*, 343, d5928. [CrossRef] [PubMed]
- Hong, Q. N., Pluye, P., Bujold, M., & Vedel, I. (2018). *Mixed methods appraisal tool (MMAT)* (version 2018) [Registration of Copyright (#1148552)]. Canadian Intellectual Property Office, Industry Canada.
- International Monetary Fund (IMF). (2024). *Fiscal monitor, April 2024: Fiscal policy in the great election year*. Available online: <https://www.imf.org/en/publications/fm/issues/2024/04/17/fiscal-monitor-april-2024> (accessed on 17 April 2026).
- Kamoun, A., Boujelbane, R., & Boujelben, S. (2025). A prediction model to detect non compliant taxpayers using a supervised machine learning approach: Evidence from Tunisia. *Journal of Business Analytics*, 8(2), 116–133. [CrossRef]
- Karataş Durmu, N., & Arıtı Erdem, İ. (2023). Vergi İdaresi 3.0: Yapay Zekâ Perspektifinden Bir İnceleme [Tax administration 3.0: An analysis in terms of artificial intelligence]. *Maliye Dergisi*, 184, 225–253.
- Khaltar, O. (2024). Tax evasion and governance quality: The moderating role of adopting open government. *International Review of Administrative Sciences*, 90(1), 276–294. [CrossRef]
- Kostova, S., & Zhelev, Z. (2024). Identifying money laundering risk indicators: Evidence from Bulgaria. *Edelweiss Applied Science and Technology*, 8, 2809–2816. [CrossRef]
- Kuandykov, N., Abil, Y., Utepov, A., Nurkey, A., & Abenova, A. (2021). Investigating the relationship between tax culture and tax communication in Kazakhstan. *Problems and Perspectives in Management*, 19(3), 197–210. [CrossRef]
- Künnapas, K., Maslionkina, P., Hinno, R., & Liberts, R. (2025). Current AI applications by Estonian tax authorities and use case scenarios. *TalTech Journal of European Studies*, 15(1), 179–208. [CrossRef]
- Lazovic Pita, L., Odzak, A., & Sceta, L. (2025). Examining individual tax morale in Europe with machine learning methods. *Central European Public Administration Review*, 23(1), 129–155. [CrossRef]
- Lee, E. Y., Leeroy, G. C., & Leeroy, W. (2024). Impact of blockchain on improving taxpayers' compliance: Empirical evidence from panel data model and agent based simulation. *Journal of Emerging Technologies in Accounting*, 21(1), 89–109. [CrossRef]

- Malashin, I. P., Masich, I. S., Tynchenko, V. S., Gantimurov, A. P., Nelyub, V. A., & Borodulin, A. S. (2025). Minimizing unnecessary tax audits using multi objective hyperparameter tuning of XGBoost with focal loss. *Frontiers in Artificial Intelligence*, 8, 1669191. [CrossRef]
- Mathew, L. A., & Devan, V. V. (2025). Leveraging AI to combat tax evasion: A GST perspective. *International VAT Monitor*, 36(3), 91–98. [CrossRef]
- Mengist, W., Soromessa, T., & Legese, G. (2020). Method for conducting systematic literature review and meta analysis for environmental science research. *MethodsX*, 7, 100777. [CrossRef] [PubMed]
- Miller, A. H. (2019). Tax strategies employed by overseas American individuals and SMEs. *Journal of Social Sciences Research*, 5, 690–699. [CrossRef]
- Mojahedi, H., Babazadeh Sangar, A., & Masdari, M. (2022). Towards tax evasion detection using improved particle swarm optimization algorithm. *Mathematical Problems in Engineering*, 2022, 1027518. [CrossRef]
- Mu, R., Fentaw, N. M., & Zhang, L. (2023). Tax evasion, psychological egoism, and revenue collection performance: Evidence from Amhara region, Ethiopia. *Frontiers in Psychology*, 14, 1045537. [CrossRef] [PubMed]
- Mudome, J., & Ndiege, J. R. A. (2025). Leveraging artificial intelligence for enhanced tax collection in developing nations: A systematic literature review. *African Multidisciplinary Tax Journal*, 5, 254–275. [CrossRef]
- Munjeji, E., & Schutte, D. (2024). Examining the critical success factors influencing the diffusion of AI in tax administration in Botswana. *Cogent Social Sciences*, 10(1), 2419537. [CrossRef]
- Munjeji, E., & Schutte, D. (2025). Development of an artificial intelligence framework to combat tax noncompliance in Botswana. *Acta Commercii*, 25(1), 12. [CrossRef]
- Murorunkwere, B. F., Tuyishimire, O., Haughton, D., & Nzabanita, J. (2022). Fraud detection using neural networks: A case study of income tax. *Future Internet*, 14(6), 168. [CrossRef]
- Neuman, E. L., & Sheu, R. J. (2022). Big Data analytics in IRS audit procedures and its effects on tax compliance: A moderated mediation analysis. *Journal of the American Taxation Association*, 44(2), 97–113. [CrossRef]
- OECD. (2020). *Tax administration 3.0: The digital transformation of tax administration*. OECD Publishing. Available online: <https://www.oecd.org/tax/forum-on-tax-administration/publications-and-products/tax-administration-3-0-the-digital-transformation-of-tax-administration.pdf> (accessed on 17 April 2026).
- Oehlert, C., Schulz, E., & Parker, A. (2022). NAICS code prediction using supervised methods. *Statistics and Public Policy*, 9, 58–66. [CrossRef]
- Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., . . . Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372, n71. [CrossRef] [PubMed]
- Paul, J., Merchant, A., Dwivedi, Y. K., & Rose, G. (2021). Writing an impactful review article: What do we know and what do we need to know? *Journal of Business Research*, 133, 337–340. [CrossRef]
- Pluye, P., & Hong, Q. N. (2014). Combining the power of stories and the power of numbers: Mixed methods research and mixed studies reviews. *Annual Review of Public Health*, 35, 29–45. [CrossRef] [PubMed]
- Prasetyo, T., Susetyo, B., & Kurnia, A. (2024). Improving the risk profile of Indonesian enterprise taxpayers using multilabel classification. *IAES International Journal of Artificial Intelligence*, 13(4), 4323–4333. [CrossRef]
- Puklavec, Z., Kogler, C., Stavrova, O., & Zeelenberg, M. (2023). What we tweet about when we tweet about taxes: A topic modelling approach. *Journal of Economic Behavior & Organization*, 212, 1242–1254. [CrossRef]
- Pylypenko, O., Matviienko, H., Putintsev, A., Vlasenko, I., & Onyshchuk, N. (2022). Government tax policy in the digital economy. *Cuestiones Políticas*, 40, 279–296. [CrossRef]
- Rabasco, M., & Battiston, P. (2023). Predicting the deterrence effect of tax audits. A machine learning approach. *Metroeconomica*, 74(3), 531–556. [CrossRef]
- Radcliffe, V., Spence, C., & Stein, M. (2025). Tax, technology, and craftsmanship. *Accounting Review*, 100, 293–316. [CrossRef]
- Ratinaud, P. (2009). *IRAMUTEQ: Interface de R pour les analyses multidimensionnelles de textes et de questionnaires* [Logiciel]. Available online: <http://www.iramuteq.org> (accessed on 17 April 2026).
- Saavedra, S., & Romero, M. (2021). Local incentives and national tax evasion: The response of illegal mining to a tax reform in Colombia. *European Economic Review*, 138, 103843. [CrossRef]
- Saponaro, F. (2025). Intelligence and taxation: The Ve.R.A. algorithm. *Revista Brasileira de Direito*, 21(1), e5240. [CrossRef]
- Saragih, A. H., Reyhani, Q., Setyowati, M. S., & Hendrawan, A. (2023). The potential of an artificial intelligence (AI) application for the tax administration system's modernization: The case of Indonesia. *Artificial Intelligence and Law*, 31(3), 491–514. [CrossRef]
- Savić, M., Atanasijević, J., Jakovetić, D., & Krejić, N. (2022). Tax evasion risk management using a hybrid unsupervised outlier detection method. *Expert Systems with Applications*, 193, 116409. [CrossRef]

- Shujaaddeen, A., Ba-Alwi, F. M., & Al-Gaphari, G. (2024). A new machine learning model for detecting levels of tax evasion based on hybrid neural network. *International Journal of Intelligent Systems and Applications in Engineering*, 12, 450–468.
- Siimon, O. R., & Lukason, O. (2021). A decision support system for corporate tax arrears prediction. *Sustainability*, 13(15), 8363. [CrossRef]
- Soled, J. A., & Thomas, K. D. (2023). AI, taxation, and valuation. *Iowa Law Review*, 108, 651–701.
- Suwarni, Akhmetshin, E. M., Okagbue, H. I., Laxmi Lydia, E., & Shankar, K. (2020). Digital economic challenges and economic growth in environmental revolution 4.0. *Journal of Environmental Treatment Techniques*, 8, 546–550.
- Tax Justice Network. (2024). *The state of tax justice 2024*. Available online: <https://taxjustice.net/reports/the-state-of-tax-justice-2024/> (accessed on 21 April 2026).
- Thanakijkhajonphaisan, D., & Sincharoonsak, T. (2025). The influence of revenue department's tax auditing technologies, social motivations and online business tax management capabilities on the success of tax administration of e commerce businesses in Thailand. *Scientific Culture*, 11(3.1), 269–285. [CrossRef]
- Topchii, V., Moroz, Y., Karpenko, N., Khoronovskyi, O., & Tarashchenko, V. (2025). Prevention of tax criminal offences as a factor in the financial stability of the state. *Theoretical and Practical Research in the Economic Fields*, 16, 170–181.
- Tuyishimire, O., & Murorunkwere, B. F. (2024). Applications of big data analytics in tax compliance monitoring: A case study of Rwanda's value added tax. *CESifo Economic Studies*, 70(4), 578–587. [CrossRef]
- Uwizeye, C. (2025). Digitalising tax compliance and elevating revenue forecasting in Rwanda: Evidence from statistical modelling and machine learning. *African Multidisciplinary Tax Journal*, 5, 311–330. [CrossRef]
- Visitpanya, N., & Samanchuen, T. (2023). Synthesis of tax return datasets for development of tax evasion detection. *IEEE Access*, 11, 48203–48220. [CrossRef]
- Wahab, R. A. S. R., & Abu Bakar, A. (2021). Digital economy tax compliance model in Malaysia using machine learning approach. *Sains Malaysiana*, 50(7), 2059–2077. [CrossRef]
- Warner, G., Wijesinghe, S., Marques, U., Badar, O., Rosen, J., Hemberg, E., & O'Reilly, U.-M. (2015). Modeling tax evasion with genetic algorithms. *Economics of Governance*, 16(2), 165–178. [CrossRef]
- Xavier, O. C., Pires, S. R., Marques, T. C., & da Silva Soares, A. (2022). Tax evasion identification using open data and artificial intelligence. *Revista de Administração Pública*, 56(3), 426–440. [CrossRef]
- Xu, C., & Kong, Y. (2024). Random forest model in tax risk identification of real estate enterprise income tax. *PLoS ONE*, 19(3), e0300928. [CrossRef] [PubMed]
- Yang, L. (2025). Predictive modeling of tax compliance risks: A comparative study of machine learning approaches. *PLoS ONE*, 20(9), e0331715. [CrossRef] [PubMed]
- Zheng, S., Trott, A., Srinivasa, S., Parkes, D. C., & Socher, R. (2022). The AI economist: Taxation policy design via two level deep multiagent reinforcement learning. *Science Advances*, 8(18), eabk2607. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.