



Journal of Frontiers in Multidisciplinary Research

Developing Explainable Artificial-Intelligence Audit Tools to Identify Revenue-Recognition, Payroll, Expense-Management, and Financial-Reporting Risks While Supporting Professional Audit Judgment

Prosper Fuadedzi

Georgia State University Atlanta, GA, USA

* Corresponding Author: **Prosper Fuadedzi**

Article Info

E-ISSN: 3050-9726

P-ISSN: 3050-9718

Volume: 07

Issue: 02

Received: 18-08-2026

Accepted: 16-09-2026

Published: 14-10-2026

Page No: 340-345

Abstract

Auditors increasingly have access to complete populations of payroll records, card transactions, and financial statements, but black-box anomaly scores are difficult to defend in audit documentation and can displace, rather than inform, professional judgment. This study develops and evaluates XAT, an explainable audit toolkit that combines transparent audit rules with peer-group statistical baselines, so that every flagged item comes with a reason an auditor can verify. XAT is applied to three real U.S. public datasets: 544,338 New York City payroll records (FY2024), 109,011 Oklahoma state purchase-card transactions (April–June 2024), and 35,166 public-company firm-years labeled with SEC enforcement actions. In payroll and expense data, where no fraud labels exist, we seed 2,098 realistic anomalies (padded overtime, salary overpayment, ghost employees, un-supported payments, split purchases, duplicates, personal-use purchases, and inflated amounts) into the real records. Within a review budget of 1% of records, XAT surfaces 31.1% of seeded payroll anomalies and 26.0% of seeded card anomalies, versus 8.0% and 9.7% for an isolation forest and 13.8% and 13.6% for rules alone. For revenue recognition, a sparse, glass-box logistic model of revenue red flags achieves an out-of-sample AUC of 0.647 against SEC-identified fraud, close to a black-box gradient-boosting model (0.671) and well above a Beneish-style revenue index (0.553). The results show that explainable tools can match or exceed opaque detectors while producing evidence that supports, rather than replaces, auditor judgment.

Keywords: explainable artificial intelligence, audit analytics, payroll fraud, purchase cards, expense management, revenue recognition, Benford's law, human-in-the-loop, professional judgment

1. Introduction

Auditing has traditionally relied on sampling: an auditor examines a few dozen payroll records or card transactions and extrapolates. Today, auditors can obtain entire populations of payroll, expense, and financial-reporting data, and audit standards and professional guidance encourage data analytics that examine every record^[1,2]. Payroll and expense-reimbursement schemes remain among the most common forms of occupational fraud^[3], and improper revenue recognition is a frequent subject of SEC enforcement^[4].

Machine learning can rank millions of records for review, but its outputs create a new problem for the audit profession. An auditor who acts on a flag must document *why* the item was selected and what evidence was obtained^[5,6]. A black-box anomaly score does not provide that reason. It also invites two failure modes documented in research on human–automation interaction: over-reliance, in which the auditor accepts the tool's output without scrutiny, and under-reliance, in which a correct flag is dismissed because it cannot be explained^[7,8]. Post-hoc explanation methods help, but they explain an opaque model rather than making the model itself transparent^[9,10].

This paper takes a different approach. We develop XAT (Explainable Audit Toolkit), in which every component is inherently interpretable: transparent audit rules that encode known fraud schemes, peer-group baselines that compare each record with similar records (for example, the same job title in the same agency), and, where labels exist, a sparse linear model whose coefficients can be read directly. Every flagged item comes with an *explanation card*: which rules fired, how far the item deviates from its peers, and what follow-up procedure is suggested. The auditor, not the tool, decides what the evidence means.

We evaluate XAT on the four areas named in its purpose:

- **Payroll:** 544,338 New York City payroll records for fiscal year 2024.
- **Expense management:** 109,011 Oklahoma state purchase-card transactions posted in April–June 2024, the last quarter of the state’s fiscal year.

Revenue Recognition and Financial Reporting: 35,166

U.S. public-company firm-years (2003–2008) labeled with SEC Accounting and Auditing Enforcement Releases (AAERs) [11].

Because payroll and card data carry no fraud labels, we follow the established practice of seeding realistic anomalies into real data and measuring how many the tool surfaces within a fixed review budget [12]. For revenue recognition, we use real enforcement labels.

The contributions are: (i) an explainable toolkit design that combines rules, peer baselines, and glass-box models, with explanation cards for audit documentation; (ii) a reproducible evaluation on three real U.S. public datasets covering payroll, expenses, and revenue; (iii) evidence that the transparent combination outperforms both rules alone and a black-box isolation forest at realistic review budgets, and that a glass-box revenue model gives up little accuracy relative to a black-box model; and

(iv) real-data findings, including the scale of rule exceptions an auditor would face, a Benford’s-law test, and a check for fiscal year-end spending.

2. Related Work

Audit data analytics. Appelbaum *et al.* [2] set out research needs for analytics in the modern audit, and the AICPA’s guide describes how analytics fit into risk assessment and substantive procedures [1]. Nigrini and Mittermaier [13] introduced Ben-ford’s law as an analytical procedure, and Nigrini [14] codified conformity thresholds that remain standard in forensic account-ing.

Anomaly detection in accounting data. Schreyer *et al.* [12] used deep autoencoders to detect anomalous journal entries and evaluated them partly with seeded anomalies. Isolation forests [15] are widely used as general-purpose detectors. These methods are powerful but opaque, and their scores do not trans-late directly into audit evidence.

Fraud prediction and explainability. Beneish [16] and Dechow *et al.* [4] built interpretable ratio models of earnings manipulation, including revenue-related signals such as receivables growth. Bao *et al.* [11] showed that ensemble learning on raw financial data improves detection. Rudin [9] argues that in high-stakes settings, inherently interpretable models should be preferred when they perform comparably. Commerford *et al.* [8] found that auditors’ reliance on AI-based evidence differs from their reliance on human specialists, underscoring the need for tools whose reasoning auditors can evaluate. XAT builds on these strands by making interpretability a design constraint rather than an add-on.

3. The Xat Toolkit

Fig. 1 shows the design. A shared engine applies two kinds of transparent tests to each module’s data, and a human-in-the-loop cycle turns flags into audit evidence.

3.1. Transparent Rules

Rules encode well-known schemes, each with a stated threshold that an audit team can adjust (Table I). For payroll they target excessive overtime, pay above salary, large unsupported payments, and the same person employed full-time in two agen-cies. For purchase cards they target split purchases (several same-day charges from one cardholder at one merchant that together exceed a single-transaction limit), possible duplicate payments, weekend or holiday use, and merchant categories

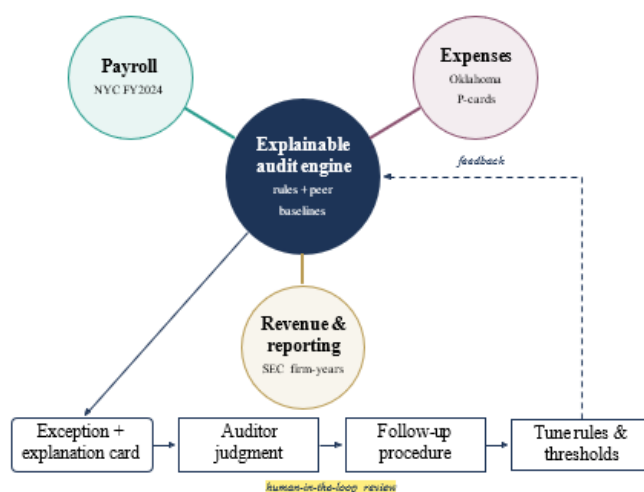


Fig 1: XAT design. A shared explainable engine (rules plus peer baselines) serves three modules. Each flag produces an explanation card that feeds a human-in-the-loop cycle: the auditor judges the evidence, performs follow-up procedures, and tunes rules and thresholds over time.

that are rare for the agency. We use \$5,000 as an illustrative single-transaction limit.

3.2. Peer-Group Baselines

Rules miss schemes that stay below thresholds. XAT therefore compares each record with its peers using robust z-scores. For feature j of record i in peer group g ,

$$z_{ij} = \max \left(0, \frac{x_{ij} - \text{med}_g(x_j)}{1.4826 \cdot \text{MAD}_g(x_j)} \right) \quad (1)$$

where med and MAD are the peer-group median and median absolute deviation, which resist distortion by the very outliers being sought^[17]. Only upward deviations, which correspond to overpayment or overspending, are scored. Peer groups are as specific as possible: agency $\times$ job title $\times$ pay basis for pay-

roll, and agency $\times$ merchant category for cards, falling back to broader groups when a group has fewer than 20 members. Payroll features are overtime hours and pay, overtime pay per hour, other pay, regular gross pay, and the ratio of gross pay to base salary. Card features are the amount, the amount relative to the cardholder's usual spending, and the rarity of the merchant category in the agency.

3.3. Combined Score and Explanation Card

The review score is $s_i = \log(1 + \max_j z_{ij}) + w \cdot r_i$, where r_i is the number of rules triggered and $w = 3$, so that a rule hit counts as much as a roughly twentyfold larger peer deviation. The score is a ranking device only. The explanation card lists the triggered rules, the top deviating features with their values, peer medi-ans, and peer-group sizes, and a suggested follow-up procedure (Fig. 4).

Table 1: Transparent Rules and Real-Data Exceptions

Rule	Description	Exceptions
Payroll rule (NYC FY2024, 534,338 real records)		
R1	Overtime > 1,500 hours in the base year	443
R2	Regular gross pay > 120% of base salary (annual basis)	17,905
R3	Other pay > regular pay and > \$20,000	1,549
R4	Same name and start date, full-time in 2+ agencies	285
	Any payroll rule	20,146 (3.7%)
Purchase-card rule (Oklahoma, 107,913 real transactions)		
E1	Possible split purchase (711 groups, \$3.53M)	3,001
E2	Possible duplicate (≤ 3 days apart, \$2.93M)	7,367
E3	Weekend or state holiday	13,005
E4	Merchant category < 0.1% of agency's transactions	2,816
	Any card rule	23,888 (22.1%)

3.4. Glass-Box Revenue Model

For revenue recognition, where enforcement labels exist, XAT uses an ℓ_1 -regularized (sparse) logistic regression^[18] on eight standardized revenue red flags: the Beneish days-sales-in-receivables index (DSRI), sales growth index, and gross-margin index^[16]; receivables and inventory relative to sales; and the change in receivables, change in cash sales, and share of soft assets^[4]. Each coefficient states directly how a red flag moves the risk. Following^[11], each test year t (2003–2008) is scored by a model trained on 1991 to $t - 2$, with serial frauds that span both periods recoded in training.

4. Data and Evaluation

4.1. Data

Payroll. The NYC Citywide Payroll dataset, published by the Office of Payroll Administration^[19], reports each employee's agency, title, pay basis, base salary, regular hours and gross pay, overtime hours and pay, and other pay. We keep FY2024 records with positive regular or overtime pay: 543,338 records across 159 agencies and 1,543 titles.

Purchase cards. Oklahoma's Office of Management and Enterprise Services publishes all state and higher-education

purchase-card transactions^[20]. We use the files posted in April, May, and June 2024: 111,988 rows, of which 3,875 re-funds are excluded, leaving 107,913 transactions worth \$49.5 million by 4,925 cardholders in 115 agencies.

Revenue and reporting. We use the public firm-year dataset of Bao *et al.*^[11], with AAER fraud labels; test years 2003–2008 contain 35,166 firm-years, 261 of them fraudulent.

Although both transaction datasets are public and name in-dividuals, this study reports only aggregates and anonymized examples.

4.2. Seeded Anomalies

We seeded realistic schemes into the real data, drawing templates from real records: in payroll, 250 cases each of padded overtime (400–900 extra hours paid at 1.5 times the hourly rate), salary overpayment (gross pay raised 15–40%), ghost or dupli-cate employees (a real employee cloned into another agency), and unsupported other pay (\$15,000–\$40,000 added). In cards, 200 cases each of split purchases (\$5,200–\$12,000 split into 2–3 same-day charges below the limit, 498 transactions), duplicate

Table 2: Detection of Seeded Anomalies in Real Data

Global z-score	0.2	2.1	0.814	0.6	16.7	25.0	0.859	10.1
Isolation forest	8.0	15.2	0.839	0.9	9.7	13.2	0.790	4.7
Rules only	13.8	25.6	0.720	2.3	13.6	29.7	0.836	7.7
Peer z-score	22.0	34.6	0.824	2.5	12.8	24.3	0.835	8.0
XAT (rules + peer z)	31.1	40.1	0.939	5.4	26.0	32.8	0.921	16.2

Top $k\%$: share (%) of seeded anomalies within the top $k\%$ of the ranked review list (random review: $k\%$). AP in percent.

payments (a charge repeated within 0–2 days), personal-use purchases (\$100–\$1,500 on a Saturday in a merchant category the agency does not use), and inflated amounts (5–15 times the original). Altered records replaced their originals; cloned and added records were appended. In total, 1,000 payroll records and 1,098 card transactions were seeded.

4.3. Baselines and Metrics

We compare XAT with rules alone, peer z-scores alone, a global z-score that ignores peer groups, and an isolation forest^[15, 21]. The main metric is the share of seeded anomalies found in the top 1% and 2% of the review list; random review would find 1% and 2%. We also report AUC and average precision (AP). For revenue, we report mean out-of-sample AUC and the share of frauds captured in the riskiest 5% of firms.

5. Results

5.1. Payroll and Expense Modules

Table II summarizes detection of seeded anomalies. Within a 1% review budget, XAT surfaces 31.1% of seeded payroll anomalies and 26.0% of seeded card anomalies, more than any baseline and about three to four times the share found by

the isolation forest. XAT also has the highest AUC and AP in both modules.

Fig. 2 breaks detection down by scheme and shows why the combination matters. Rules and peer baselines catch *different* schemes. In payroll, the peer z-score is the best detector of padded overtime (47.6%), because padded hours rarely cross the 1,500-hour rule but stand out sharply against colleagues with the same title. Rules are best for ghost employees (30.0%), because a cloned employee looks normal relative to peers. XAT keeps most of the strength of each. Two schemes remain hard: unsupported other pay (11.6%), because real other-pay amounts vary widely, and inflated card amounts, which no method de-tected well (at most 11.0%), because legitimate card amounts span several orders of magnitude even within a merchant cate-gory.

The rule weight w has a modest effect. Varying it from 1 to 5 changes XAT's top-1% payroll detection between 24.2% and 32.2%, and card detection between 25.6% and 26.9%.

5.2. What Auditors Would Face in the Real Data

Table I reports how many real records trigger each rule. These are exceptions for follow-up, not findings of fraud. The volumes are instructive: 22.1% of card transactions trigger at least one

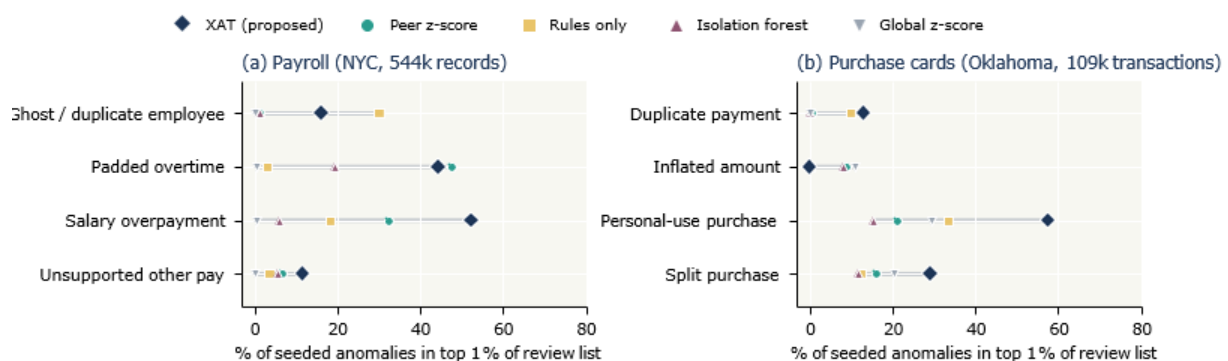


Fig 2: Share of each seeded scheme surfaced within a 1% review budget. Each row compares five methods; the gray bar spans their range. XAT (navy diamond) combines the scheme-specific strengths of rules and peer baselines.

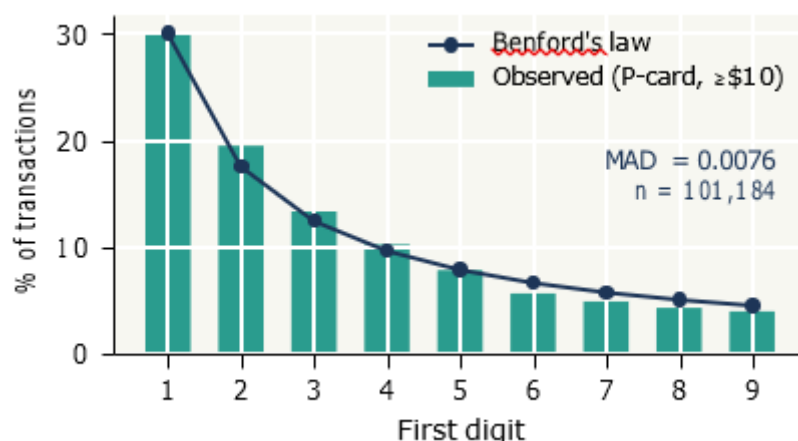


Fig 3: First-digit test of 101,184 Oklahoma purchase-card amounts of \$10 or more against Benford's law.

card rule, mostly weekend use and possible duplicates, far more than any audit team can review. This is the practical case for ranking: XAT orders exceptions by combined evidence, so the team reviews the most unusual first.

Two population-level tests found no red flags. The first-digit distribution of card amounts conforms closely to Benford's law (Fig. 3; mean absolute deviation 0.0076, within Nigrini's acceptable-conformity range^[14]). Average weekday card spending in the last two weeks of the fiscal year

(\$635,000) was slightly below the April–May average (\$692,000), so there was no year-end spending surge.

5.3. Explanation Cards

Fig. 4 shows two explanation cards for seeded anomalies, illustrating what an auditor would see. The payroll case

triggered no rule, but its overtime pay was extreme relative to its 24 peers, which placed it 102nd of 544,338 records. The card case triggered two rules and was ranked 93rd of 109,011 transactions. In both cases the card states the evidence, so the auditor can decide whether the explanation is plausible and which procedure to perform.

Table 3: Revenue-Recognition Module, Test Years 2003–2008

Method	Interpretable	AUC (s.d.)	Capture@5%
DSRI alone	yes	0.505 (0.037)	2.5%
Isolation forest	no	0.516 (0.066)	5.4%
Beneish revenue index (DSRI+SGI+GMI)	yes	0.553 (0.031)	3.8%
XAT sparse logistic (glass box)	yes	0.647 (0.033)	11.2%
Gradient boosting (black box)	no	0.671 (0.054)	18.9%

5.4. Revenue-Recognition Module

Table III reports the revenue module. The glass-box sparse logistic model achieves a mean AUC of 0.647 and outperforms the Beneish-style revenue index (0.553) in every one of the six test years (Fig. 5b). DSRI alone is no better than chance (0.505). The black-box gradient-boosting model is moderately better (AUC 0.671; 18.9% versus 11.2% of frauds captured in the riskiest 5% of firms). That gap is the price of transparency on this task: auditors can choose the glass-box model when the reason for each score must be documented, or use the black-box model as a secondary screen.

Because the model is linear, its reasoning can be read directly from Fig. 5a, and it is stable across years. The share of soft assets (assets other than cash and property, plant, and equipment) is the strongest red flag, followed by the change

in receivables and the change in cash sales, consistent with [4]. Once these are included, DSRI receives a small negative weight, a reminder that individual ratios should not be read in isolation. The sales growth index is nearly zero and was dropped entirely in one year.

6. Discussion

6.1. Supporting, Not Replacing, Judgment

XAT is designed so that the auditor remains the decision-maker. Each exception carries evidence the auditor can verify against source documents, which supports documentation requirements [5, 6] and reduces both blind reliance and unwarranted dismissal of flags [7, 8]. The results show that this transparency need not cost detection power: in payroll and card data, the transparent combination outperformed a black-box isolation

PAYROLL EXCEPTION				Rank 102 of 544,338
Record: Associate Staff Analyst, Dept. of Citywide Administrative Services, annual pay basis				
Peer group: same agency, title, and pay basis ($n = 24$)				
Rules triggered: none				
Why flagged	Value	Peer median	z	
Overtime pay	\$58,858	\$1	>100	
Overtime hours	799	0	15.8	
Other pay	\$7,671	\$3,153	1.4	
Suggested follow-up: obtain timesheets and supervisor overtime approvals; compare with access or dispatch records.				

PURCHASE-CARD EXCEPTION				Rank 93 of 109,011
Transaction: \$1,428.80, Saturday, merchant category "hotels" (Meridien)				
Agency: University of Oklahoma; this category makes up under 0.01% of the agency's card transactions				
Rules triggered: E3 weekend or state holiday; E4 rare merchant category				
Why flagged				
Merchant-category rarity in agency		32.2		
Amount vs. agency $\times$ category peers		2.5		
Amount vs. cardholder's usual spending		1.9		
Suggested follow-up: request the itemized receipt and travel authorization; confirm business purpose with the approver.				

Fig 4: Explanation cards for two seeded anomalies, as presented to an auditor. Each card states the rules triggered, the deviation from peers, and a suggested follow-up procedure; the auditor judges the evidence. Personal identifiers are omitted.

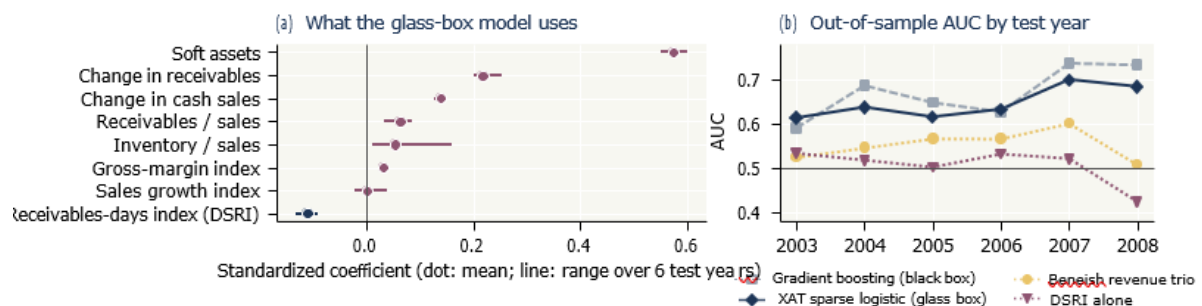


Fig 5: Revenue module. (a) Standardized coefficients of the glass-box model; dots show the mean and lines the range across the six yearly models, so an auditor can see both the direction and the stability of each red flag. (b) Out-of-sample AUC by test year.

forest by a wide margin. In the revenue module, transparency carried a moderate cost, which auditors can weigh explicitly.

6.2. Practical Implications

For internal auditors and inspectors general, the toolkit can run over complete payroll and card populations each period at negligible cost, using thresholds the audit team sets and documents. Rule exception volumes in real data, 3.7% of payroll records and 22.1% of card transactions, show that rules alone overwhelm review capacity; peer-based ranking is what makes population-level testing feasible. The same design applies to private-sector payroll and expense systems, which contain equivalent fields.

6.3. Limitations

Seeded anomalies are a controlled test, not a substitute for confirmed fraud, and some schemes were designed with the same red flags the rules encode; detection rates for those schemes therefore measure how well XAT surfaces known patterns amid real-world noise, not its ability to discover new schemes. Real exceptions were not investigated, so their fraud rate is unknown. Public payroll data report annual totals rather than pay-period detail, which limits tests such as duplicate payments within a period. Rule thresholds are illustrative and should be set from each organization's policies. Finally, the revenue module covers 2003–2008, and its performance on recent filings should be verified.

7. Conclusion

This paper developed XAT, an explainable AI audit toolkit that combines transparent rules, peer-group baselines, and glass-box models, and evaluated it on real U.S. payroll, purchase-card, and financial-reporting data. Within a 1% review budget, XAT surfaced 31.1% of seeded payroll anomalies and 26.0% of seeded card anomalies, outperforming rules alone and a black-box isolation forest, and its glass-box revenue model came close to a black-box model on SEC-identified fraud. By attaching verifiable evidence to every flag, the toolkit helps auditors use complete-population data while keeping professional judgment at the center of the audit. Future work will test the toolkit prospectively with audit teams, using investigated exceptions as labels, and extend it to pay-period and journal-entry data.

References

- American Institute of Certified Public Accountants, *Guide to Audit Data Analytics*. Durham, NC, USA: AICPA, 2017.
- Appelbaum D, Kogan A, Vasarhelyi MA. Big data and analytics in the modern audit engagement: Research needs. *Auditing: A Journal of Practice & Theory*. 2017;36(4):1-27.
- Association of Certified Fraud Examiners, *Occupational Fraud 2024: A Report to the Nations*. Austin, TX, USA: ACFE, 2024.
- Dechow PM, Ge W, Larson CR, Sloan RG. Predicting material accounting misstatements. *Contemporary Accounting Research*. 2011;28(1):17-82.
- Public Company Accounting Oversight Board, "AS 2110: Identifying and assessing risks of material misstatement," PCAOB Auditing Standards.
- Public Company Accounting Oversight Board, "AS 2401: Consideration of fraud in a financial statement audit," PCAOB Auditing Standards.
- Parasuraman R, Manzey DH. Complacency and bias in human use of automation: An attentional integration. *Human Factors*. 2010;52(3):381-410.
- Commerford BP, Dennis SA, Joe JR, Ulla JW. Man ver-sus machine: Complex estimates and auditor reliance on artificial intel-ligence. *Journal of Accounting Research*. 2022;60(1):171-201.
- Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*. 2019;1(5):206-215.
- Lundberg SM, Lee SI. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems* 30. 2017:4765-4774.
- Bao Y, Ke B, Li B, Yu YJ, Zhang J. Detecting accounting fraud in publicly traded U.S. firms using a machine learning approach. *Journal of Accounting Research*. 2020;58(1):199-235.
- Schreyer M, Sattarov T, Borth D, Dengel A, Reimer B. Detec-tion of anomalies in large scale accounting data using deep autoencoder networks. arXiv:1709.05254. 1709.
- Nigrini MJ, Mittermaier LJ. The use of Benford's law as an aid in analytical procedures. *Auditing: A Journal of Practice & Theory*. 1997;16(2):52-67.
- M. J. Nigrini, *Benford's Law: Applications for Forensic Accounting, Au-diting, and Fraud Detection*. Hoboken, NJ, USA: Wiley, 2012.
- Liu FT, Ting KM, Zhou ZH. Isolation forest. *Proc. 8th IEEE Int. Conf. Data Mining (ICDM)*. 2008:413-422.
- Beneish MD. The detection of earnings manipulation. *Financial Ana-lysts Journal*. 1999;55(5):24-36.
- Leys C, Ley C, Klein O, Bernard P, Licata L. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*. 2013;49(4):764-766.
- Tibshirani R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*. 1996;58(1):267-288.
- NYC Office of Payroll Administration, "Citywide Payroll Data (Fis-cal Year)," NYC Open Data, <https://data.cityofnyc.gov/d/k397-673e>, accessed Oct. 2026.
- Oklahoma Office of Management and Enterprise Services, "Purchase Card (PCard) Fiscal Year 2024," <https://data.ok.gov/dataset/purchase-card-pcard-fiscal-year-2024>, accessed Oct. 2026.
- F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

How to Cite This Article

Fuadedzi P, Developing explainable Artificial-Intelligence audit tools to identify Revenue-Recognition, payroll, Expense-Management, and Financial-Reporting risks while supporting professional audit judgment. *J Front Multidiscip Res*. 2026;7(2):340-345.

Creative Commons (CC) License

This is an open access journal, and articles are distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0) License, which allows others to remix, tweak, and build upon the work non-commercially, as long as appropriate credit is given and the new creations are licensed under the identical terms.