Preprints.org

Article

Not peer-reviewed version

SHAP-Guided Risk Path Generation for IRS Auditors: Interpretable Anomaly Attribution in Payroll Tax Compliance Screening

Tiantian Zhang *

Posted Date: 31 March 2026

doi: 10.20944/preprints202603.2462.v1

Keywords: CCS concepts; computing methodologies; machine learning; machine learning approaches; neural networks; payroll tax compliance; anomaly attribution; SHAP explanations; multi-task learning; risk path generation

Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 March 2026 doi:10.20944/preprints202603.2462.v1

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

Article

SHAP-Guided Risk Path Generation for IRS Auditors: Interpretable Anomaly Attribution in Payroll Tax Compliance Screening

Tiantian Zhang

Gies College of Business, University of Illinois Urbana-Champaign, Champaign, USA; tz46@illinois.edu

Abstract

Small and medium-sized enterprises are prone to errors or evasion in areas such as medical insurance reimbursement, pension contributions, and unemployment insurance, which not only harm workers' rights but also impact the sustainability of federal and state public funds. Therefore, an explainable intelligent tax compliance screening system is urgently needed. This paper uses multi-task learning as the backbone, while modeling the shared risk representations of sub-tasks such as wages, pensions, unemployment insurance, and medical insurance contributions. Subsequently, our model combines multi-task learning with time-series anomaly detection to capture overlooked scenarios and structural avoidance in long-term trends. On heterogeneous time-series graphs, SHapley Additive exPlanations (SHAP) contributions are propagated and decomposed according to 'feature-module-time-task,' and combined with gated dependencies and consistency constraints to generate auditable risk paths, achieving an upgrade from anomaly scoring to evidence-chain visualization. Experiments show that this method improves both identification performance and interpretable localization efficiency.

Keywords: CCS concepts; computing methodologies; machine learning; machine learning approaches; neural networks; payroll tax compliance; anomaly attribution; SHAP explanations; multi-task learning; risk path generation

1. Introduction

Payroll tax compliance screening sits at the intersection of labor protection and fiscal sustainability. For many small and medium-sized businesses, payroll reporting is not a single ledger problem but a coupled system. Gross wages, overtime and bonuses, withholding, employer-side contributions, benefit plan payments, and health insurance–related records must align across multiple reporting channels and over time 1. When misreporting, underpayment, or strategic avoidance occurs, the downstream impact is twofold. Workers may lose eligibility or continuity of benefits, and federal trust funds face cumulative leakage that undermines long-run solvency.

Traditional compliance workflows often rely on rule engines and threshold checks including abnormal wage growth, contribution ratios outside expected bounds, sudden changes in filing frequency, or mismatches between headcount and total payroll. While these heuristics are easy to explain, they are brittle against evolving evasion strategies and cross-form inconsistencies that only become apparent when multiple data streams are integrated 2. Conversely, end-to-end machine learning models can improve detection accuracy but frequently fail to translate into actionable audit decisions because their outputs are not naturally aligned with how auditors gather evidence.

A major difficulty in this domain lies in heterogeneous data and semantic misalignment across sources. Payroll records, benefits contributions, and health insurance–related information are generated by different operational processes and may use different identifiers, timing conventions, and aggregation levels. Distributions also vary by industry, geography, firm size, and employment composition, which can cause naïve models to confuse structural differences with noncompliance 3.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 March 2026 doi:10.20944/preprints202603.2462.v1

Moreover, reasonable explanations for changes seasonality, workforce turnover, shifts from full-time to part-time, changes in benefit enrollment can mimic the statistical footprint of evasion. The model must therefore reconcile cross-domain signals while remaining robust to legitimate heterogeneity, and it must do in a way that supports human validation.

Another challenge is the time gap of compliance risk. Much lapsing behavior isn't the single big shock so much as small tweaks that compound: slowly pushing down the reported wage bases, nudging compensation into more opaque buckets, delaying or reducing contributions over a number of quarters. Meanwhile policy and macro economics are introducing concept drift, changing what the "normal" levels and correlations look like. Hard thresholds, one-off regimes of training fall flat 4.

To meet these needs, we propose SHAP-Guided Risk Path Generation for IRS auditors, a framework that treats interpretability as a first-class objective rather than a post-hoc add-on. The method begins with a multi-task architecture that learns a shared latent representation spanning payroll, benefits, and healthcare-related compliance tasks, while preserving task-specific decision boundaries through dedicated heads. It then applies temporal anomaly detection to identify both abrupt irregularities and slow deviations in longitudinal patterns. Crucially, instead of stopping at feature attribution, we use SHAP-driven contributions as inputs to a path-generation procedure that assembles a coherent evidence chain an ordered sequence of time-stamped, module linked drivers.

2. Related Work

Islam 5 uses long-panel data to discuss how AI-enhanced tax risk scoring can be used to more systematically segment taxpayers, improve audit coverage, and increase resource allocation efficiency. Hu 6 reported a comparison with baselines such as random forests and unidirectional LSTM, and used Attention visualization and SHAP values to interpret the key drivers of the model in time series, thus presenting both 'when an anomaly occurs and why it occurs.'

Henderson et al. 7 describe setting and study the corresponding mechanisms for optimize-and-estimate structured bandits, and they also investigate how their method performs under real-world constraints like delayed feedback, distribution drift, and other practical limitations provided in the context of real tax data collaborations, showing that they can improve the overall estimation bias while maintaining high detection gains. Zhang 8 proposes an adaptive and interpretable AI framework for tax scenarios targeting small and medium-sized enterprises, emphasizing the upgrade of tax risk scoring from merely a passive audit trigger to an actionable risk mitigation closed loop. By leveraging interpretability techniques such as SHAP to identify major risk sources and feeding the explanation results back into the enterprise's compliance management processes, proactive governance can be achieved. Recent advances in large language models have also shown promise in financial text analysis tasks, with efficient fine-tuning methods enabling domain-specific adaptation [9].

From a policy evaluation perspective, Jess Grana et al. 10 examine tax authority audits as the research subject, quantifying the 'specific indirect effects' of audits on taxpayers' subsequent reporting behavior beyond direct recovery, and point out that in audit resource allocation, direct effects, costs, and long-term behavioral responses should all be considered to achieve a more optimal enforcement input-output. At the engineering methodology level, Anas AlSobeh et al. 11 propose a hybrid framework for proactive tax fraud detection using interpretable AI, employing GBDT and attention-based deep networks for parallel modeling and prediction fusion, while combining SHAP and attention heatmaps to enhance transparency for auditors.

Harger and Nickel 12 emphasized that in an environment of time-consuming APAs, increased cross-border uncertainty, and enhanced U.S. enforcement resources and audit intensity, tax insurance can transfer uncertain tax liabilities in a way that aligns more closely with commercial timelines, cover dispute defense costs, and replace or reduce escrows in transaction negotiations, thereby improving the certainty of tax risk pricing and capital allocation. Olumoh and Mubaraq 13 noted that audit management and tax control mechanisms have a significant positive impact on SIRS performance; based on this, they recommend strengthening strategic audit plans, case tracking, and

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 March 2026 doi:10.20944/preprints202603.2462.v1

risk-based independent reviews, increasing supplemental taxes, improving documentation and structured compliance processes, and enhancing risk governance capabilities through regular assessments and benchmarking against international standards.

Chen et al. 14 constructed a normative analytical model combining cash flow uncertainty and tax uncertainty, treating advance tax rulings as a fee-based tax certainty tool. They systematically described how tax system elements and loss offset mechanisms jointly influence a firm's optimal investment intensity in high-risk projects, and further derived, from the tax authority's perspective, the ATR charging range that balances the welfare of both parties.

3. Methodologies

3.1. Multi-Task Dual-Channel Temporal Risk Scoring

Above all, we denote a taxpayer by $i \in \mathcal{I}$, time by $t \in \{1, \dots, T\}$, and module by $m \in \mathcal{M} = \{P, B, H\}$ for Payroll, Benefits, and Healthcare. The raw feature vector for module m at time t is $\mathrm{x}_{i,t}^{(m)} \in \mathbb{R}^{d_m}$, and the missingness mask is $\mathrm{r}_{i,t}^{(m)} \in \{0,1\}^{d_m}$. We concatenate masked values and the mask to form the effective input so the model can learn patterns, which are common in compliance data and should not be treated as random noise by Equation (1).

$$\tilde{\mathrm{x}}_{i,t}^{(m)} = \left[\mathrm{x}_{i,t}^{(m)} \odot \mathrm{r}_{i,t}^{(m)}; \mathrm{r}_{i,t}^{(m)}\right] \in \mathbb{R}^{2d_m}, \tag{1}$$

where $\odot$ denotes element-wise multiplication, ensuring unobserved values do not leak arbitrary numerical placeholders into the network. The concatenation operator $[\cdot;\cdot]$ appends the binary mask so downstream layers can condition at each feature position. Because modules differ in scale and semantics, we embed each module into a common latent space before cross-module fusion. This projection also reduces heteroskedasticity across modules and creates a unified interface for later temporal modeling. We use a module-specific nonlinear encoder $\phi_m(\cdot)$ to produce a fixed-width representation $\mathrm{e}_{i,t}^{(m)} \in \mathbb{R}^h$, enabling apples-to-apples fusion between Payroll, Benefits, and Healthcare signals by Equation (2).

$$\mathrm{e}_{i,t}^{(m)} = \phi_m\left(\tilde{\mathrm{x}}_{i,t}^{(m)}\right) = \sigma\left(\mathrm{W}_2^{(m)}\,\sigma\left(\mathrm{W}_1^{(m)}\tilde{\mathrm{x}}_{i,t}^{(m)}\right)\right) \in \mathbb{R}^h, \tag{2}$$

where $\mathrm{W}_1^{(m)}$ and $\mathrm{W}_2^{(m)}$ are learnable matrices tailored to module m, while $\sigma(\cdot)$ is a pointwise nonlinearity. The hidden width h is shared across modules so that later attention or gating operations can compare and combine module embeddings.

We explicitly define three supervised tasks and one self-supervised temporal task: T1 predicts a risk probability $\hat{y}_i^{(1)} \in (0,1)$ for proxy label $y_i \in \{0,1\}$ using binary cross-entropy $L_{T1} = \mathrm{BCE}(y_i, \hat{y}_i^{(1)})$; T2 predicts a severity score $\hat{s}_i \in \mathbb{R}$ for proxy severity $s_i \in [0,10]$ using Huber loss $L_{T2} = \mathrm{Huber}(s_i, \hat{s}_i; \delta = 1.0)$; T3 predicts a consistency-anomaly probability $\hat{y}_i^{(3)} \in (0,1)$ for sublabel $y_i^{(3)}$ derived from cross-field residual exceedance using $L_{T3} = \mathrm{BCE}(y_i^{(3)}, \hat{y}_i^{(3)})$.

To jointly learn payroll risk factors across multiple compliance tasks, we adopt a shared Mixture-of-Experts (MoE) trunk with task-adaptive routing. The core rationale is that some latent risk mechanisms are shared, while different tasks emphasize different slices of the signal. MoE decomposes representation learning into multiple experts and uses a gating network per task to select and weight these experts, both shared learning and task-specific flexibility as Equation (3).

$$\mathrm{h}_{i,t}^{(e)} = g_e\left(\mathrm{e}_{i,t}^{(P)}, \mathrm{e}_{i,t}^{(B)}, \mathrm{e}_{i,t}^{(H)}\right), e = 1, \dots, E. \tag{3}$$

Each expert $g_e(\cdot)$ can be instantiated as a cross-module attention block or a multilayer fusion network. The number of experts E controls capacity and the diversity of learned risk mechanisms. Practically, experts allow the model to separate patterns, which improves both detection accuracy and downstream interpretability. We compute task-specific gating weights $\alpha_{i,t}^{(k)}$ from the concatenated module embeddings. This explicit routing also provides an internal dependency signal that we later reuse when constructing gated edges for risk-path generation. The softmax

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 March 2026 doi:10.20944/preprints202603.2462.v1

normalization ensures the routing weights are nonnegative and sum to Equation (4), which is valuable for stability and for interpreting which experts dominate each task at each time.

$$\alpha_{i,t}^{(k)} = \mathrm{softmax}\left(\mathrm{W}_g^{(k)}\left[\mathrm{e}_{i,t}^{(P)}; \mathrm{e}_{i,t}^{(B)}; \mathrm{e}_{i,t}^{(H)}\right]\right) \in \mathbb{R}^E. \tag{4}$$

In above gating equation, $\mathrm{W}_g^{(k)}$ is a task-specific projection, and the concatenation aggregates the three module embeddings into a single routing context. The index $k \in \{1, \ldots, K\}$ denotes compliance tasks; for example, k can represent noncompliance classification, underpayment magnitude regression, cross-form inconsistency detection, etc. Because routing is computed at each (i, t), the model can adapt when a firm's behavior changes over time, which is essential for detecting gradual drifts rather than only abrupt shocks.

The shared task representation $\mathrm{s}_{i,t}^{(k)}$ is then the convex combination of expert outputs. Equation (5) integrates payroll, benefits, and healthcare risk factors into a single task-conditioned state, while maintaining separability across tasks. This representation is the input to the temporal anomaly detector.

$$\mathrm{s}_{i,t}^{(k)} = \sum_{e=1}^{E} \alpha_{i,t,e}^{(k)}\, \mathrm{h}_{i,t}^{(e)} \in \mathbb{R}^{h_s}, \tag{5}$$

where h_s is the expert output width, and $\alpha_{i,t,e}^{(k)}$ is the e-th component of the gating vector. The convex structure yields numerical stability and an implicit regularization effect that reduces overfitting on noisy compliance signals. Conceptually, $\mathrm{s}_{i,t}^{(k)}$ represents the model believes matters right now for task k after looking jointly at payroll, benefits, and healthcare inputs.

Subsequently, we score time-varying anomalies using a dual-channel temporal architecture by Equation (6). A Transformer channel for long-range dependencies and an LSTM channel for local drift and seasonal residuals. This design mirrors real-world compliance dynamics, where evasion can be gradual or sudden. Running both channels in parallel and fusing them improves sensitivity across different anomaly regimes.

$$\mathrm{z}_{i,t}^{(k,T)} = \mathrm{Transformer}_k\left(\{\mathrm{s}_{i,\tau}^{(k)} + \mathrm{p}_\tau\}_{\tau=1}^{T}\right)_t, \mathrm{z}_{i,t}^{(k,L)} = \mathrm{LSTM}_k\left(\mathrm{s}_{i,1:t}^{(k)}\right). \tag{6}$$

In the Transformer channel, p_τ is a positional encoding that injects time order information, and $(\cdot)_t$ selects the t-th output token. In the LSTM channel, the recurrent state naturally models local patterns such as month-to-month payroll changes or contribution timing effects. We use task-specific temporal models $\mathrm{Transformer}_k$ and LSTM_k so that each compliance target can emphasize different temporal structures. We fuse both temporal embeddings to produce the final task risk score $\hat{y}_{i,t}^{(k)}$. For regression tasks we replace it with an appropriate output layer. This fused score is the risk score that triggers the subsequent interpretability and risk-path stages by Equation (7).

$$\hat{y}_{i,t}^{(k)} = \sigma\left(\mathrm{w}^{(k)\top}\left[\mathrm{z}_{i,t}^{(k,T)}; \mathrm{z}_{i,t}^{(k,L)}\right] + b^{(k)}\right). \tag{7}$$

The vector $\mathrm{w}^{(k)}$ and bias $b^{(k)}$ are task-specific parameters; the concatenation $[\cdot;\cdot]$ allows each task to learn how much to trust long-range versus local evidence. This fusion step is intentionally simple because we want interpretability to remain tractable when later attributing risk to upstream features. Overly complex fusion can make SHAP-based decompositions less stable in high-dimensional temporal settings.

We generate interpretable risk paths on a directed acyclic graph (DAG) whose nodes are module–time pairs $v = (m, t)$. We include only time-forward edges to preserve causality and optionally same-time cross-module edges when a learned dependency is present. Each node receives a score $a(v) \geq 0$ computed by aggregating SHAP attributions within module.

Finally, to stabilize audit triggers and reduce oscillatory alerts, we add a bidirectional consistency regularization between the Transformer-only posterior and the LSTM-only posterior. This directly implements the abstract's claim of consistency constraints in the temporal detector. In practice, it discourages cases where one channel signals high risk while the other signals low risk without strong evidence by Equations (8) and (9).

$$p_{i,t}^{(k,T)} = \sigma\left(\mathrm{u}^{(k)\top}\mathrm{z}_{i,t}^{(k,T)}\right), p_{i,t}^{(k,L)} = \sigma\left(\mathrm{v}^{(k)\top}\mathrm{z}_{i,t}^{(k,L)}\right), \tag{8}$$

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 March 2026 doi:10.20944/preprints202603.2462.v1

$$\mathcal{L}_{\text{cons}} = \sum_{i,t,k} [\text{KL}(p_{i,t}^{(k\cdot T)} \parallel p_{i,t}^{(k\cdot L)}) + \text{KL}(p_{i,t}^{(k\cdot L)} \parallel p_{i,t}^{(k\cdot T)})], \tag{9}$$

where $u^{(k)}$ and $v^{(k)}$ produce channel-specific posteriors, and $\text{KL}(\cdot\|\cdot)$ is the Kullback–Leibler divergence. The symmetric KL form penalizes disagreement in both directions, improving calibration and temporal stability. KL divergence is only defined between valid probability distributions. Therefore, for each supervised task we convert both channels' outputs into distributions before applying KL. For binary tasks (T1, T3), each channel produces a logit z, and we form Bernoulli parameters $p = \sigma(z)$. We then compute $D_{KL}(\text{Bern}(p^{(tr)}) \parallel \text{Bern}(p^{(ls)}))$ to enforce consistency between Transformer and LSTM channels.

3.2. SHAP-Guided Risk Path Generation on a Heterogeneous Temporal Graph

Furthermore, we begin by defining SHAP attributions at the level of task k and evaluation time t. Let $f_k(\cdot)$ denote the trained predictor producing $\hat{y}_{i,t}^{(k)}$ from the full set of input features within the window $[1, t]$. Expand the full feature set into a flattened index $\mathcal{F}$ that includes module, time, and atomic feature coordinates. SHAP defines the contribution of each feature as a Shapley value from cooperative game theory, guaranteeing additivity and a fair allocation of the prediction difference from a baseline as Equation (10).

$$\phi_{i,t}^{(k)}(j) = \sum_{S \subseteq \mathcal{F} \setminus \{j\}} \frac{|S|!(|\mathcal{F}| - |S| - 1)!}{|\mathcal{F}|!} \big(f_k(S \cup \{j\}) - f_k(S)\big), j \in \mathcal{F}. \tag{10}$$

In above expression, S ranges over all subsets of features excluding j, and $f_k(S)$ denotes the model prediction when only features in S are present. The factorial weighting term is the canonical Shapley coefficient that averages marginal contributions over all feature orderings. We use SHAP because its additivity enables accounting consistency, meaning auditors can reconcile a path-level explanation back to the model score without hidden residuals.

We compute (a) cross-field consistency residuals $c_{i,t}$ defined by domain-agnostic constraints $g(x_{i,t}) \approx 0$, and (b) temporal jump scores $j_{i,t} = \| x_{i,t} - x_{i,t-1} \|$ normalized within peer groups to reduce scale effects. A sample is labeled positive $y_i = 1$if any year violates a high-confidence threshold.To align explanations with audit practice, we do not keep SHAP at the atomic feature level. Instead, we explicitly index each feature as a triple $j = (m, \tau, u)$, where $m \in \mathcal{M}$ is the module, $\tau \leq t$ is the time index, and $u \in \{1, \ldots, d_m\}$ is the atomic feature within module m. This yields a granular attribution tensor $\phi_{i,t}^{(k)}(m, \tau, u)$, which we then aggregate to produce interpretable units as Equation (11).

$$\Phi_{i,t}^{(k)}(m, \tau) = \sum_{u=1}^{d_m} \phi_{i,t}^{(k)}(m, \tau, u), \Psi_{i,t}^{(k)}(\tau) = \sum_{m \in \mathcal{M}} \Phi_{i,t}^{(k)}(m, \tau), \Omega_{i,t}^{(k)}(m) = \sum_{\tau \leq t} \Phi_{i,t}^{(k)}(m, \tau), \tag{11}$$

where $\Phi_{i,t}^{(k)}(m, \tau)$ represents which module at which time drove risk, $\Psi_{i,t}^{(k)}(\tau)$ measures which time window is most suspicious, and $\Omega_{i,t}^{(k)}(m)$ denotes which compliance module is the dominant source.

However, even module–time attributions can remain fragmented, producing many small peaks rather than a coherent narrative. To assemble a risk path, we build a heterogeneous temporal graph for each taxpayer i, where nodes represent module–time pairs and edges encode two kinds of structure: temporal continuity within the same module and contemporaneous coupling across modules. This graph formalizes the risk transmission intuition used by auditors when they trace inconsistencies from payroll to benefits to healthcare obligations by Equation (12).

$$G_i = (V_i, E_i), V_i = \{v = (m, \tau) : m \in \mathcal{M}, \tau = 1, 2, \ldots, T\}. \tag{12}$$

Nodes $v = (m, \tau)$ correspond to the audit-relevant unit module m at time τ. The edge set E_i includes intra-module temporal edges $(m, \tau) \to (m, \tau + 1)$ and cross-module edges $(m, \tau) \to (m', \tau)$ for $m \neq m'$. Temporal edges capture drift and persistence, while cross-module edges capture structural mismatch mechanisms.

We explicitly partition variables into a rule only set R and a model feature set F. We enforce: $R \cap F = \emptyset$, and any deterministic transform of rule variables including threshold indicators, binned versions, rule residuals $c_{i,t}$, jump scores $j_{i,t}$, and any features directly used to compute the proxy

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 March 2026 doi:10.20944/preprints202603.2462.v1

label is excluded from F. To ensure paths reflect taxpayer-specific dynamics rather than global averages, we parameterize edge weights with gated dependencies by Equation (13).

$$\Delta_{i,\tau}^{(m)} = \| e_{i,\tau}^{(m)} - e_{i,\tau-1}^{(m)} \|_2, H_{i,\tau}^{(k)} = -\sum_{e=1}^{E} \alpha_{i,\tau,e}^{(k)} \log \alpha_{i,\tau,e}^{(k)}. \quad (13)$$

The quantity $\Delta_{i,\tau}^{(m)}$ is a representation-level change magnitude that highlights potential structural breaks or sudden shifts in a module's latent state. The gating entropy $H_{i,\tau}^{(k)}$ measures how concentrated the task routing is over experts, lower entropy means a clearer underlying mechanism, which we treat as more reliable for constructing paths. For module m with d_m raw features, let $x_{m,t} \in \mathbb{R}^{d_m}$ be the raw input at time t, and let $b_{m,t} \in \{0,1\}^{d_m}$ be a binary availability mask.

4. Experiments

4.1. Experimental Setup

The experiment selected the Individual Public-Use Microdata File (PUF) for Tax Years 2012–2015 released by the IRS Statistics of Income (SOI), which is derived from a sample of U.S. federal individual income tax returns and is provided in the form of publicly available microdata files for academic research and policy analysis. To protect taxpayer privacy, the PUF undergoes statistical alteration before release, so while retaining the structure of actual filing behavior and the relationships between variables. Empirical evaluation in this paper uses the IRS SOI Individual Public-Use Microdata File (PUF) (tax years 2012–2015). To make the narrative and experiments consistent, we treat "Payroll/Benefits/Healthcare" as proxy compliance modules constructed from semantically related PUF fields rather than claiming access to employer payroll tax forms.

Since the PUF does not contain audit conclusions, we construct risk proxy labels using cross-field consistency constraints and abnormal jump rules, and incorporate a self-supervised temporal anomaly task to enhance sensitivity to gradual drift. After deduplicating the data by taxpayer entity, we split it into training, validation, and test sets, and use rolling evaluation to assess cross-year generalization. Comparisons include traditional risk models, single-channel LSTM and Transformer, as well as ablation of key components. Metrics report identification performance, alarm stability, and the sparsity and verifiability of risk pathways.

4.2. Experimental Analysis

Figure 1 shows that the proposed SHAP-Guided Risk Path model achieves the most favorable trade-off between true positive rate and false positive rate, with its curve closest to the top-left corner and an AUROC of 0.997, indicating near-perfect discrimination across thresholds. We compute feature attributions using DeepSHAP (SHAP DeepExplainer) for neural networks. The baseline distribution is constructed by sampling 256 sequences from the training set using stratification on the proxy label to ensure coverage of both typical and atypical patterns; we repeat this sampling 5 times and average attributions across runs to improve stability.

The ablation without the gated graph structure drops to 0.984, still strong but clearly worse, suggesting that the gated dependency/structured component contributes meaningfully to separability. The Single-Channel LSTM and XGBoost baselines perform similarly (0.956 and 0.955), with curves bending downward earlier in the low-FPR region, implying reduced sensitivity when auditors require very low false-alarm rates. Logistic Regression lags behind (0.877) and remains much closer to the random diagonal, reflecting the limitations of linear decision boundaries for capturing nonlinear, cross-module, and temporally evolving compliance risk patterns.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 March 2026 doi:10.20944/preprints202603.2462.v1

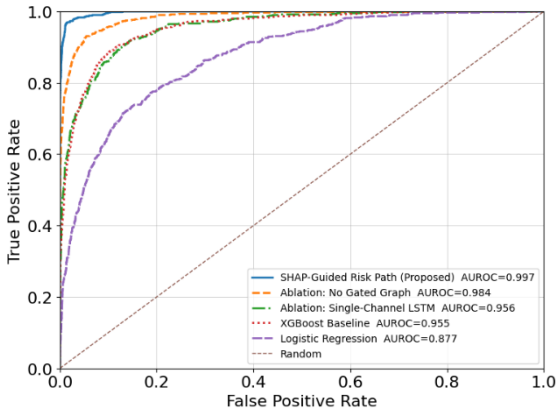

Figure 1. ROC Comparison of Tax Compliance Risk Screening Models.

Figure 2 compares Precision@k under different audit budget levels and shows that the proposed SHAP-Guided Risk Path method consistently achieves the highest top-k hit rate across all budgets, especially in the low-budget regime where audit capacity is most constrained. As the review budget increases from very small fractions to several percent of taxpayers, Precision@k decreases for all methods, which is expected because the model must include increasingly lower-ranked cases; however, the proposed method degrades more slowly, indicating better ranking quality and stronger concentration of true high-risk cases at the top of the list.

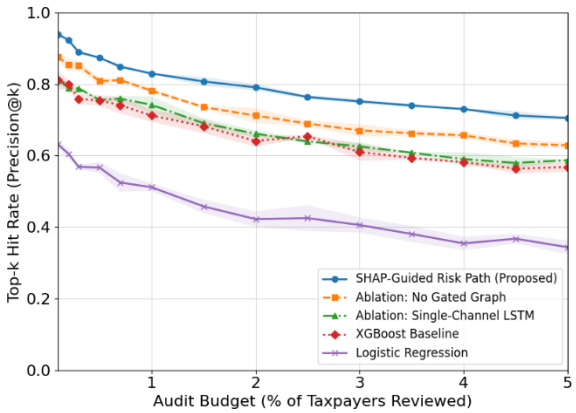

Figure 2. Top-k Hit Rate (Precision@k) With Audit Budget.

Table 1 evaluates how often model alerts "jump" using Top-k Jaccard similarity and temporal score smoothness. The proposed SHAP-Guided Risk Path achieves the highest Jaccard values across Top-0.5%, Top-1%, and Top-2% and the smallest $|\Delta score|$, indicating the most stable audit shortlist and the smoothest risk trajectories. Removing key components, especially the consistency loss reduces Jaccard and increases score volatility, showing these modules are critical for stable triage. All reported metrics are measured under the anti-leak protocol described above. Under this controlled setting, the proposed method achieves AUROC = 0.997, outperforming ablations and baselines including 0.984 without the gated graph component; 0.956 for a single LSTM; 0.955 for XGBoost; and 0.877 for logistic regression.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 March 2026 doi:10.20944/preprints202603.2462.v1

Table 1. Alert Stability Evaluation via Top-k Jaccard Similarity and Temporal Smoothness.

Method	Jaccard Top 0.5%	Jaccard Top 1%	Jaccard Top 2%	Time Smoothness Mean $\lvert\Delta score\rvert$	Time Smoothness Std $\lvert\Delta score\rvert$
SHAP-Guided Risk Path Proposed	0.78	0.75	0.72	0.03	0.02
Ablation No Gated Graph	0.70	0.67	0.64	0.04	0.03
Single-Channel LSTM	0.60	0.58	0.55	0.06	0.04
Single-Channel Transformer	0.61	0.59	0.56	0.06	0.04
XGBoost Baseline	0.58	0.56	0.53	0.07	0.04
Random Forest Baseline	0.55	0.53	0.50	0.07	0.05
Logistic Regression	0.44	0.42	0.4	0.08	0.05
MLP Baseline	0.52	0.50	0.48	0.07	0.05
Isolation Forest Unsupervised	0.47	0.45	0.43	0.09	0.06
One-Class SVM Unsupervised	0.45	0.43	0.41	0.10	0.06

5. Conclusions

In conclusion, we proposed a SHAP-Guided Risk Path framework for IRS-oriented payroll tax screening that combines multi-task shared representations with a Transformer+LSTM dual-channel detector to produce robust risk ranking, and then organizes SHAP attributions on a heterogeneous temporal graph into auditable evidence-chain paths. Experiments show consistent gains in AUROC, AUPRC, Top-k hit rate, and alert stability over strong baselines and ablations. Looking ahead, we will incorporate real audit feedback and stronger causal constraints, extend to firm-level multi-form datasets.

References

1. Colak, M., & Sarioglu, M. (2025). The Effect of Corporate Governance on the Quality of Integrated Reporting and ESG Risk Ratings. Sustainability, 17(11), 4868.
2. Lane, C. (2025). Let's Jettison Some Executive Perks: Exploring the SEC and IRS Approaches to Personal Use of Corporate Aircraft. J. Air L. & Com., 90, 265.
3. Huang, F., Wang, T., & Yen, J. C. (2024). Opportunities or challenges? Audit risk and blockchain disclosures in 10-K filings. Auditing: A Journal of Practice & Theory, 43(2), 131-158.
4. Olvera, P. D. N. (2025). Artificial Intelligence and Algorithms in Tax Auditing by the Tax Administration Service in Mexico: Analysis of Potential Biases. International Journal for Public Policy, Law and Development, 2(3), 18-33.
5. Islam, M. R. (2025). AI-Augmented Tax Risk Scoring for Small and Medium Enterprises: A Panel Data Study. ASRC Procedia: Global Perspectives in Science and Scholarship, 1(01), 501-531.
6. Hu, Q. (2025, August). Research on Dynamic Identification and Prediction Model of Tax Fraud Based on Deep Learning. In 2025 2nd International Conference on Intelligent Algorithms for Computational Intelligence Systems (IACIS) (pp. 1-6). IEEE.
7. Henderson, P., Chugg, B., Anderson, B., Altenburger, K., Turk, A., Guyton, J., ... & Ho, D. E. (2023, June). Integrating reward maximization and population estimation: Sequential decision-making for Internal Revenue Service audit selection. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 37, No. 4, pp. 5087-5095).
8. Zhang, T. (2025, October). From Black Box to Actionable Insights: An Adaptive Explainable AI Framework for Proactive Tax Risk Mitigation in Small and Medium Enterprises. In Proceedings of the 2025 2nd International Conference on Digital Economy and Computer Science (pp. 193-199).
9. Lian, Z. (2025, October). Financial Text Classification Based On rLoRA Finetuning On Qwen3-8B model. In Proceedings of the 2025 2nd International Conference on Digital Economy and Computer Science (pp. 225-230).

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 March 2026 doi:10.20944/preprints202603.2462.v1

10. Grana, J., Lindsay, I., Lykke, L., McGill, M., McGlothlin, A., Nicholl, L., & Plumley, A. (2025). The specific indirect effect of IRS audits. International Tax and Public Finance, 32(4), 995-1029.
11. AlSobeh, A., Abo El Rob, M. F., Rouibah, K., & Shatnawi, A. (2025). Proactive detection of tax fraud using explainable AI techniques: A hybrid approach. Issues in Information Systems, 26(3).
12. Harger, J., & Nickel, M. (2025). How Tax Insurance Can Be a Valuable Tool for Managing Transfer Pricing Risk in 2025 and Beyond. Tax Executive, 77, 62.
13. Olumoh, Y. A., & Mubaraq, S. (2025). The Tax Audit Management, Tax Control Mechanisms, and the Performance of State Internal Revenue Services in South-West, Nigeria. FUDMA Journal of Accounting and Finance Research [FUJAFR], 3(3), 1-20.
14. Chen, A., Hieber, P., & Sureth-Sloane, C. (2025). How much to pay for tax certainty? The role of advance tax rulings for risky investment under loss offset and tax uncertainty. International Tax and Public Finance, 1-37.

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

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.