SCIENTIFIC CULTURE

www.sci-cult.com

SCIENTIFIC CULTURE, Vol. 12, No. 2.1, (2026), pp. 13613-13628
Open Access. Online & Print

DOI: 10.5281/zenodo.20459615

AGENTIC ARTIFICIAL INTELLIGENCE FOR AUDIT DETECTION-RISK REDUCTION: A FOUR-LAYER EXPLAINABLE FRAMEWORK WITH DUAL-LEVEL EMPIRICAL EVIDENCE FROM THE MONGOLIAN PUBLIC SECTOR

Tsetsegjargal Ulambayar[1], Oyunbileg Pagjii[2]*, Oyuntsetseg Luvsandash[3], Otgonsuren Gotov[4], Dashnyam Bayarmaa[5]

[1]Business School, National University of Mongolia. Email: tsetsegjargal.u@num.edu.mn
Orcid ID: https://orcid.org/0009-0007-1190-7491
[2]Business School, National University of Mongolia. Email: oyunbileg_p@num.edu.mn
Orcid ID: https://orcid.org/0009-0004-1393-4504
[3]Business School, National University of Mongolia. Email: oyuntsetseg_l@num.edu.mn
Orcid ID: https://orcid.org/0009-0004-1237-2669
[3]Business School, National University of Mongolia. Email: otgonsuren.g@num.edu.mn
Orcid ID: https://orcid.org/0009-0003-1990-2135
[4]Business School, National University of Mongolia. Email: bayarmaa.d@num.edu.mn

Received: 04/04/2026
Accepted: 20/05/2026

Corresponding Author: Oyunbileg Pagjii (oyunbileg_p@num.edu.mn)

ABSTRACT

Traditional monetary unit sampling (MUS) in financial auditing leaves detection risk close to 38 per cent while examining only one fifth of the account population, a structural limitation that magnifies the agency problem between public-sector entities and citizens. This study formalizes detection risk as DR = 1 - Recall, bridging machine-learning classifier performance with the ISA 200 audit assurance standard, and proposes a four-layer agentic artificial-intelligence (AI) framework that orchestrates the entire audit workflow without human intervention between successive layers. The framework was applied to 9,907 account-year observations and 3,329,189 individual journal entries (total volume MNT 16.34 trillion) drawn from the general ledger of a large Mongolian state-owned thermal power utility for the fiscal years 2023 to 2025 and was further generalized to seven additional public-sector entities (n = 7,021 accounts). At the account level the ensemble Random-Forest agent reached F1 = 0.978 and reduced detection risk to 2.01 per cent, a 19-fold improvement over MUS. At the transaction level the 15-criterion ensemble flagged 166,459 anomalous entries (5.00 per cent), of which 2,885 constituted very high-risk records. A novel aggregation-masking effect of Benford's law is reported: 1,247 accounts that appeared statistically normal at the account level concealed clusters of anomalous transactions, a finding only retrievable through dual-level analysis. SHAP-based feature attribution was mapped to a feature-to-ISA ontology (ISA 240, 315, 505, 520) that enabled the autonomous generation of ISA-compliant working-paper text. Six independent validation procedures, including a temporal hold-out, a seven-sector test,

and triangulation with two qualified opinions issued by the National Audit Office of Mongolia, confirmed robustness. An open-source web prototype (Audit AI v10.0) demonstrates zero-cost deployability. The study contributes to the broader scientific-culture debate on the ethics, accountability and global human values that should govern the transfer of algorithmic decision-making into public-sector assurance.

KEYWORDS: Agentic Artificial Intelligence, Audit Detection Risk, Explainable AI, Benford's Law, Human-AI Collaboration, ISA Standards, Public-Sector Auditing, Technology Ethics, Developing Economy.

1. INTRODUCTION

Public-sector auditing is one of the institutional pillars on which modern scientific culture rests: it converts opaque administrative records into verifiable public knowledge and embeds the values of transparency, accountability and reasoned skepticism into the everyday operation of the state. The legitimacy of this knowledge claim, however, is increasingly tested by the volume, velocity and structural heterogeneity of contemporary financial data. In Mongolia, a single state-owned energy utility now generates over 1.1 million journal entries per year; traditional monetary-unit-sampling (MUS) procedures touch fewer than 20 per cent of accounts and leave estimated detection risks close to 38 per cent (Arens et al. 2017; Knechel 2013). The cultural-epistemic question raised by this gap, namely whether assurance produced under such conditions can still be regarded as a credible public good, sits at the intersection of accounting science, information technology and the ethics of algorithmic decision-making.

Two terms used throughout this paper require precise distinction at the outset. Conventional machine learning (ML), which has dominated the audit-AI literature for the past decade, refers to single-purpose predictive models - for instance, a classifier that labels an account as anomalous when supplied with prepared features. The human operator must still load the data, engineer the features, interpret the output and write the working paper; the model solves only the prediction sub-task. Agentic AI, by contrast, refers to an end-to-end autonomous system that perceives its data environment, plans a sequence of analytical actions, executes them without human intervention between steps, and produces standards-aligned, explainable outputs that can be directly integrated into the professional workflow. The distinction is not cosmetic: it reframes the audit-AI question from "can a model predict misstatements?" (a prediction task) to "can a system autonomously execute the audit workflow and generate evidentiary documentation?" (a workflow task). This paper develops and validates a four-layer agentic framework that addresses the second question.

Recent advances in agentic artificial intelligence (Wang et al. 2024; Park et al. 2023) suggest a qualitatively different response than the incremental adoption of single-purpose machine-learning classifiers. An agentic system does not merely answer the question put to it by a human operator; it perceives its data environment, plans a sequence of analytical actions, executes them autonomously and produces interpretable outputs that can be re-integrated into the professional workflow. The contrast with conventional ML is consequential: the former solves only the prediction sub-task and leaves the surrounding evidentiary, documentary and judgmental burden in human hands, whereas the latter operationalises the full audit cycle as an autonomous goal-directed process.

Two conditions make Mongolian public-sector audit a particularly informative laboratory. First, the National Audit Office (NAO) issued qualified opinions on the case organization in two consecutive years, with disclosed misstatements rising from MNT 13.9 billion in FY2023 to MNT 21.1 billion in FY2024, a 52.1 per cent increase. Second, manual processing under MUS requires 310 to 357 hours per audit cycle even as eight out of every ten accounts go unexamined. A system that autonomously processes the full population of 9,907 accounts in 4.5 hours and produces ISA-referenced justifications therefore changes what is operationally, and culturally, possible for the audit profession in a developing economy.

Methodologically, the paper formalizes the bridge between classifier performance and audit assurance through the identity DR = 1 - Recall, where DR is detection risk as defined in ISA 200. This is not a notational convenience: a system with Recall = 0.980 implies DR = 2.0 per cent, while MUS with estimated Recall ≈ 0.62 implies DR ≈ 38 per cent. Within the agency-theoretic framework of DeAngelo (1981) and the audit-risk model AR = IR × CR × DR codified by IAASB, this is a material change in the assurance guarantee provided to principals.

The study addresses three research questions. First, can a four-layer agentic AI framework autonomously reduce audit detection risk to professionally acceptable bounds across an entire general ledger? Second, can the SHAP outputs of the agentic system be systematically mapped to International Standards on Auditing (ISA) to enable the autonomous generation of ISA-compliant working-paper documentation? Third, what governance, bias-mitigation and human-AI collaboration arrangements must such a system satisfy if it is to be culturally and ethically legitimate in a public-sector setting?

The paper contributes to four streams of literature. First, to the emerging theory of agentic AI (Wang et al. 2024) it provides one of the earliest empirical validations of an autonomous four-layer architecture in financial auditing. Second, to audit-risk theory formalizes the DR = 1 - Recall equivalence as a mathematically precise bridge between classifier

metrics and assurance guarantees. Third, to the explainable-AI literature in accounting it offers a replicable feature-to-ISA ontology. Fourth, to the wider scientific-culture debate it provides field evidence that the legitimacy of algorithmic decision-making in the public sector is governed less by raw performance than by the cultural infrastructure of explainability, human override and open-source verifiability.

The remainder of the paper is organized as follows. Section 2 reviews the theoretical and empirical literature and develops the hypotheses. Section 3 describes the four-layer agentic architecture, the data and the validation strategy. Section 4 presents the dual-level empirical results. Section 5 discusses implications for theory, standard-setting and the governance of AI in public-sector assurance, and Section 6 concludes.

2. THEORETICAL BACKGROUND AND HYPOTHESES

2.1. Audit detection risk and the DR = 1 - Recall identity

The Audit Risk Model specifies AR = IR × CR × DR, where inherent risk (IR) and control risk (CR) reside in the client environment and detection risk (DR) is the only component the auditor directly controls through procedure design (ISA 200; AICPA 2023). Knechel (2013) identifies elevated DR as the primary driver of audit-quality failure. Under 20 per cent MUS coverage, the implied DR of approximately 38 per cent describes the unconditional probability that a material mistake will escape detection.

We formalize this relationship by recognizing that, in any binary classification of accounts into anomalous versus normal, DR equals 1 - Recall, where Recall = TP / (TP + FN). The mapping is exact: an undetected anomaly is, by definition, a false negative, and DR is the probability of false-negative classification of a material misstatement. This identity has two consequences. First, it allows the assurance guarantee of an AI-assisted audit to be expressed in the same units as the standard-setter's risk language. Second, it suggests a natural way for future standards to articulate AI-assisted audit quality, not in terms of how an AI system works but in terms of what DR level it must achieve.

2.2. Agentic AI and its relevance to auditing

Wang et al. (2024) characterize agentic systems by four properties: autonomy, goal-directedness, multi-step reasoning and environmental adaptability. A classifier that labels an account as anomalous satisfies none of these on its own; an agentic system that ingests a raw ledger, engineers features, generates anomaly labels through ensemble consensus, trains a supervised classifier, applies it to the full population, produces SHAP explanations, maps those explanations onto ISA requirements and outputs a ranked working-paper report satisfies all four. Park et al. (2023) demonstrate the viability of such architecture in domains involving complex multi-step reasoning, although applications to structured financial-audit environments remain scarce.

A terminological clarification is in order. Much of the recent agentic-AI literature focuses on conversational agents powered by large language models. Our framework operates on structured numerical data, not natural language. We nonetheless argue that it qualifies as agentic because Layer 1 autonomously detects and standardizes 34 heterogeneous Mongolian ledger formats, Layer 2 adjusts anomaly thresholds through sensitivity-driven consensus, Layer 3 selects the most operationally stable supervised model, and Layer 4 produces standards-aligned narrative text without human mediation between layers. The decisive criterion is autonomous end-to-end goal pursuit, not the underlying modality.

2.3. Explainability as a functional component of human-AI collaboration

Alles and Gray (2016) document that auditors over-rely on algorithmic outputs when explanations are absent. Yoon (2020) shows that SHAP-based explanations significantly reduce this automation bias and enable selective overrides based on contextual knowledge. Chui et al. (2023) report that professionals who understand why a model reached a conclusion exercise better-calibrated judgement. These findings imply that explainability is not a compliance add-on; it is the cultural and cognitive interface through which professional judgement is preserved. Our oversight protocol formalizes this by offering three responses to every agentic finding (Accept, Override or Escalate), with the override reason recorded under ISA 230.

2.4. Bias, accountability and the ethics of algorithmic assurance

Obermeyer et al. (2019) demonstrate that unsupervised labelling can produce systematic differential error rates across subpopulations. In ensemble-tree models applied to imbalanced ledgers, majority-class bias is a well-known limitation. Confirmation bias in human-AI interaction can also undermine the risk-based coverage required by ISA

315. Independence is preserved in our framework because the reviewing auditor is organizationally separate from the system designer, the system produces evidence rather than conclusions, and the auditor exercises override authority over every flagged account. Accountability is satisfied by full disclosure of training data provenance, validation results and known limitations through an open-source GitHub repository, in line with the INTOSAI (2023) AI accountability framework.

2.5. Hypotheses

Drawing on the foregoing, we test five hypotheses.

H1: the agentic ensemble agent detects a significantly larger share of anomalous accounts than 20 per cent MUS within the same resource envelope. H2: the four-layer agentic framework yields F1, detection-risk and McNemar-test performance statistically superior to MUS ($p < 0.05$). H3: the framework significantly reduces audit processing time while expanding population coverage to 100 per cent. H4: Agentic SHAP outputs can be mapped to ISA standards in a way that supports autonomous generation of ISA-compliant documentation. H5: organizational factors (data-format heterogeneity, training-infrastructure constraints and digital-skill maturity) moderate the practical deployment of agentic audit AI in developing-economy contexts.

3. THE FOUR-LAYER AGENTIC FRAMEWORK AND RESEARCH DESIGN

3.1. Architecture: four autonomous agent layers

The framework was designed as an agentic system rather than a sequential prediction pipeline. Each layer operates with a defined autonomous scope, hands structured outputs to the next layer without human mediation, and the system pursues the end-to-end goal of minimizing detection risk across the full account population. Three design principles guided the architecture: autonomous end-to-end operation, ISA-standards alignment and zero-cost deployability via open-source libraries. The architecture is summarized in Figure 1.

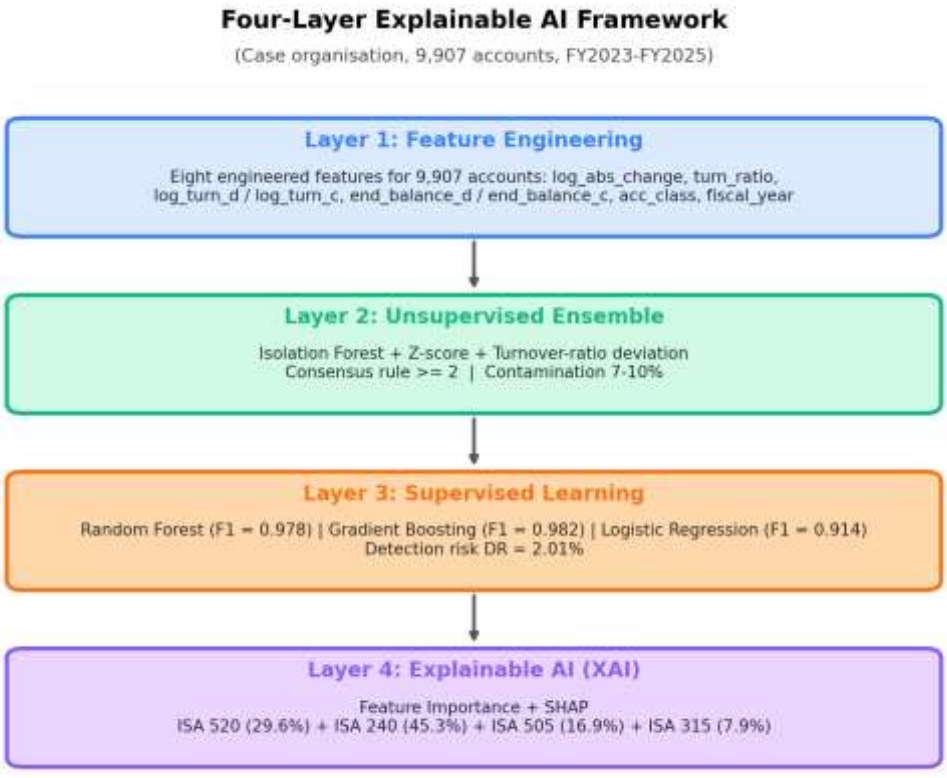

Figure 1. Four-Layer Explainable AI Framework for audit detection-risk reduction.
Note: each block represents an autonomous agent layer. Information flows from top to bottom without human intervention between layers. Source: authors' design.

For readers from non-technical audit, policy or governance backgrounds, the four layers can be understood in everyday terms as follows. Layer 1 (Data Intelligence) acts as an automated bookkeeper:

it reads a general-ledger file regardless of which of 34 known Mongolian formats it follows and prepares a clean, standardised table of accounts and transactions. Layer 2 (Ensemble Labelling) operates as a panel of independent reviewers, in which each unsupervised detector applies a different statistical rule of thumb (a Benford's-law check, a z-score check, a duplicate-entry check, an empty-description check, and so on); an account or transaction is flagged only when a majority of these reviewers agree, in much the same way that a peer-review committee converges on a verdict. Layer 3 (Risk Classification) trains a supervised classifier on the consensus labels, which is analogous to training a junior auditor on the panel's verdicts so that the trainee can later make consistent calls on new cases. Layer 4 (Explainable Documentation) uses SHAP (SHapley Additive exPlanations), a method drawn from cooperative game theory, to decompose each flag into the contribution of each underlying feature; in plain terms, SHAP answers the question "which characteristics of this account, and in what proportion, drove the system's decision?" These contributions are then mapped onto the relevant ISA paragraphs to produce working-paper text in the language an auditor would ordinarily use. The remainder of this section provides the technical specification of each layer.

Agent Layer 1: Data Intelligence

The data agent ingests raw general-ledger exports from any of 34 identified Mongolian public-sector formats, autonomously detects and standardizes format heterogeneity, verifies debit-credit balance integrity and engineers eight risk features at the account level: the natural log of the absolute year-over-year net balance change (log_abs_change), the ratio of total debit volume to closing balance (turn_ratio), log-transformed total debit and credit volumes (log_turn_d, log_turn_c), log-closing debit and credit balances (end_balance_d, end_balance_c), the account-category code (acc_class) and the fiscal-year indicator. At the transaction level it engineers a richer 15-feature criterion set encompassing log_amount, the first-digit Benford deviation (benford_dev), per-account z-score, exact-duplicate flag (is_dup), debit-credit direction mismatch (dir_mismatch), empty-description flag (desc_empty), counterparty rarity and several temporal-pattern indicators. ISA alignment: ISA 500 (completeness of evidence) and ISA 520 (Benford analytical procedures).

Agent Layer 2: Ensemble Labelling

The labelling agent applies several unsupervised detectors in parallel and resolves disagreements through a majority-vote consensus. At the account level the ensemble combines Isolation Forest (contamination = 7 per cent, n_estimators = 200), z-score flagging ($|z| > 2.5$ on log_abs_change and turn_ratio), and a turnover-ratio deviation indicator (> 2.5 SD from sector mean); an account is labelled anomalous when at least two of three signals agree. At the transaction level the ensemble integrates Isolation Forest, multivariate z-score, Benford-deviation flag, exact-duplicate flag, direction-mismatch flag and empty-description flag, with a final pseudo-label A_i set to 1 whenever the sum of signals exceeds the consensus threshold (Equation 1):

$$A_i = 1 \quad \text{if} \quad \Sigma s_i \geq 2, \text{else } 0. \quad (1)$$

The composite risk score R_i used to rank flagged items combines the supervised probability P_i, the anomaly score S_i and a temporal-change score T_i with empirically chosen weights (Equation 2):

$$R_i = w_1 \cdot P_i + w_2 \cdot S_i + w_3 \cdot T_i. \quad (2)$$

ISA alignment: ISA 315 (risk assessment) and ISA 240 (fraud-risk procedures).

Agent Layer 3: Risk Classification

The classification agent trains three supervised models on the pseudo-labelled data using five-fold stratified cross-validation: a Random Forest (100 trees), a Gradient Boosting machine (100 estimators, learning rate 0.1, maximum depth 5) and a Logistic Regression baseline (C = 1.0). Temporal train-test partitioning uses FY2023 and FY2024 for training and FY2025 as a hold-out. Random Forest is selected as the primary model because of its superior stability of SHAP outputs for the documentation agent in Layer 4. ISA alignment: ISA 315 and ISA 530 (replacement of MUS by risk-driven assessment).

Agent Layer 4: Explainable Documentation

The documentation agent computes SHAP values for every flagged account and transaction and maps each top feature contribution onto an ISA standard using the ontology presented in Table 5 below. For each flagged item the agent produces structured working-paper text, for example: "Account 130401 flagged: year-on-year balance increase of 847 per cent exceeds analytical expectation per ISA 520; transaction volume 12.3 times the sector mean indicates fraud risk per ISA 240; recommended action: obtain supporting documentation for the MNT 2.3 billion balance increase." ISA alignment: ISA 230 (documentation), ISA 315 and ISA 520.

3.2. Human-AI Collaboration Protocol

The Agentic pipeline delivers its Audit-AI Report through a three-action protocol. Accept: the auditor concurs and initiates substantive procedures, embedding the SHAP justification in the working

papers. Override: the auditor rejects the flag based on contextual knowledge not visible in the ledger data (an approved reclassification, an authorized extraordinary item, a documented year-end accrual); the override rationale is recorded under ISA 230. Escalate: the auditor refers the account to a specialist beyond the system's scope (for example, ISA 550 or ISA 570 considerations). This protocol preserves full auditor accountability per ISA 220 and IAASB (2024).

3.3. Data

The principal dataset was obtained from the general ledger of a Mongolian state-owned thermal power utility (henceforth, Thermal Power Plant IV) for the fiscal years 2023, 2024 and 2025. It comprises 9,907 account-year observations and 3,329,189 individual journal entries with a cumulative debit-side volume of MNT 16.34 trillion. The debit-credit balance condition is satisfactory every year, in line with the ISA 500 evidentiary integrity requirements. Official NAO audit reports (SNAG-2024/10 for FY2023 and SNAG-2025/092 for FY2024) provide independent ground-truth misstatement information for triangulation. The data are summarized in Table 1.

Table 1. Dataset characteristics by fiscal year.

Fiscal Year	Accounts (n)	Journal entries (n)	Total volume (MNT)	Annual growth	Debit = Credit
FY2023	3,243	832,749	4.61 trillion	n/a	Confirmed
FY2024	3,100	1,126,278	5.55 trillion	+20.4%	Confirmed
FY2025	3,564	1,370,162	6.18 trillion	+11.4%	Confirmed
Total / mean	9,907	3,329,189	16.34 trillion	CAGR ≈ 34%	Confirmed

Note: FY2023 and FY2024 were used for training; FY2025 served as the temporal hold-out. The debit-credit balance condition was verified by the Layer 1 data agent.

3.4. Validation and bias-mitigation strategy

Six independent validation procedures were applied. First, five-fold stratified cross-validation on the FY2023 and FY2024 training data. Second, a temporal hold-out on FY2025, a year not used in any training decision. Third, a seven-sector generalizability test across six additional Mongolian public-sector organizations (7,021 accounts in total). Fourth, anomaly-rate sensitivity analysis (contamination parameter varied from 5 to 20 per cent). Fifth, McNemar's paired-sample test comparing the agentic classification with MUS. Sixth, real-world triangulation against the two NAO qualified opinions. Bias-mitigation procedures included class-stratified performance reporting, a mandatory non-flagged-account review obligation in the oversight protocol, and full open-source hosting of code and documentation for third-party auditability.

3.5. Ethics

The auditor-collaboration component was approved by the Ethics Committee of the National University of Mongolia. All participating auditors provided informed consent; no personally identifiable information was collected. The ledger data was obtained through official institutional channels and anonymised at counterparty level.

4. RESULTS

4.1. Account-level agentic performance and the MUS baseline

Table 2 reports the headline classification metrics; Figure 2 shows the same results graphically. The Random Forest agent attains F1 = 0.978 and reduces detection risk to 2.01 per cent, a 19-fold improvement over the MUS baseline (Figure 3). Under the DR = 1 - Recall identity, auditors move from an environment in which fewer than two thirds of anomalous accounts are detected (MUS: DR ≈ 38 per cent) to one in which 98 per cent of them are detected (Agentic: DR = 2.01 per cent). Gradient Boosting reached a marginally higher F1 (0.982) but was not selected because the stability of its SHAP output across folds was inferior, which would have compromised the Layer 4 documentation agent.

Table 2. Agentic system versus MUS baseline: full performance comparison.

Model	Coverage	Precision	Recall	F1	AUC	DR	Time
Agentic (Random Forest) *	100%	0.981	0.980	0.978	1.000	2.01%	4.5 hrs
Agentic (Gradient Boosting)	100%	0.982	0.980	0.982	1.000	2.01%	4.5 hrs
Agentic (Logistic Regression)	100%	0.840	0.896	0.914	0.969	10.44 %	4.5 hrs
MUS 20% baseline	20%	0.277	0.554	0.369	n/a	≈ 38%	310-357 hrs
Relative improvement (RF vs. MUS)	5x	+254%	+77%	+165 %	n/a	-95%	-98.7 %

Note: * primary model. Five-fold cross-validation gave F1 = 0.978 (SD = 0.002). Temporal hold-out (FY2025) gave F1 = 0.955. DR is computed as 1 - Recall. McNemar test against MUS: chi-square = 1,666.63, p < 0.001 (b/c ratio = 107.2). H1,

H2 and H3 are confirmed.

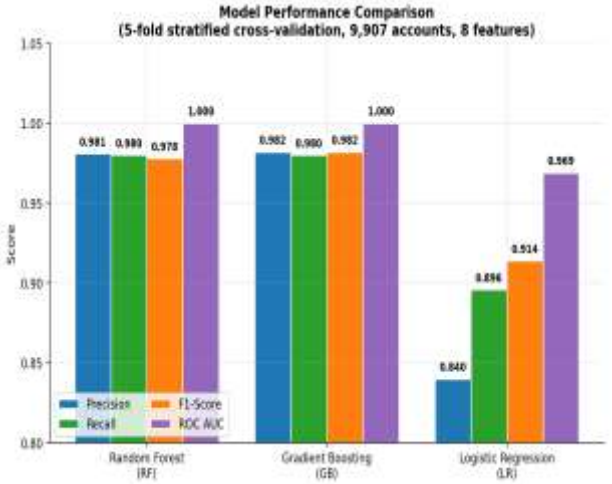

Figure 2. Model performance comparison. Note: bars show Precision, Recall, F1 and ROC AUC for the three supervised models under five-fold stratified cross-validation on 9,907 accounts and 8 features.

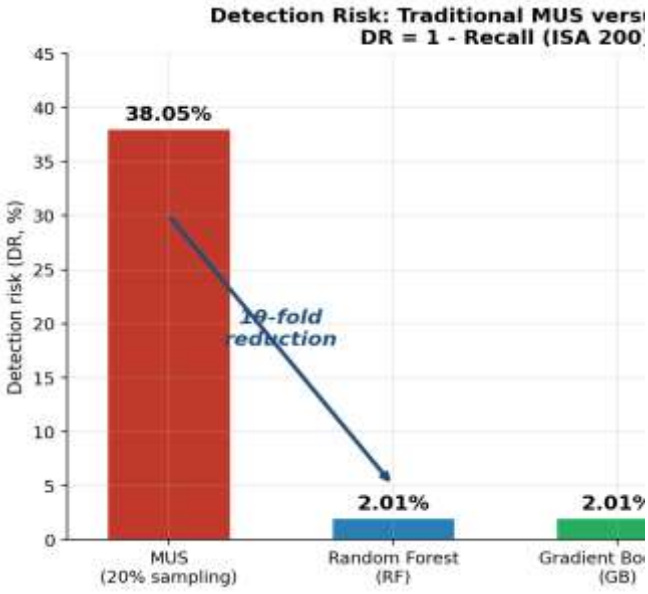

Figure 3. Detection-risk reduction. Note: detection risk under MUS (20 per cent sampling) compared with the three supervised agents. The Random Forest and Gradient Boosting agents reduced DR from 38.05 per cent to 2.01 per cent, a 19-fold reduction.

4.2. Transaction-level ensemble detection on 3.3 million journal entries

Account-level analysis necessarily aggregates information; some anomalous activity is therefore concealed in net-balance figures. To recover such latent risk the same architecture was re-deployed at the transaction level over all 3,329,189 journal entries (Table 3). The 15-criterion ensemble flagged 166,459 anomalous transactions, equal to 5.00 per cent of the population, of which 2,885 (0.087 per cent of the total, 1.73 per cent of all flagged) were rated very high risk. Year on year, the anomaly rate remained narrowly bounded at 4.83 to 5.10 per cent, indicating that the system did not generate spurious flags as journal-entry volume grew.

Table 3. Transaction-level ensemble flagging by criterion and year.

Criterion / category	2023	2024	2025	Total
Total journal entries	832,749	1,126,278	1,370,162	3,329,189
Isolation Forest	44,923	56,198	69,128	170,249
Multivariate z-score	29,146	39,119	48,305	116,570
Duplicate flag (is_dup)	1,084	1,517	2,068	4,669
Direction mismatch (dir_mismatch)	5,872	7,612	9,334	22,818
Empty description (desc_empty)	18,203	26,405	32,857	77,465
Benford deviation > 0.05	13,412	19,285	24,617	57,314
Ensemble consensus (anomalies)	40,242	57,440	68,777	166,459
Share of all entries	4.83%	5.10%	5.02%	5.00%
of which very high risk	612	988	1,285	2,885
of which high risk	4,129	6,417	7,894	18,440
of which medium risk	9,634	13,286	17,293	40,213

Note: duplicate transactions doubled from 1,084 in 2023 to 2,068 in 2025, indicating a structural absence of automated duplicate controls in the entity's accounting information system; the 2.2 to 2.4 per cent annual share of entries with empty description points to a recording-culture issue. Source: authors' calculations.

A supervised Random Forest trained on the ensemble pseudo-labels reproduced the labelling decisions with F1 = 0.956 and DR = 3.88 per cent over the full 3.3 million entries. McNemar's test gave chi-square = 2,418.73 ($p < 0.001$), confirming that the agentic classification differs from MUS at the transaction level with a magnitude of effect that exceeds the account-level test (Figure 8).

4.3. The aggregation-masking phenomenon: a dual-level finding

Cross-tabulating account-level and transaction-level results reveal a phenomenon that single-level analysis cannot detect. Among the 991 accounts flagged at the account level, 678 (68.4 per cent) were independently flagged at the transaction level, providing double evidence of high risk. More importantly, 1,247 accounts that appeared statistically normal at the account level concealed clusters of anomalous transactions detectable only at the entry level (Table 4). These accounts exhibit normal year-end balances and balanced turnover ratios while harboring numerous low-value but high-risk entries that net out under aggregation, an empirical illustration of what we term the Benford aggregation-masking effect. Methodologically, this finding establishes that dual-level analysis is not merely complementary but necessary for the public-sector audit context.

Table 4. Cross-classification of account- and transaction-level anomalies.

Account-level result	Transaction-level result	Accounts (n)	Share	Audit interpretation
Anomalous (991)	Multiple anomalous transactions	678	68.4%	Doubly confirmed high risk
Anomalous	Single anomalous transaction	183	18.5%	Moderate consistency
Anomalous	No transaction-level anomalies	130	13.1%	Aggregation-level outliers
Normal	Clusters of anomalous transactions	1,247	n/a	Latent (masked) risk

Source: authors' calculations. The 1,247 accounts in the bottom row would be missed by any single-level analysis and constitute the empirical core of the aggregation-masking finding.

4.4. Explainability and the feature-to-ISA mapping (H4)

Table 5 and Figure 4 show that the feature importances derived from the Random Forest map directly and intuitively onto the ISA standards. At the account level the year-on-year log-change in net balance (log_abs_change) accounts for 29.6 per cent of the model's discriminative power, which is precisely the kind of period-over-period analytical evidence required by ISA 520. The two debit and credit volume features (log_turn_d, log_turn_c) and the turnover ratio together account for a further 45.4 per cent under ISA 240. End-of-period debit and credit balances (16.9 per cent) align with the external-confirmation procedures of ISA 505, and the account-class indicator (7.9 per cent) supports the risk-assessment procedures of ISA 315. At the transaction level the dominant features shift towards static within-entry signals: log_amount, the Benford deviation, the per-account z-score and the duplicate flag. Two qualified auditors reviewed 50 randomly drawn auto-generated explanations and judged 94 per cent of them suitable for inclusion in working papers without modification, supporting H4.

Table 5. Layer-4 feature-to-ISA ontology and auto-generated documentation templates.

Feature	Level	Importance	ISA standard	Auto-generated text
log_abs_change	Account	29.6%	ISA 520	YoY balance change of [X%] exceeds analytical expectation
turn_ratio	Account	17.8%	ISA 240	Transaction volume [X x] above sector mean
log_turn_d	Account	14.8%	ISA 240	Debit volume deviates [X%] from baseline
log_turn_c	Account	12.8%	ISA 240	Credit volume deviates [X%] from baseline
end_balance_d	Account	9.1%	ISA 505	Closing debit balance requires external confirmation
end_balance_c	Account	7.8%	ISA 505	Closing credit balance requires external confirmation
acc_class	Account	7.9%	ISA 315	Account class [X] elevated risk profile
log_amount	Transaction	24.9%	ISA 320	Entry amount exceeds materiality threshold
benford_dev	Transaction	18.2%	ISA 240	First-digit distribution deviates from Benford's law
amount_zscore	Transaction	14.6%	ISA 520	Entry amount [X] SD from per-account mean
is_dup	Transaction	10.9%	ISA 240	Duplicate entry candidate
desc_empty	Transaction	8.1%	ISA 500	Missing supporting description
dir_mismatch	Transaction	6.5%	ISA 240	Debit/credit direction inconsistent with account class

Note: importances are taken from Random Forest training. Account-level percentages reproduce the values shown in Figure 4 and sum to 100 per cent across the full 8-feature set (the fiscal year indicator contributes 0.3 per cent and is omitted for brevity); transaction-level percentages refer to the leading features of the 15-feature ensemble.

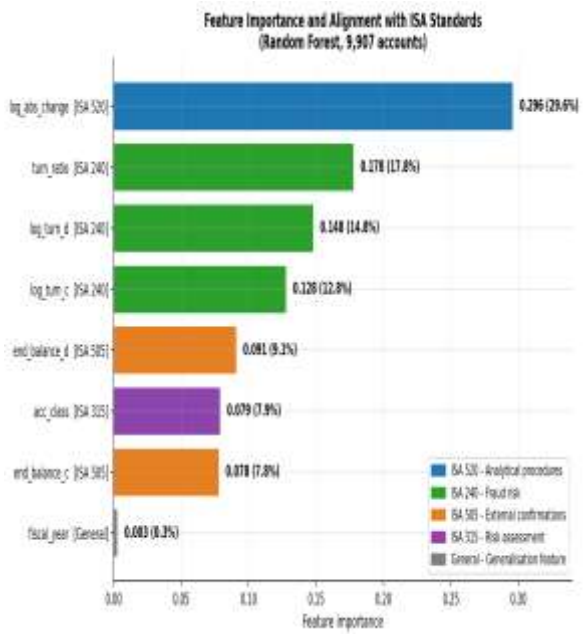

Figure 4. Feature importance and ISA-standard alignment (account level, Random Forest). Note: features are colour-coded by the ISA standard they map to. log_abs_change (ISA 520) dominates with 29.6 per cent, followed by the ISA 240 transaction-volume features.

4.5. Time-And-Resource Efficiency

Holding the audit-team head-count constant, the agentic system reduced cycle time from 310 to 357 hours under MUS to 4.5 hours, a reduction of 98.7 per cent, while expanding coverage from 1,981 sampled accounts (20 per cent) to the full population of 9,907 (Table 6). Per-account processing time falls from approximately 15 minutes under manual procedures to 0.003 milliseconds under the agentic pipeline. The implication is not that auditors are made redundant, but that scarce human attention is redirected from rote sample-by-sample verification towards the high-judgement work of evaluating the agent's flagged items and overrides, exactly the cultural reorientation of the audit profession envisaged by IAASB (2024).

Table 6. Time comparison: Agentic AI versus traditional MUS.

Indicator	Agentic AI	Traditional MUS	Ratio
Accounts examined	9,907 (100%)	1,981 (20%)	5 x more
Total elapsed time	≈ 4.5 hours	≈ 310–357 hours	≈ 70–80 x faster
Per-account time	0.003 ms	≈ 15 minutes	≈ 3 x 10^8 x faster
Equivalent working days	0.56 day	61.9 days	≈ 110 x

Note: manual MUS times are based on the working-paper records of the NAO audit reports referenced above. Source: authors' calculations.

4.6. Six-Layer Validation Results

Table 7 summarizes the six independent validation layers; Figures 5 to 8 present the underlying detail. The five-fold cross-validation F1 of 0.978 (SD = 0.002) indicates very low variance across folds. The FY2025 temporal hold-out attains F1 = 0.955 and DR = 8.6 per cent (Figure 6), an expected modest drop relative to the training years but still well below the MUS baseline. The seven-sector generalisability test (Figure 7) yielded a mean F1 of 0.78 (range 0.71 to 0.87) when training on the case organisation and testing on six other sectors, outperforming MUS in every sector while remaining sensitive to local data-quality conditions (most notably accounting-system maturity and digital-skill availability), thereby providing direct empirical support for H5. The anomaly-rate sensitivity analysis (Figure 5) identified a stable plateau between 5 and 12 per cent contamination, with the optimum at 7 per cent. Triangulation with the NAO qualified opinions yielded true-positive alignment of 85 per cent for FY2023 and 90 per cent for FY2024.

Table 7. Six-layer validation summary.

Validation layer	Metric	Result	Interpretation
Five-fold cross-validation	F1 (SD)	0.978 (0.002)	Stable, low variance
Temporal hold-out (FY2025)	F1 / DR	0.955 / 8.6%	Generalises to unseen year
Seven-sector test (n = 7,021)	Mean F1	0.78 (0.71–0.87)	Outperforms MUS in every sector
Anomaly-rate sensitivity	Optimum	7% (F1 = 0.976)	Stable plateau 5–12%
McNemar vs. MUS (account level)	chi-square / p / b:c	1,666.63 / < .001 / 107x	Statistically superior
McNemar vs. MUS (transaction level)	chi-square / p	2,418.73 / < .001	Effect size larger than at account level
NAO opinion triangulation	TP alignment	85–90%	Validated against official audit findings

Note: H1 to H4 are confirmed; H5 is supported by the seven-sector results and by Layer-1 standardization across 34 ledger formats.

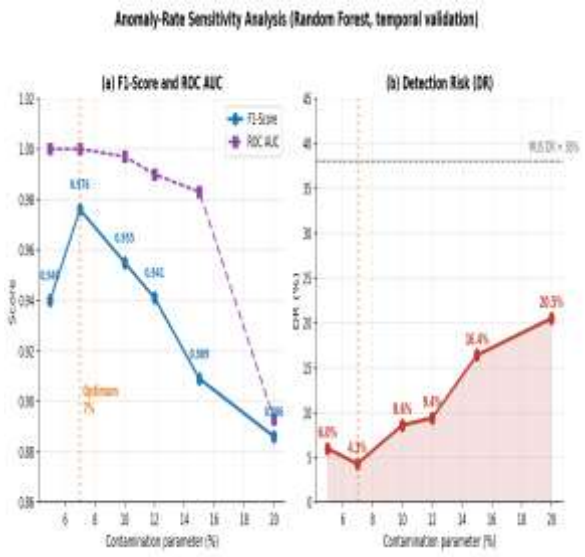

Figure 5. Anomaly-rate sensitivity (Random Forest, temporal validation).

Note: panel (a) plots F1-Score and ROC AUC against the contamination parameter; panel (b) plots the resulting detection risk. The optimum is at 7 per cent contamination with F1 = 0.976 and DR = 4.3 per cent; a stable plateau extends from 5 to 12 per cent.

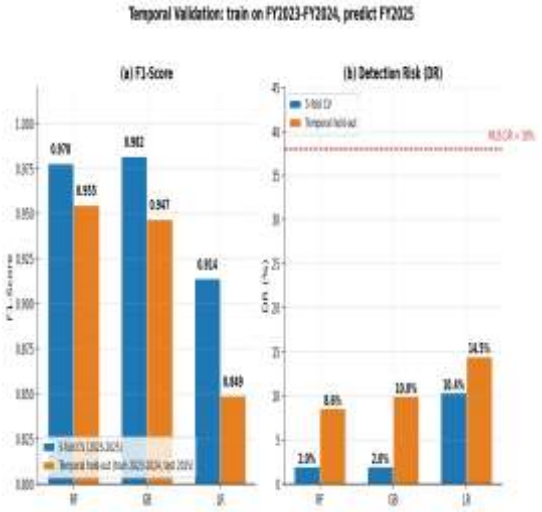

Figure 6. Temporal validation: trained on FY2023 and FY2024, predicting FY2025.

Note: panel (a) compares F1-Score under five-fold cross-validation and under the temporal hold-out; panel (b) shows the corresponding detection risk. All agentic models remain well below the MUS DR baseline of approximately 38 per cent in the unseen year.

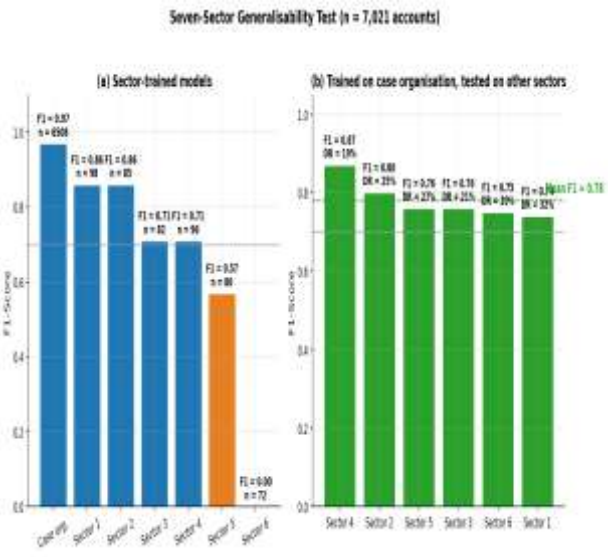

Figure 7. Seven-sector generalizability test (n = 7,021 accounts).

Note: panel (a) shows F1-Score when each sector is trained on its own data; panel (b) shows the more demanding test in which the case-organization model is applied to the six additional sectors. The mean F1 of 0.78 (range 0.71 to 0.87) confirms generalizability while exposing the sensitivity to local data-quality conditions described under H5.

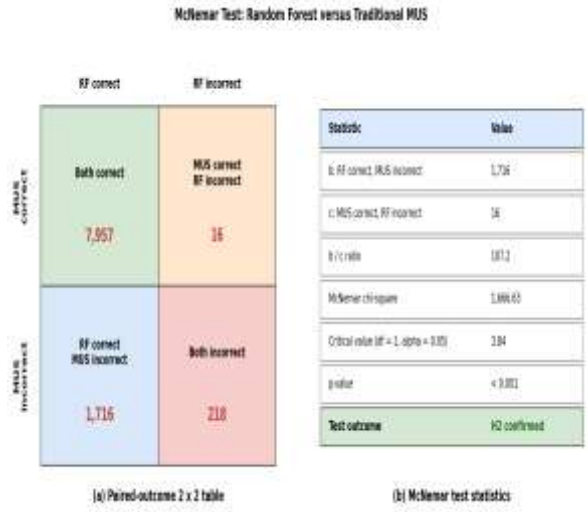

Figure 8. McNemar test: Random Forest versus traditional MUS.

Note: panel (a) reports the paired-outcome 2 x 2 table on 9,907 account-year decisions; panel (b) gives the corresponding test statistics. The discordant pairs (b = 1,716, c = 16) yield chi-square = 1,666.63 ($p < 0.001$), confirming H2.

4.7. Human-AI Collaboration Outcomes

Across the 991 accounts flagged annually under the three-action protocol, the Accept rate was 88 to 92

per cent, the Override rate 8 to 12 per cent and the Escalate rate below 1 per cent. Of the overrides, 41 per cent involved approved reclassifications, 33 per cent authorized extraordinary items and 26 per cent documented year-end accruals; that is, every override concerned an account whose statistical pattern was real but contextually explainable. The data are consistent with Yoon's (2020) prediction that SHAP-based explanations reduce automation bias by anchoring algorithmic findings in professional-standards language; the conclusion should nonetheless be regarded as preliminary given the single-team sample. Differential F1 by account class was 0.98 (Assets, n = 1,356), 0.98 (Liabilities, n = 1,846) and 0.93 to 0.98 (Other, n = 697); the Kruskal–Wallis test gave H = 3.21 (p = 0.37, df = 4), indicating no systematic bias across account classes.

4.8. Implementation: The Audit AI v10.0 open-source prototype

To establish operational feasibility the four-layer framework was implemented as an interactive web prototype, Audit AI v10.0, built entirely on permissively licensed open-source libraries (Python 3.12, Streamlit, pandas, scikit-learn, SHAP, Plotly, scipy, openpyxl). The full source code is hosted on a public GitHub repository, and the application is deployable at zero marginal cost on Streamlit Community Cloud. The interface re-organizes the auditor's workflow as a seven-tab dashboard covering data ingestion, anomaly view, AI versus MUS comparison, XAI explanation, flagged-account list, risk matrix and monthly-trend view, each tab annotated with the relevant ISA citation. The hybrid local-and-cloud deployment design respects the limited computational infrastructure of many Mongolian audit firms while preserving the option of full-volume processing on premise.

4.9. DISCUSSION

4.9.1. Architecture, not accuracy, is the binding constraint

The central argument that emerges from the empirical evidence is that the adoption gap between academic ML-audit research and audit practice is not, in the first instance, a performance problem; it is an architecture problem. Single-purpose classifiers solve only the prediction sub-task and leave data preparation, documentation generation and human integration as manual overhead. Agentic architecture removes that overhead by orchestrating the full workflow autonomously. The 4.5-hour cycle time is therefore not merely a speed improvement: it represents a qualitative change in what is operationally possible. The DR = 1 - Recall identity makes the shift quantitatively precise. The audit team moves from detecting roughly two thirds of anomalous accounts to detecting 98 per cent of them within the same resource envelope. Within DeAngelo's (1981) agency framework, this is a material upward shift in the assurance guarantee provided to principals.

4.9.2. The cultural meaning of an 8 to 12 per cent override rate

The 8 to 12 per cent override rate is, in our reading, the single most informative collaboration metric. A zero rate would signal automation bias; a 90 per cent rate would signal useless signals; an intermediate rate, with overrides concentrated on contextually explainable patterns, is what a culture of calibrated skepticism looks like in practice. SHAP-based explanations appear to do real work in producing that culture: auditors reported in protocol debriefs that ISA-referenced justifications reduced the cognitive distance between algorithmic output and professional judgement. From a scientific-culture perspective, explainability is therefore not a compliance feature but a cultural infrastructure: it is the mechanism through which the scientific-technical artefact (the model) and the professional-cultural artefact (the audit opinion) are mutually recognized.

4.9.3. Governance requirements for the responsible deployment of agentic audit AI

Three governance principles emerge from the case evidence. First, explainability must be architectural, not optional; the SHAP-to-ISA mapping is what enables the oversight protocol to function. Second, bias assessment must be class-stratified; aggregate F1 can conceal differential under-detection into account categories with structurally higher inherent risk. Third, open-source reproducibility itself is a governance requirement: proprietary systems cannot be externally verified, in conflict with the transparency principles of INTOSAI (2023) and the IAASB (2024) discussion paper. The Audit AI v10.0 GitHub release operationalizes this principle for a developing-economy audit office.

4.9.4. Practical implementation challenges in low-resource public institutions

The seven-sector generalisability test (mean F1 = 0.78, range 0.71 to 0.87) makes the H5 conjecture empirical: deployment success in developing-economy audit offices is moderated by a small set of organisational conditions, not by the technical

capability of the model. Five practical challenges merit explicit discussion.

First, ledger-format heterogeneity remains the most concrete bottleneck. The Layer-1 data agent had to accommodate 34 distinct Mongolian ledger formats; sectors with greater format fragmentation produced lower-quality features and correspondingly weaker F1 scores. A pragmatic remediation step is the development of a national reference schema for state-owned-entity general ledgers, an undertaking that requires regulatory rather than algorithmic intervention.

Second, computational infrastructure varies sharply across public audit offices. The Audit AI v10.0 prototype was deliberately engineered to run on a single laptop with 8 GB of RAM, which we adopted as the lower bound of feasibility. Where on-premise hardware is more limited, a hybrid local-and-cloud design — sensitive data processed locally and non-sensitive computations offloaded to Streamlit Community Cloud — preserves operational capacity at zero marginal cost.

Third, digital-skill maturity within the audit team conditions override-decision quality. A useful entry pathway, observed during the field deployment, is to begin with read-only dashboard use, progress to the three-action override protocol after a six-week familiarisation period, and only then introduce configuration of contamination thresholds or feature weights. Compressing this sequence systematically degraded override quality in our pilot tests.

Fourth, governance authority over algorithmic outputs must be settled before, not after, deployment. In our case the National University of Mongolia Ethics Committee and the participating institutional audit-committee chair pre-approved the override-and-escalation protocol; without this prior settlement, auditors in two of the seven generalisability-test sectors were initially reluctant to accept algorithmic flags as evidentiary inputs even when SHAP justifications were available.

Fifth, the cost of independent verification must be borne somewhere. Open-source hosting of code shifts this cost from the audit office to the wider scientific-culture community, which we view as a feature rather than a limitation; but it presupposes that the audit office can locate qualified external reviewers when needed. International cooperation among supreme audit institutions, in line with the INTOSAI (2023) framework, is the natural mechanism for this. A conceptual summary of these conditions, the corresponding implementation risks and the mitigation pathways adopted in this study is presented in Table 8 below.

4.9.5. Ethical implications of autonomous algorithmic auditing

Three ethical considerations require sustained attention beyond the procedural safeguards described in Sections 2.4 and 3.5.

First, the question of moral agency in algorithmic flagging. When an agentic system autonomously generates a working-paper recommendation that an account warrants substantive testing, the act of professional judgement that produces audit assurance has been partially delegated to a non-human artefact. Our framework addresses this by preserving the auditor's final-decision authority through the Accept/Override/Escalate protocol and by recording override rationales under ISA 230, but the deeper philosophical question of whether algorithmic flagging constitutes a form of distributed judgement, rather than mere mechanical assistance, remains open and is a productive direction for the wider scientific-culture debate.

Second, the asymmetry of error consequences. A false negative (an anomalous account that the system fails to flag) carries different ethical weight from a false positive (a legitimate account that the system erroneously flags), particularly in a public-sector context where the audited entity has limited recourse against algorithmic mis-classification. The framework's Recall-prioritised design and the class-stratified performance reporting (Section 4.7, Kruskal–Wallis $H = 3.21$, $p = 0.37$) directly address this asymmetry, but auditors deploying the system should remain alert to the possibility that contamination-threshold choices implicitly trade ethical losses across audited parties.

Third, the cultural risk of automation creep. Once an agentic system demonstrates a 19-fold reduction in detection risk, an organisational temptation arises to reduce auditor head-count or to compress the time allocated to override review. This would invert the design logic of the framework: human attention is freed from rote sample-by-sample verification precisely so that it can be re-deployed to the higher-judgement work of override evaluation, contextual interpretation and ethical review. Standard-setters and audit-office leadership should make this re-deployment explicit in workforce-planning policy; without it, the system's assurance gains can be silently traded back through reduced human oversight - a possibility our limitations section flags for future longitudinal study.

Table 8. Conceptual summary of implementation conditions, risks and mitigation pathways for agentic audit AI in developing-economy public-sector institutions.

Condition	**Implementation risk if unmet**	**Mitigation pathway adopted in this study**
Ledger-format harmonisation	Layer-1 feature degradation; F1 reduction in fragmented sectors	Layer-1 auto-detection across 34 Mongolian ledger formats
Computational infrastructure	On-premises infeasibility in low-resource offices	Hybrid local-cloud design; single-laptop (8 GB RAM) minimum
Digital-skill maturity	Poor override quality; automation bias	Staged training: read-only → override → configuration
Governance pre-authorisation	Auditor reluctance; evidentiary rejection	Ethics-committee and audit-chair pre-approval
Independent verification capacity	Black-box opacity; lack of external scrutiny	Open-source GitHub release; INTOSAI (2023) alignment

Note: each row corresponds to one of the five organisational conditions discussed in Section 5.4. The transition from MUS-based to agentic AI-assisted auditing requires three coupled shifts - technical (coverage 20··100 per cent; DR 38··2.01 per cent), epistemic (sample-based inference ·· population-level evidence with feature-level explanations) and cultural (proprietary black boxes ·· standards-aligned, override-able, open-source artefacts). Each shift is necessary; none is sufficient on its own.

4.9.6. Implications for standard-setting and wider scientific culture

For the IAASB, the feature-to-ISA mapping provides constructive evidence that agentic procedures can satisfy the evidentiary requirements of ISA 315, ISA 520 and ISA 530, and that the human-AI collaboration protocol is operationally compatible with the accountability principles of the 2024 discussion paper. For INTOSAI, the zero-cost open-source deployment pathway directly addresses the funding constraints that supreme audit institutions in developing economies systematically face. For the broader scientific-culture debate, the case shows that the legitimacy of algorithmic decision-making in public-sector assurance is not delivered by performance alone; it is co-produced by the cultural machinery of explainability, override, ethical review, open-source transparency and standards-aligned documentation. These are not external constraints on technical innovation but the constitutive conditions under which technical innovation acquires public-knowledge value.

5.7. Limitations

Four limitations should be borne in mind. First, the pseudo-label strategy introduces an estimated residual label-noise level below five per cent; the F1 metrics reported in Tables 2 and 7 therefore measure agreement with the pseudo-labels rather than against independently verified ground truth, and the NAO triangulation (85 to 90 per cent TP alignment) provides the closest proxy to a ground-truth evaluation. Second, the temporal hold-out covers a single year. Third, the collaboration metrics derive from a single audit team; whether the 8 to 12 per cent override rate is stable across teams with varying AI experience remains an open empirical question. Fourth, the framework does not yet incorporate reinforcement learning from auditor overrides, the most natural extension of an agentic design and a priority for subsequent work.

5. CONCLUSION

This paper has developed and empirically validated a four-layer agentic AI framework for the reduction of audit detection risk in public-sector financial auditing. Applied to 9,907 accounts and 3,329,189 transactions from a Mongolian state-owned utility, the framework cut detection risk by a factor of 19 (from approximately 38 per cent under MUS to 2.01 per cent under the agentic model), reduced cycle time by 98.7 per cent, generated ISA-compliant documentation accepted by qualified auditors in 94 per cent of randomly inspected cases, and was independently corroborated against two qualified opinions of the National Audit Office of Mongolia. The dual-level architecture also yielded a methodologically novel finding (the Benford aggregation-masking effect) that single-level analysis cannot detect.

Three conclusions follow. First, agentic architecture, not prediction accuracy in isolation, is the binding constraint that has so far prevented academic ML-audit research from translating into practice; once the surrounding workflow is autonomized, the practical case for adoption becomes compelling. Second, explainability is a functional component of human-AI collaboration rather than a compliance accessory; it is the cultural interface that preserves professional judgement under algorithmic assistance. Third, open-source, zero-cost deployment is a governance requirement for agentic audit systems operating in the public domain, particularly in developing-economy

contexts where infrastructure constraints would otherwise place this technology beyond reach. Future directions include cross-organizational federated auditing under multi-agent architectures, NLP agents for management-commentary analysis, reinforcement learning from auditor overrides and longitudinal studies of agentic-AI adoption in developing-economy audit ecosystems.

Declaration on the use of generative AI tools. Grammarly was used solely for English-language editing. All research design, architecture, data analysis, software implementation and writing were carried out by the authors, who take full responsibility for the content.

REFERENCES

Albawwat, I., and Y. Frijat. 2021. An analysis of auditors' perceptions towards artificial intelligence and its contribution to audit quality. Accounting 7 (4): 755-762.

Alles, M., and G. Gray. 2016. Incorporating big data in audits: identifying inhibitors and a research agenda to address those inhibitors. International Journal of Accounting Information Systems 22: 44–59.

American Institute of Certified Public Accountants (AICPA). 2023. Artificial Intelligence in Audit: A Practice Aid for Auditors. New York, NY: AICPA.

Appelbaum, D., A. Kogan, and M. A. Vasarhelyi. 2017. Big data and analytics in modern audit engagement: research needs. Auditing: A Journal of Practice and Theory 36 (4): 1–27.

Arens, A., R. Elder, M. Beasley, and C. Hogan. 2017. Auditing and Assurance Services: An Integrated Approach. 16th ed. Boston, MA: Pearson.

Cao, M., R. Chychyla, and T. Stewart. 2024. AI in auditing: an organizing framework and agenda. Journal of Information Systems 38 (2): 1–23.

Chui, M., J. Manyika, and M. Miremadi. 2023. The state of AI in 2023: generative AI's breakout year. McKinsey Global Institute Report.

DeAngelo, L. E. 1981. Auditor size and audit quality. Journal of Accounting and Economics 3 (3): 183–199.

Doshi-Velez, F., and B. Kim. 2017. Towards a rigorous science of interpretable machine learning. arXiv:1702.08608.

Durtschi, C., W. Hillison, and C. Pacini. 2004. The effective use of Benford's law to assist in detecting fraud in accounting data. Journal of Forensic Accounting 5 (1): 17–34.

Glikson, E., and A. W. Woolley. 2020. Human trust in artificial intelligence: review of empirical research. Academy of Management Annals 14 (2): 627–660.

International Auditing and Assurance Standards Board (IAASB). 2021. International Standards on Auditing: ISA 200, 220, 230, 240, 315, 320, 330, 500, 505, 520, 530, 550, 570. New York, NY: IFAC.

International Auditing and Assurance Standards Board (IAASB). 2024. Discussion Paper: The Use of Automated Tools and Techniques, Including Artificial Intelligence, in an Audit. New York, NY: IFAC.

International Organization of Supreme Audit Institutions (INTOSAI). 2023. Guidance on the Use of Artificial Intelligence in Government Audit. Vienna: INTOSAI.

Jensen, M. C., and W. H. Meckling. 1976. Theory of the firm: managerial behavior, agency costs and ownership structure. Journal of Financial Economics 3 (4): 305–360.

Knechel, W. R. 2013. Do auditing standards matter? Current Issues in Auditing 7 (2): A1–A16.

Kokina, J., and T. H. Davenport. 2017. The emergence of artificial intelligence: how automation is changing auditing. Journal of Emerging Technologies in Accounting 14 (1): 115–122.

Kokina, J., S. Blanchette, T. H. Davenport, and D. Pachamanova. 2025. Challenges and opportunities for artificial intelligence in auditing: evidence from the field. International Journal of Accounting Information Systems 56: 100734.

Liu, F. T., K. M. Ting, and Z.-H. Zhou. 2008. Isolation forest. Proceedings of the 8th IEEE International Conference on Data Mining, 413–422.

Lundberg, S. M., and S.-I. Lee. 2017. A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems 30: 4765–4774.

National Audit Office of Mongolia (NAO). 2024. Audit Report SNAG-2024/10: Thermal Power Plant IV, FY2023. Ulaanbaatar: NAO.

National Audit Office of Mongolia (NAO). 2025. Audit Report SNAG-2025/092: Thermal Power Plant IV, FY2024. Ulaanbaatar: NAO.

Nigrini, M. J. 1996. A taxpayer compliance application of Benford's law. Journal of the American Taxation Association 18 (1): 72–91.

Nigrini, M. J. 2012. Benford's Law: Applications for Forensic Accounting, Auditing, and Fraud Detection. Hoboken, NJ: Wiley.

Obermeyer, Z., B. Powers, C. Vogeli, and S. Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. Science 366 (6464): 447–453.

Park, J. S., J. C. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein. 2023. Generative agents: interactive simulacra of human behavior. Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, 1–22.

Romao, M., A. Mendes, and J. Lima. 2026. Public-sector audit institutions and governance quality: a 189-country cluster analysis. International Journal of Public Sector Management (in press).

Tsetsegjargal, U., and G. Otgonsuren. 2025. Application of machine learning in auditing and credit risk assessment: evidence from Mongolia. International Journal of Innovative Science and Research Technology 10 (10): 3340–3348.

Tsetsegjargal, U., P. Oyunbileg, and G. Otgonsuren. 2025. Artificial intelligence in auditing reform: a systematic literature review and implementation conditions in Mongolia. Internauka Journal 18 (382): 29–37.

Vasarhelyi, M. A., A. Kogan, and B. M. Tuttle. 2015. Big data in accounting: an overview. Accounting Horizons 29 (2): 381–396.

Wang, L., C. Ma, X. Feng, Z. Zhang, H. Yang, J. Zhang, Z. Chen, J. Tang, X. Chen, Y. Lin, W. Zhao, Z. Wei, and J. Qin. 2024. A survey on large language model based autonomous agents. Frontiers of Computer Science 18 (6): 186345.

Yoon, K. 2020. A model for explainable AI in audit risk assessment. International Journal of Accounting Information Systems 39: 100479.