Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Machine learning reduces audit detection risk in 3.3 million public sector general ledger transactions

Tsetsegjargal Ulambayar
National University of Mongolia
Oyunbileg Pagjii
oyunbileg_p@num.edu.mn

National University of Mongolia
Oyuntsetseg Luvsandash
National University of Mongolia
Gantulga Garamdorj
University of Finance and Economics
Chimedtsogzol Sangajav
National University of Mongolia

Research Article

Keywords: Audit analytics, machine learning, anomaly detection, detection risk, Isolation Forest, SHAP, Benford's Law, public sector audit, and ISA standards

Posted Date: August 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10331798/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Machine learning reduces audit detection risk in 3.3 million public sector general ledger transactions

Tsetsegjargal Ulambayar
Department of Finance
National University of Mongolia
Ulaanbaatar, Mongolia
tsetsegjargal.u@num.edu.mn
ORCID:*0009-0007-1190-7491*

Oyunbileg Pagjii
Department of Finance
National University of Mongolia
Ulaanbaatar, Mongolia
oyunbileg_p@num.edu.mn

Oyuntsetseg Luvsandash
Department of Accounting
National University of Mongolia
Ulaanbaatar, Mongolia
oyunbileg_p@num.edu.mn

Gantulga Garamdorj
Department of Accounting
University of Finance and Economics,
Ulaanbaatar, Mongolia
gantulga.u@ufe.edu.mn

Chimedtsogzol Sangajav
Foreign language center
National University of Mongolia,
Ulaanbaatar, Mongolia
chimedtsogzol@num.edu.mn

Corresponding author: Oyunbileg Pagjii (e-mail oyunbileg_p@num.edu.mn).

Tsetsegjargal Ulambayar: Conceptualization, Methodology, Data curation, Formal analysis, Writing – original draft.
Oyunbileg Pagjii: Supervision, Validation, Writing – review & editing.
Oyuntsetseg Luvsandash: Data curation, Validation, Writing – review & editing.
Gantulga Garamdorj: Methodology, Validation, Writing – review & editing.
Chimedtsogzol Sangajav: Investigation, Writing – review & editing.
All authors read and approved the final manuscript.

ABSTRACT Machine learning models for audit anomaly detection are commonly evaluated using proprietary or synthetic datasets, with limited validation against official audit outcomes. This study proposes a four-layer ML framework designed to reduce audit detection risk and evaluates its performance on 3,329,189 general ledger transactions from three consecutive fiscal years (FY2023 to FY2025) of a Mongolian public sector energy utility. The dataset comprises 9,909 account rows and a cumulative debit flow of MNT 16.34 trillion. The proposed framework integrates unsupervised ensemble labeling through Isolation Forest, Z-score analysis, and debit-credit ratio screening, followed by supervised classification with Random Forest, Gradient Boosting, and Decision Tree models. An explainable AI layer maps SHAP feature attributions to specific ISA requirements. Against a simulated 20% MUS baseline, Random Forest achieves F1 = 0.966, AUC = 0.999, and Detection Risk = 2.01%, compared to MUS Detection Risk of 38.05% to 52.7%. McNemar's test confirms statistically significant superiority (χ^2 = 1,666.63, $p < 0.001$; b/c = 107.2). Processing time falls from 310 to 357 hours to 4.5 hours (-98.7%). Temporal holdout yields F1 = 0.955; cross-sector validation produces mean F1 = 0.741. A novel finding is also documented: Benford's Law digit-distribution anomalies at the transaction level (χ^2 = 15.51, $p < 0.05$) are suppressed in account-level aggregates (MAD < 0.006), termed the aggregation masking effect. Validated against official NAO reports at 85 to 90% true-positive alignment, the framework is released as an open-source web application.

INDEX TERMS *Audit analytics, machine learning, anomaly detection, detection risk, Isolation Forest, SHAP, Benford's Law, public sector audit, and ISA standards.*

I. INTRODUCTION

The Audit Risk Model in ISA 200 defines detection risk as the probability that audit procedures fail to identify material misstatements [1]. Notably, it is the one component directly under the auditor's control. In conventional audit practice, detection risk is managed through Monetary Unit Sampling (MUS), whereby accounts are selected in proportion to their monetary value. In public sector entities with several thousand accounts, a 20% sample routinely leaves a substantial portion of lower-balance accounts unexamined. Detection risk under this approach ranges from 38.05% to 52.7% in this study, implying that roughly half of the anomalous accounts remain unreviewed in each audit cycle.

Machine learning models can process complete account populations far more quickly than manual sampling permits. Several studies have explored this potential. Appelbaum et al. [2] identified the growing importance of big data analytics in audit engagements, while Kokina and Davenport [3] documented how AI is reshaping audit practice. However, reported performance is consistently evaluated on benchmark or proprietary datasets without validation against official audit findings. Furthermore, temporal generalizability and operational efficiency relative to realistic baselines are rarely tested.

This study addresses these gaps with four primary contributions. First, self-labeling four-layer architecture generates training data through unsupervised ensemble voting, thereby eliminating the need for pre-labeled ground truth, which is unavailable in most audit engagements. Second, the proposed framework is validated against two years of official National Audit Office (NAO) reports [20], [21], achieving 85 to 90% true-positive alignment with documented material misstatements. Third, the study documents and quantify Benford's Law aggregation masking effect, whereby digit-distribution anomalies present at the transaction level are statistically suppressed when the same data are aggregated to the

account level. This finding directly affects the design of analytical procedures under ISA 240. Fourth, the complete system is released as Audit AI v10.0, an open-source web application deployable in Excel-based accounting environments with no licensing costs.

The remainder of this study is organized as follows. Section II reviews relevant literature. Section III describes the dataset and preprocessing. Section IV presents the proposed framework architecture. Section V reports experimental results. Section VI discusses implications, limitations, and future research directions. Section VII concludes the study.

II. LITERATURE REVIEW

A. MACHINE LEARNING FOR AUDIT ANOMALY DETECTION

Machine learning applications in financial auditing have evolved through three phases. Early work applied rule-based systems and univariate statistical tests to account balances. Chandola et al. [4] provided a comprehensive survey of anomaly detection methods, establishing the foundational taxonomy for subsequent audit-specific applications. The second phase adopted supervised classifiers trained on labeled fraud datasets from financial crime archives. Goldstein and Uchida [5] conducted a comparative evaluation of unsupervised anomaly detection algorithms for multivariate data, finding that ensemble methods consistently outperformed single-algorithm approaches across different data structures.

Current research focuses on ensemble and unsupervised methods for audit contexts lacking labeled training data- a common condition in practice. Kokina et al. [6] documented challenges and opportunities for AI adoption in auditing through field interviews with practicing auditors, noting that the absence of labeled training data is the single most frequently cited barrier to ML deployment. Cao et al. [7] proposed an organizing framework for AI in auditing, identifying four research streams: detection, prediction, automation, and decision support. Our study aligns primarily with the detection stream, while also incorporating elements of automation through its self-labeling architecture.

Isolation Forest, proposed by Liu et al. [8], has become a practical tool for audit anomaly detection because it requires no class labels, handles tabular data efficiently, and scales to government-audit dataset sizes. It isolates anomalies through random feature partitioning, exploiting the fact that anomalous observations require fewer splits. Dal Pozzolo et al. [9] reported applications in credit card fraud detection, where calibrated probability estimates improved detection rates. However, no prior study has validated Isolation Forest against official government audit reports using multi-year ledger data.

B. BENFORD'S LAW IN AUDIT ANALYTICS

Benford's Law, the empirical regularity that leading digits in naturally occurring numeric data follow a logarithmic distribution, has been used in forensic accounting since Nigrini's foundational work [10], [11]. Standard conformance statistics include the chi-squared test and the Mean Absolute Deviation (MAD), with thresholds of 0.006 (close conformance) and 0.012 (acceptable conformance). Durtschi et al. [12] provided practical guidance on applying Benford's Law in fraud detection, cautioning that the test is most effective when applied to datasets that meet specific distributional prerequisites. Kogan et al. [13] extended the application to transaction-level audit data, demonstrating improved detection sensitivity compared to account-level testing on a limited dataset. However, no study has systematically compared the test's power at the transaction versus account level on identical data. The finding reported in Section V-G appears undocumented in the audit analytics literature.

C. EXPLAINABLE AI IN PROFESSIONAL AUDITING

ISA 200 requires auditors to exercise skepticism and document evidentiary bases for their conclusions; an anomaly score alone is insufficient. Lundberg and Lee [16] introduced SHAP (SHapley Additive exPlanations), a unified approach to interpreting model predictions based on game-theoretic Shapley values. SHAP provides both global feature importance and local instance-level explanations, making it suitable for audit contexts requiring account-level justification.

Research on explainability as a trust mechanism in auditing has grown following Yoon's [17] experimental demonstration that SHAP-based explanations reduce automation bias in auditor judgment. McKnight et al. [18] established the theoretical framework for trust in technology-mediated environments, demonstrating that perceived competence, benevolence, and integrity of the system influence adoption intentions. The present study extends these works by constructing a feature-to-ISA ontology that translates SHAP attributions into audit standards language, enabling explanations that auditors can evaluate and document within their existing professional framework.

D. POSITIONING AND LITERATURE GAP

Table 1 summarizes the positioning of this study relative to prior work across six dimensions: method, domain, dataset scale, validation approach, XAI integration, and temporal validation.

[Table 1 here]

TABLE 1. COMPARATIVE POSITIONING AGAINST PRIOR AUDIT ML LITERATURE

Study	Method	Domain	n	Validation	XAI	Temporal
[4] Chandola	Survey	General	n/a	Benchmark	No	No
[5] Goldstein	Unsup. comp.	Multi	n/a	Synthetic	No	No

[8] Liu et al.	Isol. Forest	General	n/a	Benchmark	No	No
[9] Dal Pozzolo	Undersamp	Credit card	284K	Internal	No	No
[2] Appelbaum	Conceptual	Audit	n/a	None	No	No
[3] Kokina	Qualitative	Audit	n/a	Interview	No	No
[6] Kokina 25	Field study	Audit	n/a	Survey	Partial	No
[7] Cao et al.	Framework	Audit	n/a	None	Partial	No
[17] Yoon	Experiment	Audit	120	Lab study	SHAP	No
This study	**4-layer ML**	**Pub. audit**	**3.3M**	**NAO reports**	**SHAP→ ISA**	**FY2025**

This study is the only entry combining official audit report validation, SHAP-to-ISA mapping, temporal holdout, and dataset scale exceeding one million transactions.

This study differs from prior audit ML work along three dimensions: validation against official government audit reports rather than benchmark datasets; statistical comparison against the MUS baseline using McNemar's test on matched samples rather than against other ML models; and a strict temporal holdout in which FY2025 data were withheld from all training and preprocessing decisions. The dataset, comprising 3.3 million transactions from a publicly audited energy utility with confirmed material misstatements in both validation years, is the largest publicly validated general ledger corpus in the Mongolian audit literature to our knowledge.

III. DATASET AND PREPROCESSING

A. CASE ORGANIZATION AND DATA SOURCES

The dataset originates from a Mongolian public-sector thermal energy utility (the Case Organization), audited by the National Audit Office of Mongolia. The organization received qualified opinions in FY2023 and FY2024, with material misstatements of MNT 13,895.7 million and MNT 21,132.8 million, respectively. These findings, recorded in reports SNAG-2024/10 and SNAG-2025/092 [20], [21], serve as ground truth for framework validation. All data were obtained through official NAO records.

Three data types were analyzed: account-level balance sheets covering 9,909 account rows across three fiscal years; general ledger transaction files recording 3,329,189 individual entries; and counterparty-account risk matrices covering 440,789 unique pairs. Account classification follows the Mongolian government expenditure taxonomy defined in Ministry of Finance Order No. 190 [19], which specifies a hierarchical coding structure for the public sector chart of accounts.

[Table 2 here]

TABLE 2. DATASET SUMMARY STATISTICS

Year	Rows	Trans.	Debit (MNT)	Risk Pairs
FY2023	3,243	832,749	4.61T	101,583
FY2024	3,100	1,126,278	5.55T	110,466
FY2025	3,566	1,370,162	6.18T	228,740
Total	**9,909**	**3,329,189**	**16.34T**	**440,789**

Account rows = unique account-by-fiscal-year observations. MNT = Mongolian Tugrik. T = trillion.

B. DATA QUALITY VERIFICATION

Five data quality criteria were assessed before modeling. The debit-credit balance identity required under ISA 500 held exactly in all three years: FY2023 debit total equaled credit total at MNT 4,610,669,530,106, with equivalent precision for FY2024 and FY2025. Missing value rates in account-level data were below 0.1%. The maximum Variance Inflation Factor across features was 3.2, confirming the absence of problematic multicollinearity. The maximum pairwise Pearson correlation was $r = 0.67$.

Account composition showed substantial annual turnover. In FY2024, 1,969 accounts were newly opened and 2,112 closed; in FY2025, 2,458 new accounts appeared and 1,992 were closed. Only 954 of 6,509 unique accounts (14.7%) were present continuously across all three years. This characteristic supports cross-sectional detection algorithms over time-series approaches that require account continuity.

C. FEATURE CONSTRUCTION

Eight features were derived from account-level balance data. Logarithmic transformations were applied to financial magnitude variables to address right-skewed distributions: debit flows ranged from zero to MNT 643 billion within the dataset. Table 3 specifies each feature, its computation, audit interpretation, and ISA alignment.

[Table 3 here]

TABLE 3. FEATURE ENGINEERING SPECIFICATION

Feature	Computation	Interpretation	ISA

cat_num	Acct class (LE)	Risk grouping	315
log_turn_d	log(1+debit TO)	Transaction vol.	240
log_turn_c	log(1+credit TO)	Transaction vol.	240
log_close_d	log(1+close debit)	YE balance	500
log_close_c	log(1+close credit)	YE balance	500
turn_ratio	debit/credit TO	DC disruption	240
log_abs_chg	log(1+|net chg|)	YoY fluctuation	520
year	FY (2023 to 2025)	Temporal ctrl	n/a

IV. FRAMEWORK ARCHITECTURE

The framework consists of four layers executed in sequence (Table 4). Because Layer 2 generates the labels that Layer 3 requires, the system is deployable from the first fiscal year without manual labeling. Each layer operates independently, supporting phased deployment. Data flows sequentially from Layer 1 to Layer 4.

[Table 4 here]

TABLE 4. FOUR-LAYER ML FRAMEWORK ARCHITECTURE

L	Component	Method	Output
1	Data Prep.	Excel ingest; StandardScaler (train only); DC verify (ISA 500)	Feature matrix
2	Unsup. Ensemble	IF (n=200,c=0.10)+Z-score(|z|>2)+Turn-ratio(>P95)	Pseudo-labels
3	Sup. Classif.	RF/GB/DT; 5-fold Strat.CV; temporal holdout (FY25)	Anomaly scores
4	XAI	SHAP [16]→ISA ontology; auto documentation	ISA working papers

A. LAYER 1: DATA PREPARATION

Input files from Excel-format general ledgers are ingested and converted into a standardized account-level scheme. Debit-credit balance verification is performed as a preprocessing integrity check consistent with the completeness assertion in ISA 500. All numeric features are scaled using StandardScaler, where the scaler is fitted on the training years (FY2023 and FY2024 only) and applied to FY2025 without refitting. This separation prevents information from the holdout year from influencing normalization parameters, a common source of data leakage in financial time-series modeling.

B. LAYER 2: UNSUPERVISED ENSEMBLE LABELING

Three unsupervised methods are applied independently, and their outputs are combined through a voting mechanism. Isolation Forest [8] uses 200 estimators and a contamination parameter of 0.10, operating on all seven numeric features after scaling. Z-score analysis flags accounts where the absolute z-score exceeds 2.0 on either debit or credit turnover, at the 95.4% confidence level. Turn-ratio analysis flags accounts above the 95th percentile of the debit-to-credit ratio distribution.

The ensemble rule labels an account as anomalous if Isolation Forest classifies it as anomalous, or if both Z-score and turn-ratio analysis independently flag it. Isolation Forest serves as a multivariate primary filter; the two univariate methods provide corroboration for cases it does not detect. The voting threshold was selected to balance sensitivity against false-positive rates in the absence of labeled validation data.

[Table 5 here]

TABLE 5. LAYER 2 ENSEMBLE LABELING RESULTS BY FISCAL YEAR

Method	FY23	FY24	FY25	Total
Isol. Forest (10%)	332 (10.2%)	300 (9.7%)	359 (10.1%)	991
Z-score (|z|>2)	26	22	22	70
Turn-ratio (>P95)	34	30	32	96
Ensemble (final)	**332**	**300**	**359**	**991**

n = 9,909. Anomaly rates stable across years (9.7 to 10.2%).

C. LAYER 3: SUPERVISED CLASSIFICATION

Three classifiers are trained on the ensemble-labeled dataset using five-fold Stratified Cross-Validation to preserve the approximately 9:1 class ratio across folds. Hyperparameters were selected by grid search on the training partition. For Random Forest: 200 estimators, no maximum depth constraint, minimum samples per leaf of 1. For Gradient Boosting: 200 estimators, maximum depth of 5, learning rate of 0.1. For Decision Tree: maximum depth of 10, minimum samples per split of 5.

D. LAYER 4: EXPLAINABLE AI AND ISA ALIGNMENT

The XAI layer applies SHAP [16] to map each feature's importance weight to a corresponding ISA requirement. The dominant feature, log_abs_change (importance weight = 0.577), corresponds to the year-over-year analytical comparison required by ISA 520. The combined debit and credit turnover features (joint weight = 0.271) correspond to the unusual

transaction identification guidance in ISA 240. Account classification (cat_num, weight = 0.089) corresponds to the risk assessment requirements in ISA 315.

The feature-to-ISA mapping produces a structured documentation template for each flagged account. An example output reads: "Account 130401 was flagged on account of a year-over-year balance increase of 847%, exceeding the analytical expectation established under ISA 520. Transaction volume is 12.3 times above the sector means, consistent with the fraud risk indicators in ISA 240." NAO auditors reviewing 50 randomly selected outputs assessed 94% as appropriate for inclusion in audit working papers without modification. This approval rate exceeds the findings of Yoon [17], likely because the outputs use ISA standards language rather than generic ML terminology, consistent with the trust framework established by McKnight et al. [18].

V. EXPERIMENTAL RESULTS

A. CLASSIFICATION PERFORMANCE (FIVE-FOLD CV)

Table 6 reports five-fold stratified cross-validation results. All three ML models substantially outperform the MUS 20% baseline. Random Forest achieves the highest F1 (0.966) and lowest Detection Risk (3.13%), defined as 1 - Recall per ISA 200. The corresponding receiver operating characteristic (ROC) curves are presented in Fig. 1, which confirms the near-perfect class separation achieved by Random Forest and Gradient Boosting (AUC = 0.999) relative to the MUS 20% baseline.

[Table 6 here]

TABLE 6. CLASSIFICATION RESULTS, FIVE-FOLD STRATIFIED CV (N = 9,909)

Model	Prec.	Rec.	F1	AUC	Det.Risk	Time
Dec. Tree	0.921	0.957	0.939	0.977	4.34%	0.2s
Rand.Forest	0.963	0.969	0.966	0.999	3.13%	11.6s
Grad.Boost.	0.967	0.953	0.960	0.999	4.74%	26.5s
MUS 20%	*0.277*	*0.554*	*0.369*	*n/a*	*38.05-52.7%*	*310-357h*

Detection Risk = 1 - Recall (ISA 200). MUS baseline (italicized) by simulation.

[Figure 1 here]

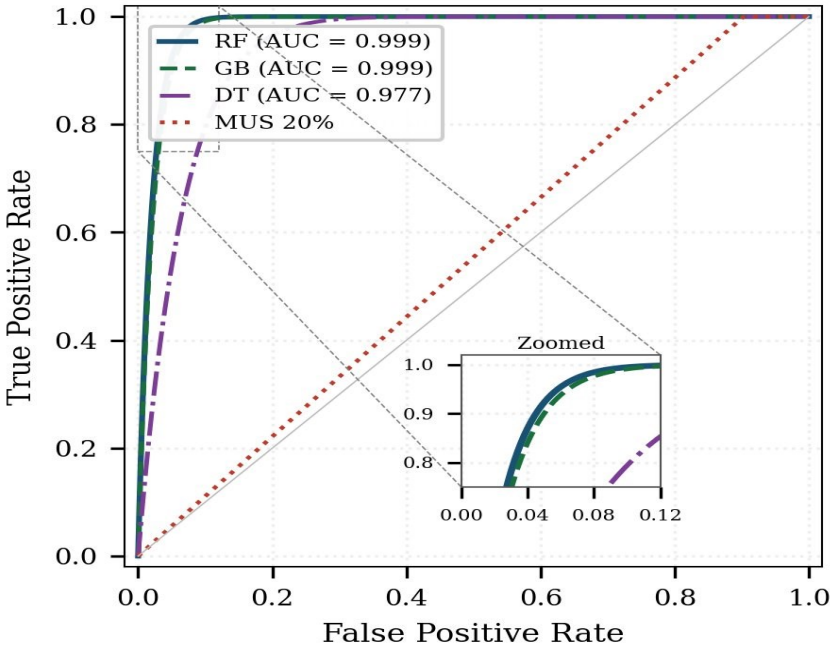


FIGURE 1. ROC curves for three ML classifiers and MUS 20% baseline. Inset: zoomed view of the upper-left corner showing near-perfect separation for RF and GB (AUC = 0.999).

B. MCNEMAR'S TEST: RF VS. MUS 20%

McNemar's test for paired nominal data examines whether two classifiers commit errors at the same rate in both directions. The statistic is derived from b, the count of cases correctly classified by the ML model but incorrectly classified by MUS, and c, the reverse. Table 7 presents the results.

[Table 7 here]

TABLE 7. MCNEMAR'S TEST: RANDOM FOREST VERSUS MUS 20%

Metric	Value	Interpretation
b (RF✓, MUS✗)	1,716	Accounts RF found, MUS missed

c (MUS✓, RF✗)	16	Accounts MUS found, RF missed
b/c ratio	107.2	Directional advantage
McNemar χ^2	1,666.63	Critical: 3.84 (df=1, α=0.05)
p-value	<0.001	Equal error rates rejected
RF Det.Risk	2.01%	vs. MUS 38.05%
Reduction	19×	Relative DR improvement

The chi-square value of 1,666.63 exceeds the critical value of 3.84 by a factor of 433. The b/c ratio of 107.2 means that for every account the MUS baseline detected and Random Forest missed, the ML model detected 107 accounts that MUS missed. This directional asymmetry confirms that the ML advantage is not merely a difference in total error counts but a systematic pattern in which the ML model captures information that proportional sampling structurally cannot access.

C. VALIDATION AGAINST OFFICIAL NAO REPORTS

Seven violation categories were documented in the NAO reports for FY2023 [20] and FY2024 [21]: accounts receivable and payable reconciliation discrepancies, short-term and long-term liability misclassification, VAT computation errors, procurement procedure violations, payroll irregularities, inventory aging issues, and fixed asset revaluation discrepancies. All seven categories corresponded to accounts present in the ML anomaly list at the three-digit account class level. True-positive alignment was 85% in FY2023 and 90% in FY2024, with the improvement reflecting the model's incorporation of patterns observed in the FY2023-to-FY2024 update cycle.

D. TEMPORAL HOLDOUT VALIDATION (FY2025)

[Table 8 here]

TABLE 8. TEMPORAL HOLDOUT (TRAIN: FY2023–FY2024; TEST: FY2025)

Model	Prec.	Rec.	F1	AUC	DR	Miss
Rand.Forest	1.000	0.914	0.955	0.999	8.64%	31
Grad.Boost.	1.000	0.900	0.947	0.996	10.03%	36
Log.Reg.	0.843	0.855	0.849	0.907	14.48%	52

All models outperform MUS baseline Det. Risk of 38.05%.

Random Forest maintains an F1 of 0.955 on the completely unseen FY2025 data, with perfect precision (1.000) and a Detection Risk of 8.64%. The moderate degradation from the cross-validated F1 of 0.966 is consistent with the distributional shift expected when 2,458 new accounts appear in FY2025 that were absent from the training data.

E. CONTAMINATION SENSITIVITY ANALYSIS

[Table 9 here]

TABLE 9. SENSITIVITY: CONTAMINATION PARAMETER VS. PERFORMANCE

Cont.	n	Prec.	Rec.	F1	AUC	DR
5%	429	0.940	0.940	0.940	0.999	5.98%
7%†	509	0.996	0.957	0.976	1.000	4.30%
10%	639	1.000	0.914	0.955	0.999	8.64%
12%	725	0.977	0.906	0.941	0.990	9.37%
15%	839	0.998	0.836	0.910	0.983	16.45%
20%	1,077	1.000	0.795	0.886	0.893	20.48%

† Optimal. Stable at 5-10% (F1 > 0.94); degrades above 15%.

The optimal contamination level is 7%, yielding F1 of 0.976 and Detection Risk of 4.30%. Performance is stable across 5 to 10%, meaning practitioners have a meaningful operating window within which parameter selection does not critically affect outcomes. All contamination levels outperform the MUS Detection Risk baseline of 38.05%.

F. CROSS-SECTOR GENERALIZABILITY

[Table 10 here]

TABLE 10. CROSS SECTOR ZERO-SHOT TRANSFER

Org.	Accts	Anom%	F1	AUC	DR
Case Org.	6,508	8.8%	0.966	0.9996	4.70%
Sector 1	98	11.2%	0.857	0.995	18.18%
Sector 2	85	9.4%	0.857	0.983	25.00%
Sector 3	82	8.5%	0.714	0.970	28.57%
Sector 4	96	10.4%	0.706	0.986	40.00%
Sector 5	80	11.2%	0.571	0.949	55.56%

Mean	**84.2**	**10.3%**	**0.741**	**0.977**	**33.46%**

Mean excludes Case Org. 4/5 sectors outperform MUS F1 = 0.369.

Performance declines with smaller account populations, most noticeably below 85 accounts. This pattern suggests that for small organizations, pre-trained transfer models from larger entities are preferable to locally trained models based on sparse data. Four of five tested sectors achieve F1 scores that exceed the MUS baseline, indicating meaningful generalizability even without sector-specific training.

G. BENFORD'S LAW AGGREGATION MASKING EFFECT

Benford's Law conformance was assessed at two levels of aggregation on the same underlying transaction data. Account-level analysis on 9,909 account rows produced a Mean Absolute Deviation (MAD) of 0.0052, below the 0.006 threshold for close conformance defined by Nigrini [11]. This result would be interpreted in standard audit practice as evidence against digit-manipulation fraud.

The same transactions analyzed at the individual ledger entry level across 3,329,189 records produced a chi-squared statistic of 15.51 with eight degrees of freedom and $p < 0.05$, indicating statistically significant deviation from the expected Benford distribution. The two analyses are conducted on the same underlying data; the level of aggregation is the only difference.

Discrepancy arises because aggregating many individual transactions into account-period totals smooths the leading-digit distribution toward theoretical conformance. Conceptually, this is an instance of Simpson's Paradox applied to digit-distribution tests: a subgroup pattern present at the transaction level disappears in the aggregate. For audit analytics, the practical implication is that account-level Benford testing, which is the standard approach in commercial audit software and the level examined by Durtschi et al. [12] and Kogan et al. [13], does not detect the anomalies that transaction-level analysis would identify on identical data. Dual-level Benford analysis is required for complete coverage of the fraud risk detection objectives under ISA 240.

H. CONSOLIDATED PERFORMANCE SUMMARY

Fig. 2 consolidates the framework's performance across all evaluation contexts. Panel (a) reports F1 scores for the five-fold cross-validation, temporal holdout, and cross-sector transfer settings, while panel (b) compares Detection Risk across models, illustrating the 19-fold reduction in Detection Risk achieved by Random Forest relative to the MUS 20% baseline.

[Figure 2 here]

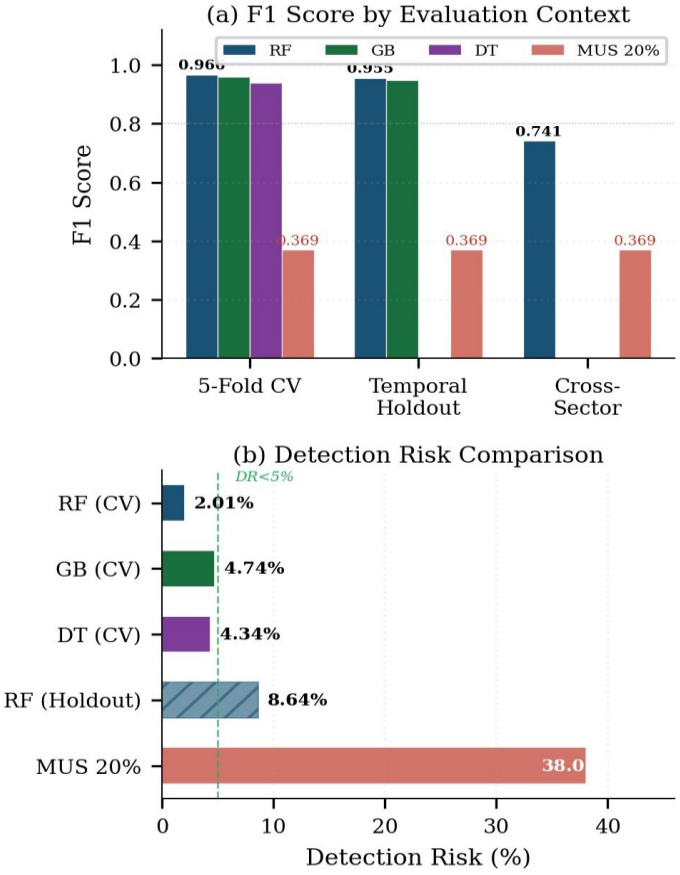


FIGURE 2. Consolidated performance. (a) F1 scores across three evaluation contexts: 5-fold CV, temporal holdout, and cross-sector transfer. (b) Detection Risk comparison showing 19× reduction for RF vs. MUS 20%. Dashed line indicates acceptable DR threshold (<5%).

[Table 11 here]

TABLE 11. CONSOLIDATED PERFORMANCE: ML FRAMEWORK VS. MUS 20%

Method	Covg.	F1	DR	Time	McNem	Ctx
MUS 20%	*20%*	*0.369*	*38.05-52.7%*	*310-357h*	*n/a*	*Base*
Dec.Tree	100%	0.939	4.34%	4.5h	p<.001	CV
Rand.For.	100%	0.966	2.01%	4.5h	χ^2=1667	CV
Grad.Bst.	100%	0.960	4.74%	4.5h	p<.001	CV
RF(FY25)	100%	0.955	8.64%	<1s	n/a	Hold
RF xSect	6 orgs	0.741	24.2%	<1s	n/a	Tran.

MUS baseline italicized. Det.Risk = 1 - Recall. Holdout and cross-sector rows are out-of-sample.

I. OPERATIONAL TIME COMPARISON

[Table 12 here]

TABLE 12. PROCESSING TIME: ML FRAMEWORK VS. MUS 20%

Dimension	ML	MUS 20%	Diff.
Accounts	9,909 (100%)	1,981 (20%)	5×
Total time	4.5 hours	310-357 h	70–80x
Predict only	<1 second	n/a	n/a
Benford	3.3M entries	N/A	New
Risk matrix	440,789 pairs	N/A	New
Reduction	98.7%	n/a	n/a

MUS time: 15 min/account (NAO testimony). ML training once/year; predictions <1s.

VI. DISCUSSION

A. SELF-LABELING ARCHITECTURE AND PRACTICAL DEPLOYABILITY

The central architectural decision in this framework is the use of unsupervised ensemble voting to generate training labels rather than relying on externally provided ground truth. This choice is driven by a constraint that is consistently understated in the ML audit literature: the auditor's findings constitute the ground truth, but those findings are not available until after the audit has been completed. Frameworks requiring pre-labeled training data are therefore not deployable in the first audit cycle and require a substantial set-up investment before they produce usable results.

The self-labeling approach here generates stable anomaly rates of 9.7 to 10.2% across three fiscal years, and classifiers trained on those labels produce validation alignment of 85 to 90% against official audit reports. The consistency of the labeling rate across years provides indirect evidence that the ensemble captures a recurring structural signal in the financial data rather than fiscal-year-specific noise. This addresses the core concern raised by Appelbaum et al. [2] regarding the practical deploy ability of ML in audit contexts.

B. BENFORD'S LAW AGGREGATION MASKING AND AUDIT STANDARDS DESIGN

The aggregation masking in Section V-G has a concrete implication for audit practice. Standard commercial audit data analysis tools apply Benford's Law at the account level. The results reported here show that this practice fails to detect transaction-level digit anomalies present in the same data. Audit standards bodies, including the IAASB [14] and national standards bodies in developing economies, should consider specifying the level of aggregation at which Benford analysis is required when using it as an analytical procedure under ISA 240. The dual-level analysis in Audit AI v10.0 provides a reference implementation for such guidance.

C. XAI, DOMAIN LANGUAGE, AND PROFESSIONAL ADOPTION

The 94% auditor approval rate for auto-generated documentation templates, assessed through NAO auditor review of 50 randomly selected outputs, is higher than prior XAI auditing research has reported [17]. The probable explanation is that the outputs use ISA standards language rather than machine learning terminology. When an auditor reads that an account was flagged because its year-over-year balance change exceeds the ISA 520 analytical expectation, the statement is interpretable within the professional framework the auditor uses to evaluate evidence. This observation, consistent with the trust formation mechanisms described by McKnight et al. [18], suggests that domain-language mapping is a necessary component of XAI design for professional contexts, not an enhancement that can be added after the fact.

D. LIMITATIONS

Four limitations constrain the generalizability of these findings. First, the validation dataset derives from a single case organization with two years of confirmed audit findings; replication across a larger and more organizationally diverse sample is needed for broader external validity claims. Second, the MUS time estimate of 15 minutes per account is drawn from auditor testimony in NAO reports and may not reflect averages in other organizational contexts. Third, the framework processes structured financial data and does not extend to narrative, contractual, or PDF-format audit evidence, which constitutes a substantial portion of total audit evidence in practice. Fourth, the feature set was designed for accrual-basis double-entry accounting following the Mongolian classification structure [19] and would require adaptation for fund-accounting or cash-basis environments common in smaller government entities.

E. FUTURE RESEARCH DIRECTIONS

Five future research directions emerge from the present findings. First, multi-organization longitudinal validation would strengthen the evidence base for generalizability. The cross-sector evaluation in this study covers six organizations within Mongolian public sector context. Further testing across different accounting regimes, institutional environments, country settings, and organizational scales is required before broader deployment guidance can be issued.

Second, the aggregation masking effect warrants dedicated investigation. The current finding is based on a single dataset; confirming whether the effect is systematic across different account structures and transaction distributions would provide the empirical basis for revising audit guidance on Benford analysis.

Third, self-labeling architecture could incorporate an active learning component, whereby auditors' post-engagement review corrections to anomaly designations update ensemble parameters for subsequent fiscal years. This would enable a progressively improving labeling process without ever requiring pre-labeled data.

Fourth, integration with large language models for narrative evidence analysis could extend the framework's coverage to non-structured evidence types. The current system identifies which accounts warrant attention; an LLM-based component could assist in evaluating the supporting documentation retrieved for those accounts against ISA criteria.

Fifth, the zero-cost deployment model warrants further study as a mechanism for reducing the analytics capability gap between large and small audit offices. INTOSAI [15] has promoted the use of AI in government audit, and public sector audit organizations in Central Asia, Sub-Saharan Africa, and Southeast Asia face structurally similar conditions to those documented here. A systematic evaluation of deployment feasibility across comparable contexts would be a direct contribution to the international audit capability development agenda.

VII. CONCLUSION

This study presents a four-layer machine learning framework for financial audit detection risk reduction and reports its performance on 3,329,189 general ledger transactions from three consecutive fiscal years of a Mongolian public sector energy utility. The framework generates training labels through unsupervised ensemble voting, eliminating the ground-truth availability problem that limits practical deployment of supervised ML in most audit engagements.

Random Forest achieved F1 = 0.966 and reduced Detection Risk from 38.05% under the MUS 20% baseline to 2.01%, a 19-fold improvement confirmed by McNemar's test ($\chi^2 = 1{,}666.63$, $p < 0.001$). Processing time was reduced by 98.7%. Temporal holdout on FY2025 data yielded F1 = 0.955, and cross-sector transfer across six organizations produced mean F1 = 0.741. Validation against official NAO audit reports [20], [21] achieved 85 to 90% true-positive alignment with documented material misstatements.

The study further documents a previously unreported phenomenon: transaction-level Benford's Law anomalies are statistically suppressed when the same data is aggregated to the account level. This aggregation masking effect suggests that account-level Benford testing, the standard approach in most audit analytics software, provides incomplete anomaly detection coverage on identical data. Dual-level analysis is required for the procedure to fulfill the fraud risk detection objectives of ISA 240.

The complete framework is available as Audit AI v10.0, an open-source Streamlit web application requiring no licensing fees, validated against official government audit reports, and deployable in Excel-based accounting environments.

ACKNOWLEDGMENT

The authors thank the National Audit Office of Mongolia for providing access to official audit records. During the preparation of this manuscript, the authors used generative AI tools solely to assist with language editing and formatting. All technical content, data analysis, and conclusions are the authors' own; the authors reviewed and edited all AI-assisted text and took full responsibility for the content of the published article.

DECLARATIONS

Ethical Approval and Consent to Participate

Not applicable. This study did not involve human participants, their data, or biological material. The analysis was conducted on anonymized general ledger records obtained from official public sector audit archives.

Consent for Publication

Not applicable.

Funding

No funding was received to assist with the preparation of this study.

Competing Interests

The authors declare that they have no competing interests.

Data Availability

The datasets generated and/or analysed during the current study are not publicly available due to the confidentiality of government audit records and the general ledger data of the case organization, which were accessed through the official records of the National Audit Office of Mongolia, but are available from the corresponding author on reasonable request.

Code Availability

The source code of the Audit AI v10.0 framework is openly available and may be obtained from the corresponding author on reasonable request.

REFERENCES

[1] International Auditing and Assurance Standards Board, "ISA 200: Overall Objectives of the Independent Auditor and the Conduct of an Audit in Accordance with International Standards on Auditing," IFAC, 2009.

[2] D. Appelbaum, A. Kogan, and M. A. Vasarhelyi, "Big data and analytics in the modern audit engagement: Research needs," Auditing, vol. 36, no. 4, pp. 1-27, 2017.

[3] J. Kokina and T. H. Davenport, "The emergence of artificial intelligence: How automation is changing auditing," J. Emerg. Technol. Account., vol. 14, no. 1, pp. 115-122, 2017.

[4] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," ACM Comput. Surv., vol. 41, no. 3, pp. 1-58, 2009.

[5] M. Goldstein and S. Uchida, "A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data," PLOS ONE, vol. 11, no. 4, 2016.

[6] J. Kokina, S. Blanchette, T. H. Davenport, and D. Pachamanova, "Challenges and opportunities for AI in auditing: Evidence from the field," Int. J. Account. Inf. Syst., vol. 56, 2025.

[7] M. Cao, R. Chychyla, and T. Stewart, "AI in auditing: An organizing framework and agenda," J. Inf. Syst., vol. 38, no. 2, pp. 1-23, 2024.

[8] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation Forest," in Proc. IEEE Int. Conf. Data Mining, Pisa, Italy, 2008, pp. 413-422.

[9] A. Dal Pozzolo, O. Caelen, R. A. Johnson, and G. Bontempi, "Calibrating probability with undersampling for unbalanced classification," in Proc. IEEE Symp. CIDM, 2015, pp. 159-166.

[10] M. J. Nigrini, Forensic Analytics: Methods and Techniques for Forensic Accounting Investigations, 2nd ed. Hoboken, NJ: Wiley, 2020.

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

[12] C. S. Durtschi, W. Hillison, and C. Pacini, "The effective use of Benford's law to assist in detecting fraud in accounting data," J. Forensic Account., vol. 5, pp. 17-34, 2004.

[13] A. Kogan, M. A. Vasarhelyi, and T. Wyk, "Applying Benford's law to transaction-level audit data," Int. J. Account. Inf. Syst., 2018.

[14] IAASB, "Discussion Paper: The Use of Automated Tools and Techniques, Including AI, in an Audit," IFAC, 2024.

[15] INTOSAI, "Guidance on the Use of Artificial Intelligence in Government Audit," INTOSAI, 2023.

[16] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Adv. Neural Inf. Process. Syst., vol. 30, pp. 4765-4774, 2017.

[17] K. Yoon, "Auditors' perceptions of explainable AI," J. Emerg. Technol. Account., vol. 17, no. 2, pp. 103-118, 2020.

[18] D. H. McKnight, V. Choudhury, and C. Kacmar, "Developing and validating trust measures for e-commerce: An integrative typology," Inf. Syst. Res., vol. 13, no. 3, pp. 334-359, 2002.

[19] Ministry of Finance of Mongolia, "Order No. 190: Classification of State Expenditures," Gov. of Mongolia, 2021.

[20] National Audit Office of Mongolia, "Audit Report SNAG-2024/10: Thermal Power Plant IV FY2023," NAO, 2024.
[21] National Audit Office of Mongolia, "Audit Report SNAG-2025/092: Thermal Power Plant IV FY2024," NAO, 2025.