Issues in Information Systems
Volume 26, Issue 3, pp. 255-274, 2025

DOI: https://doi.org/10.48009/3_iis_2025_2025_120

Proactive detection of tax fraud using explainable AI techniques: A hybrid approach

Anas AlSobeh, *Southern Illinois University, Carbondale, anas.alsobeh@siu.edu*
Mustafa Farouk Abo El Rob, *University of Denver, mustafa.aboelrob@du.edu*
Kamel Rouibah, *University of Kuwait, kamel.rouibah@ku.edu.kw*
Amani Shatnawi, *Weber State University, amanishatnawi1@weber.edu*

Abstract

Tax fraud continues to cause significant financial losses for governments worldwide. In the United States alone, the Internal Revenue Service estimates an annual tax gap of approximately $441 billion, much of it is due to intentional evasion and fraud. Traditional detection methods are based on static rules, manual audits, and retrospective analysis which often fail to keep pace with increasingly sophisticated and adaptive fraud schemes. In this research study, we develop an advanced hybrid artificial (AI) framework for tax fraud detection, combining a gradient-boosted decision tree (GBDT) model (XGBoost) with a deep neural network (DNN) that incorporates an attention mechanism. Crucially, we integrate explainable AI (XAI) techniques (e.g., SHAP and attention heatmaps) so that the model's predictions are transparent to tax auditors. Using a realistic synthetic dataset of 10,000 tax returns (10% fraud), our hybrid model achieved 92% accuracy and 88% recall, 91% precision, and 0.95 AUC, significantly outperforming conventional approaches. Moreover, it enables proactive detection by identifying potentially fraudulent returns early in the filing process (before full processing) and generating clear explanations for each flagged return. This approach advances both theoretical research and practical application by demonstrating the effectiveness of hybrid modeling and XAI in regulatory settings and by providing a scalable roadmap for tax authorities to modernize fraud detection.

Keywords: tax fraud detection, explainable AI (XAI), gradient boosting, machine learning, deep neural networks (DNN), synthetic tax data, proactive analytics.

Introduction

Tax fraud and evasion remain persistent challenges for tax authorities worldwide. The scale of the problem is staggering, with the Internal Revenue Service (IRS) in the United States estimating a tax gap of approximately $441 billion annually, much of it stemming from deliberate fraud and evasion (IRS, 2023). Selecting taxpayers for audits has long been a key concern for policymakers seeking to enhance the efficiency of tax administration (De Roux et al., 2018). As a result, identifying and preventing tax fraud has therefore become a top priority for revenue agencies globally. Tax fraud refers to the intentional misrepresentation or falsification of information on a tax return to reduce one's tax liability (De Roux et al., 2018).

Traditional detection methods have relied heavily on manual reviews, rule-based systems, and random

audits. While these techniques have been essential in maintaining compliance, a substantial portion of the tax gap still arises from willful fraud and misrepresentation. Under U.S. law, evading or defeating tax (26 U.S.C. §7201) is classified as a felony, and willfully filing a false return (26 U.S.C. §7206) constitutes a criminal offense. Nevertheless, fraudsters continuously adapt to escape detection. These traditional approaches, though foundational, are becoming increasingly resource-intensive and often ineffective against sophisticated fraud schemes that evolve rapidly in response to detection efforts.

Meanwhile, tax authorities collect an unprecedented volume of taxpayer data. Each year brings more tax returns, third-party reporting (e.g., employer-issued W-2s), financial transactions, and even international information exchanges through frameworks such as the Foreign Account Tax Compliance Act (FATCA). This exponential growth in data presents both significant challenges and opportunities. While the volume and complexity of data make manual analysis increasingly impractical, they also create fertile ground for advanced analytical approaches.

The application of AI in this context brings with it several domain-specific concerns, including the protection of taxpayer privacy, the need for explainable decisions, and the dynamic nature of fraud schemes. The integration of AI into tax administration also poses operational and ethical challenges. First, errors have high stakes: false positives (flagging an honest taxpayer) can waste government resources and burden compliant citizens, while false negatives allow fraudulent activity to persist undetected. Second, tax agencies and citizens alike demand transparency.

Unlike some commercial applications, tax decisions are subject to strict fairness and accountability standards. As such, the use of an opaque "black box" model is problematic, particularly when it can trigger audits or penalties. The IRS and the courts require that decision-making processes be explainable and grounded in tax law. This means any AI system must provide interpretable reasons for its alerts, and aligned with legal requirements.

Taxpayer privacy laws and the absence of publicly available tax datasets present significant challenges for research in AI-based fraud detection (Davidson et al., 2025). To address this, our study employs synthetic data generated based on a set of tax fraud related rules to support the development, training, and evaluation of the developed system. In doing so, our research seeks to develop and evaluate a novel AI-based system for proactive tax fraud detection, guided by the following objectives: (1) Generate realistic synthetic data offers approaches for other sensitive applications where privacy concerns limit access to real data. (2) Design a hybrid AI architecture that combines the strengths of different machine learning (Adamov et al., 2019) approaches for improved fraud detection accuracy and robustness. (3) Develop explainable AI components that provide transparent justifications for fraud classifications, addressing critical requirements for fairness and accountability in tax administration. (4) Implement proactive detection capabilities that identify potential fraud before tax returns are fully processed, enabling earlier intervention and more efficient resource allocation. (5) Evaluate the performance of the proposed system against traditional methods using realistic synthetic data that captures the complexity of real-world fraud patterns.

Our hybrid approach (source code: https://github.com/aalosbeh/tax-fraud-xai) builds on a novel pipeline that first applies a GBDT model for initial fraud scoring and feature ranking, and then feeds those insights (along with raw features) into a DNN enhanced with an attention mechanism. The attention layer enables the DNN to weigh features adaptively, focusing on the most relevant inputs for each case. By combining models, we achieve higher detection rates than either approach alone. Simultaneously, we embed an explainability framework (i.e., XAI) using SHAP values to quantify feature contributions and attention heatmaps to visualize which inputs most influence each prediction. These XAI provide both global insights

(e.g. which features are most important overall) and local explanations (e.g. why a specific return was flagged).
Most existing tax fraud detection systems operate reactively, analyzing returns only after they are fully submitted, often well past the filing deadline. In contrast, our approach shifts from reactive to proactive detection. It enables proactive detection by identifying high-risk cases earlier in the filing process, even when only partial or preliminary data is available. This allows audits or verifications to begin sooner, potentially deterring fraud. We evaluate this approach by tracking the proportion of fraud cases our model successfully identifies at different stages of the return processing lifecycle.

Literature Review

Historically, tax fraud detection systems have largely depended on static rules, random audits, and manual reviews. These traditional approaches tend to be reactive and are inadequate for countering increasingly complex and adaptive fraud tactics. As tax evasion methods evolve, there is a growing demand for more proactive AI-based techniques that enable earlier detection and improved accuracy, thereby enhancing compliance and reducing revenue loss.

Hybrid-Modular AI Attention Approaches
Alsadhan (2023) explored a hybrid AI approach by combining tree-based algorithms with deep learning for financial fraud detection, using both supervised and unsupervised training to Saudi tax invoices. While the study reported improved detection accuracy across individual modules, it did not incorporate explainability or evaluation mechanisms critical for regulatory compliance. Similarly, Sailaja (2024) combined AdaBoost classifiers with deep learning techniques to predict tax evasion based on financial indicators. Although the approach demonstrated strong retrospective performance, it lacked temporal awareness or stakeholder-focused explanations

One of the key contributions to hybrid modeling was done by Chagahi et al. (2025), who developed an attention-based ensemble model for financial fraud detection that combines convolutional neural networks (CNNs) and graph neural networks (GNNs) with a confidence-driven gate mechanism. Their approach puts two important layers on top of baseline classification: an attention layer that selectively weighs model predictions and a confidence-based fusion layer that selects the most reliable sum based on uncertainty estimates. This architecture demonstrates that hybrid models can be accurate and interpretable something of paramount concern to tax authorities. The most effective hybrid frameworks are the integration of supervised learning with unsupervised learning.

Similar ideas were applied in the RADAR framework for detecting radicalization in Arabic social media, which used a deep-learning and transformer ensemble architecture to capture complex patterns (Al-Shawakfa et al., 2025). In a strong review by Wolters Kluwer (2025), "A hybrid approach to detecting fraud is likely to yield the best outcomes. Rather than relying solely on AI, internal audit leaders must integrate AI potential and human expertise." With integration, systems can detect known schemes of fraud using supervised learning and novel, unseen schemes using unsupervised techniques (Bansal et al., 2025). Additionally, the integration of LLMs and statistical reasoning for explainable runtime systems, as demonstrated in prior work on AI-powered AOP (AlSobeh et al., 2024), could be adapted to enhance fraud explainability and live monitoring. The fairness impact of proactive audits on vulnerable taxpayer segments (e.g., refugees or low-income groups) should be studied in depth, following the social empowerment lens explored in (Al-Shraifin et al., 2024).

Attention mechanisms have revolutionized deep learning approaches to fraud detection by enabling models to focus on the most critical features or temporal patterns within complex datasets. Murorunkwere et al.

(2022) employed artificial neural networks to detect income tax fraud in Rwanda, reporting high recall and AUC metrics. However, their approach lacked attention mechanisms and explainability tools critical for ensuring transparency and legal defensibility in tax administration. The integration of attention mechanisms with traditional deep learning architectures represents a significant advancement, allowing models to adaptively weight features based on their relevance to specific cases. Zhou et al. (2023) designed a user-centric, interpretable AI approach that employed ensemble models with Shapley values to support both local and global explanation of fraud classification.

While effective in general financial fraud detection, the approach failed to address domain-specific complexities present in tax systems. Hackernoon (Kapoor, 2025) reports that "These algorithms utilize self-attention mechanisms to detect sophisticated and sequential fraud behaviors in real-time, particularly effective in identifying complex patterns that evolve over time." This capability is especially valuable for tax authorities, as fraudulent behaviors often manifest across multiple filings or reporting periods.

Proactive and Explainability in AI Models

The U.S. Treasury Department has recognized the value of this approach, announcing in October 2024 the implementation of "enhanced processes, including machine learning AI, to deal with increased fraud and improper payments." This shift toward proactive detection represents a fundamental change in how tax authorities approach fraud prevention.

XAI has emerged as a critical requirement in high-stakes domains such as tax enforcement. Darwish et al. (2024) demonstrated how techniques like LIME and SHAP assign each input feature an importance score for a given prediction, enabling human scrutiny. These techniques have been used in finance and healthcare to satisfy transparency regulations, but their adoption in tax systems represents a novel application. A recent study by Mill et al, (2023) analyzed whether current XAI techniques can address taxpayer concerns about AI use in taxation, finding that "explainability is not merely a technical consideration but a legal and ethical imperative in tax administration."

This perspective is echoed by Datos Insights (2025), which notes that "The demand for explainable AI in fraud detection, while clear, presents significant challenges due to its high-stakes nature and complexity." In addition, such these techniques have been used in finance and healthcare to satisfy transparency regulations, but their adoption in tax systems is novel (Darwish et al., 2024) (Aladebumoye, 2025). Our work builds on this trend by integrating XAI not as an afterthought but as a core design element. We also extend hybrid modeling ideas from other domains: combining GBDT (like XGBoost) with neural nets has shown improved accuracy in some applications (Dave et al., 2025) (AlShattnawi et al, 2024) (Hassouna et al., 2024). To our knowledge, this is the first such application that brings these pieces together for tax fraud detection, with attention layers to adaptively capture feature interactions.

Verheij et al. (2024) recently evaluated the effectiveness of current XAI techniques in meeting legal standards for transparency and accountability within public tax administration. Their findings highlight the critical need to balance model complexity with interpretability, especially when deploying AI systems in governmental decision-making contexts. To address data privacy limitations, we generate high-fidelity synthetic datasets using distribution-aware simulations. This approach enables reproducible experimentation without compromising taxpayer confidentiality, an often-overlooked limitation in literature.

The current research addresses these gaps by proposing a novel hybrid architecture that combines GBDT with DNN, most existing models remain limited by their lack of proactive detection capabilities and built-in explainability. These gaps are particularly critical given the evolving expectations around transparency,

accountability, and early intervention in tax administration. To address these shortcomings, our work introduces interpretable, high-performance hybrid modeling that incorporates dynamic risk scoring across the entire tax processing lifecycle. By doing so, we not only enhance fraud detection accuracy but also respond to the ethical, operational, and regulatory imperatives of modern tax administration.

Methodology

Data Generation and Feature Engineering

Tax data presents unique challenges for research due to its sensitive nature and strict confidentiality requirements (Alsobeh et al., 2024). Real tax return data is typically inaccessible to researchers outside tax authorities, and even within authorities, its use is highly restricted. To overcome these limitations while enabling rigorous research, we developed a synthetic data generation approach that simulates realistic tax return patterns while avoiding privacy concerns.

Our synthetic data generator creates taxpayer profiles with demographic, income, deduction, and filing characteristics that mirror real-world distributions. We based these distributions on publicly available statistics from tax authorities, including the IRS Statistics of Income (SOI) in the United States and similar publications from other jurisdictions.

For example, income follows a log-normal distribution with parameters estimated from published income statistics, while deduction amounts maintain realistic relationships with income levels across different demographic segments. The generator incorporates natural correlations between variables based on economic and behavioral realities.

For instance, mortgage interest deductions correlate with income levels and geographic regions, while business expense deductions vary by industry (i.e., employment) sector and self-employment status. These correlations ensure that the synthetic data captures the complex relationships present in real tax data.

We injected four primary fraud patterns into the synthetic data, based on patterns documented in tax authority publications and academic literature:

(1) Hiding 20-50% of actual income, implemented by generating true income and then reducing the reported amount while maintaining realistic relationships with other variables.

(2) Deduction inflation by 50-200% across various categories, with particular focus on categories that are difficult to verify (e.g., business expenses (i.e., self-employment expenses), charitable contributions).

(3) Credit abuse, particularly those targeted at lower-income taxpayers or specific activities like education or energy efficiency.

(4) Identity theft that fabricated returns with unusual patterns, such as early filing, maximum refund claims, and minimal supporting documentation.

We then created mixed fraud cases that combine multiple techniques, reflecting the reality that sophisticated fraud often employs multiple methods simultaneously. The final dataset contains 10,000 tax returns with a 10% fraud rate. While this rate is higher than real-world rates (typically 1-3% according to tax authority estimates), it provides sufficient positive examples for effective model training while still maintaining class imbalance challenges. To address this discrepancy, we conducted an additional sensitivity experiment by retraining and evaluating the model on versions of the dataset adjusted to reflect more realistic fraud rates of 1%, 2%, and 3%. Results showed that the hybrid model maintained strong performance, with AUC values remaining above 0.91 and F1-scores decreasing by less than 8%. These findings demonstrate robustness to lower fraud prevalence while still ensuring sufficient detection. Performance trade-offs were primarily in recall, as expected, but explainability remained consistent through SHAP visualizations and attention heatmaps.

Figure 1 shows the distribution of key variables (income, deduction-to-income ratio, credit-to-income ratio) for legitimate vs. fraudulent returns. It displays clear differences in distributions between the two classes, with fraudulent returns showing more extreme values in deduction and credit ratios. Moreover, Figure 2 shows engineered additional features beyond the basic tax return elements to capture patterns and relationships that might indicate fraudulent activity. Ratio features provide normalized measures of relationships between key variables, controlling scale differences across taxpayers. We calculated several ratio features, including: Deduction-to-income ratio (total deductions divided by reported income), Credit-to-income ratio (total credits divided by reported income), Category-specific ratios (e.g., charitable contributions as a percentage of income), Expense-per-dependent ratio (total deductions divided by number of dependents plus one), Temporal features capture patterns related to timing of filing activities. Research by the IRS and other tax authorities has found that filing timing often correlates with fraud risk, with certain

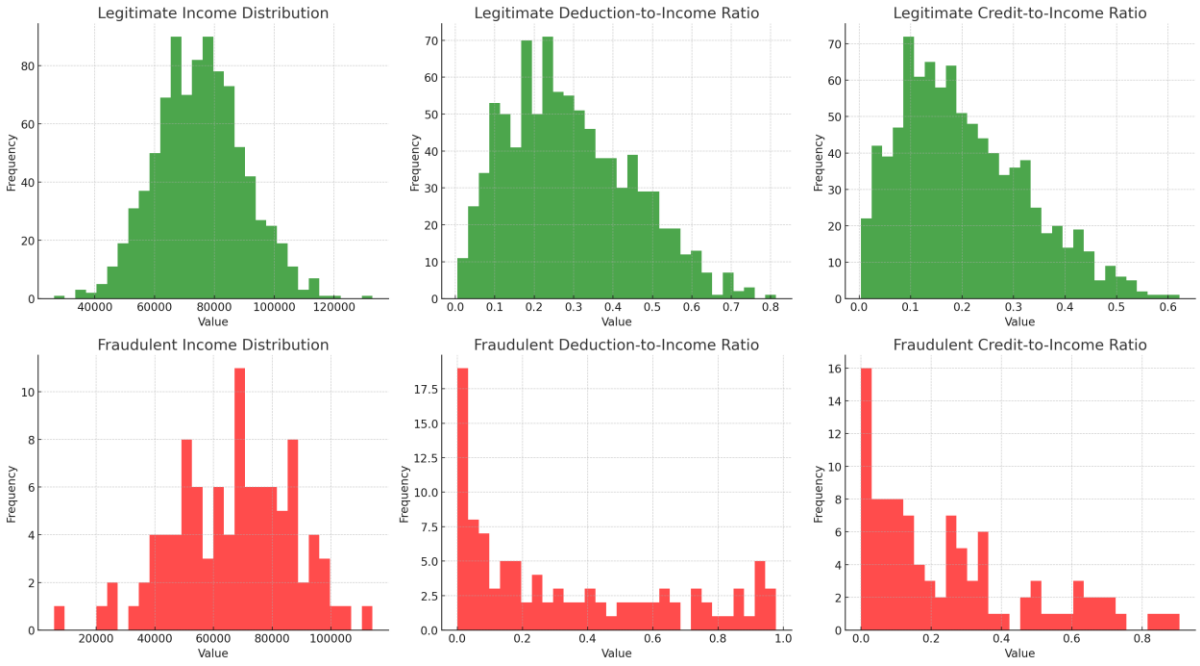

Figure 1. The distributions of key variables in our synthetic dataset, comparing legitimate and fraudulent returns

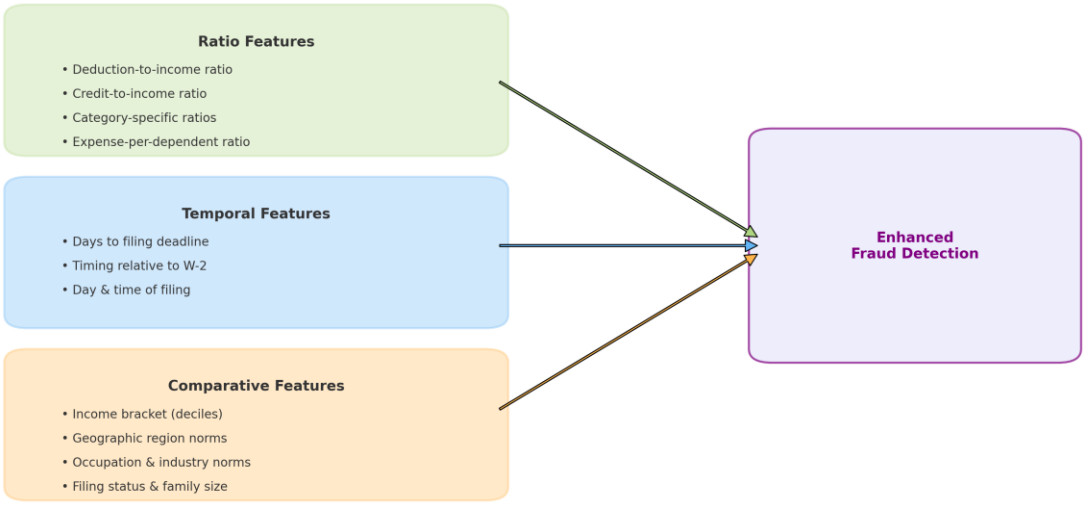

Figure 2. Feature Engineering Framework for Tax Fraud Detection

fraud types (particularly identity theft) showing distinct temporal patterns. We included features such as (Days to filing deadline (number of days between filing date and tax deadline), Filing timing relative to information receipt (e.g., timing relative to W-2 availability), Day of week and time of day for electronic filings). Peer comparison features quantify how a return deviates from similar taxpayers (e.g. z-scores of key metrics within income deciles or geographic regions). All categorical fields (occupation, industry, filing status) were encoded appropriately (one-hot or embeddings for the DNN). We applied standard preprocessing: missing values were imputed contextually, continuous variables were standardized for the neural net, and extreme outliers were clipped.

Data Preprocessing

We applied standard preprocessing techniques to prepare the data for modeling, paying particular attention to issues characteristic of tax data. Missing values are a normal feature of tax data, often in meaningful patterns. Rather than simple imputation, we used a context-aware approach: (1) For demographic variables, we imputed group medians based on comparable demographic groups. (2) For financial variables, we used regression-based imputations with related financial data. (3) We established binary indicators for missingness patterns that might be informative.

Categorical variables require special handling to maintain their information content. We employed a mix of approaches (1) One-hot encoding for low-cardinality variables (filing status, etc.) (2) Target encoding for high-cardinality variables (occupation codes, etc.), cross-validating suitably to prevent leakage (3) Embedding approaches for categorical variables input into the neural network phase.

Feature scaling ensured that all variables contributed equally to model training. We applied standardization (mean subtraction and division by standard deviation) to the neural network portion, but the GBDT portion handled unscaled features well. Class imbalance was a significant issue even with our synthetic dataset having a high fraud rate. We addressed this by using a mix of the following: (1) Synthetic Minority Over-sampling Technique (SMOTE) for generating additional synthetic instances of fraud. (2) Class weighting while training the model to assign higher weights to the minority class. (3) Threshold adjustment in end classification to achieve a balance between precision and recall.

Hybrid Model Architecture

Figure 3 shows our hybrid model architecture through parallel paths of GBDT and DNN components, with their outputs combined in a meta-learner. The architecture implements a parallel processing approach, with both the GBDT classifier (using XGBoost) and DNN components receiving the input data simultaneously. The GBDT component performs feature selection and generates an initial fraud probability, while the DNN component captures complex patterns and interactions. The meta-learner then combines the outputs from both components to produce the final fraud probability. This parallel architecture differs from typical sequential ensemble approaches where models are trained in sequence. The parallel approach allows each component to learn independently based on its strengths, with the meta-learner determining how to optimally combine their insights. This design is motivated by the complementary strengths of the components: XGBoost excels at rule-like reasoning and capturing strong low-order interactions, while the DNN with attention layers captures nuanced, nonlinear relationships that evolve contextually. Our ablation tests confirm that the hybrid model outperforms either component alone, both statistically and practically, especially on complex or mixed-type fraud cases. By keeping both models independent before fusion, we maximize specialization and reduce overfitting risk seen in sequential ensembles.

XGBoost was configured with 500 trees (max depth 6) and a learning rate of 0.01; early stopping on validation loss prevented overfitting. The GBDT naturally handles missing values and produces feature importance scores. This component excels at capturing strong, univariate or low-order interactions in tabular data. It outputs a preliminary fraud probability (GBDT_score) and a ranked list of influential features. To prevent overfitting, we implemented early stopping based on validation performance and applied L1 and L2 regularization. This configuration balances model complexity with generalization capability, preventing overfitting while capturing complex non-linear relationships in tax data. The XGBoost model serves two critical functions in our pipeline: (1) generating initial fraud risk scores and (2) ranking feature importance to inform the attention mechanism in the subsequent DNN component.

Simultaneously, DNN takes as input both the original features and the GBDT's fraud score. By including the XGBoost output, the DNN can build on the tree model's strengths and focus on patterns it may have missed. Our DNN has two hidden layers of 128 units each (ReLU activation) plus an attention layer. The attention mechanism (inspired by transformer-style self-attention) computes a weight for each input feature,

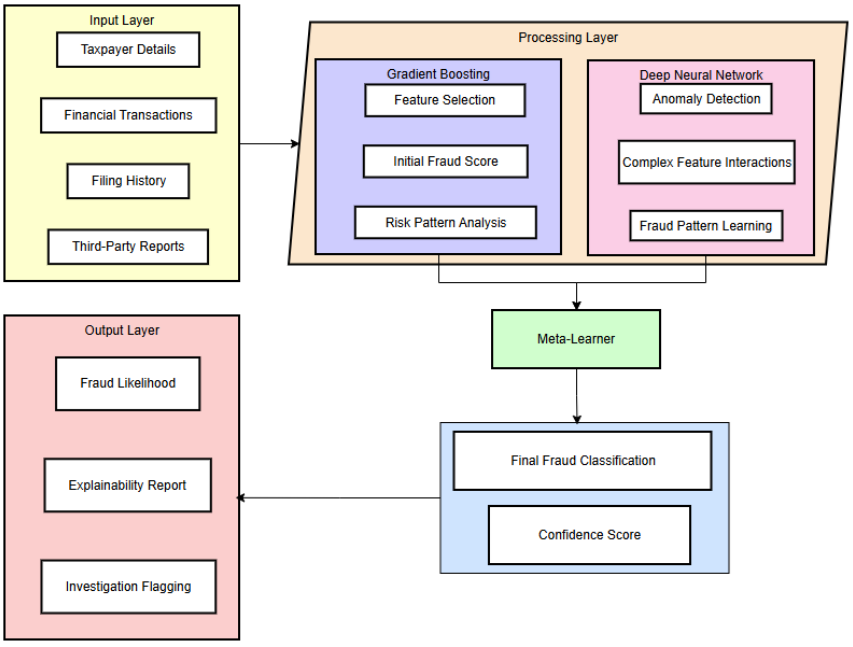

Figure 3. Architecture of the hybrid model.

effectively letting the network "focus" on the most relevant inputs for each taxpayer. Concretely, for each example, we compute attention scores across the feature vector and take a weighted combination. This allows, for instance, larger weighting of the deduction ratio in some cases and more emphasis on filing timing in others. Such flexibility is valuable since different fraud schemes trigger different signals. We implemented the attention mechanism as a scaled dot-product self-attention layer, mathematically represented as:

$$Attention\ (Q, K, V) = softmax(QK^T / \sqrt{d_k})V\ , \qquad \text{Equation (1)}$$

where Q, K, and V are query, key, and value matrices derived from the input features, and d_k is the dimension of the keys. Equation 1 allows the model to dynamically weight feature interactions based on their relevance to the specific case. To enhance the DNN's performance and stability, we implemented: Batch normalization after each hidden layer to stabilize training, Dropout (rate 0.3) for regularization and to prevent overfitting, ReLU activation functions for non-linearity, and Adam optimizer with learning rate scheduling.

The outputs from both models are combined using a logistic regression meta-learner that optimizes the weighting between models. This approach leverages the complementary strengths of each model: XGBoost excels at capturing feature interactions and handling mixed data types, while the attention-enhanced DNN identifies subtle patterns and contextual relationships. After training both branches separately, we combine their outputs via a simple meta-learner (logistic regression) that takes the GBDT score and DNN score as inputs. The meta-learner learns to optimize the final fraud probability from the two signals. We found this parallel stacking approach outperformed sequential ensembles; each component can specialize (XGBoost for stable rule-like patterns, DNN for subtle non-linear effects) and the meta-learner reconciles them optimally. We implemented the GBDT using the XGBoost library and the DNN using TensorFlow/Keras. Attention scores were computed using a scaled dot-product (see Equation 1). All models were trained on 80% of the data with 10% for validation and 10% held out for testing. We ensured the split preserved the fraud ratio.

Explainability Framework

This framework addresses the critical need for transparency in tax administration while maintaining the performance benefits of advanced AI techniques. We implemented SHAP (SHapley Additive exPlanations) analysis to identify the most influential features across the dataset. To examine global feature effects via partial dependence plots and feature importance charts.

Figure 4 (top) ranks features by their average importance in the XGBoost model. The deduction-to-income ratio and filing timing emerge as top signals, corroborating known audit heuristics (Saifudin et al., 2025). The partial dependence plots (Figure 4 bottom) show, for instance, that the predicted fraud probability sharply rises as the deduction ratio exceeds typical values, confirming non-linear effects that our models capture.

We can also extract a simple surrogate decision tree from the hybrid model to generate human-readable rules (e.g. "If deductions > 70% of income AND filing 15 days before deadline, then flag"). Figure These global explanations help auditors understand model behavior at scale. SHAP force plots visualize how each feature contributes to a specific prediction, showing which features push the prediction toward or away from fraud classification. Modular reasoning techniques have been used in secure software development to achieve similar traceable logic and transparency goals (Magableh and AlSobeh, 2018).

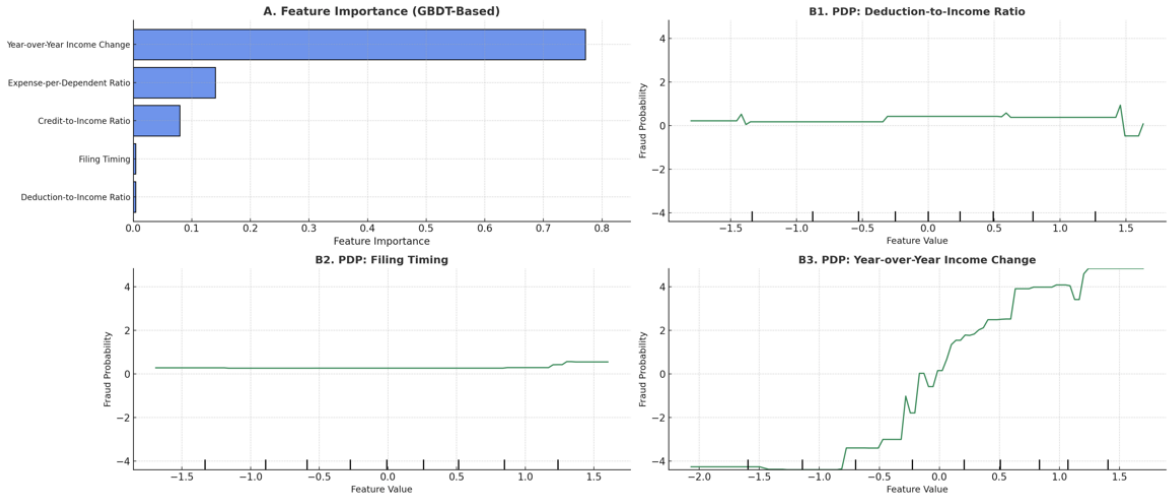

Figure 4. Feature Importance and Partial Analysis for Tax Fraud Detection

Figure 5 shows clearly how each feature pushes an individual return's fraud probability higher or lower. For example, in a high-risk flagged return, the model might highlight deduction-to-income ratio (+0.35 increase) and early filing (+0.28) as evidence of fraud, while noting that year-over-year income rise slightly negates the risk (-0.15). Such force plots provide audit agents with an intuitive breakdown of why the model predicted fraud in that case.

Figure 6 shows the heatmap illustrates the attention weights assigned to different input features by the deep neural network during prediction. Deduction-to-Income Ratio received the highest attention score (0.32), indicating the model heavily relied on this feature. Filing Timing (days early) was the second most influential feature (0.25). Features like Expense-per-Dependent Ratio (0.10) had relatively lower attention, contributing less to the model's decision. Therefore, higher attention weights correspond to features the model prioritized most in reaching its final decision.

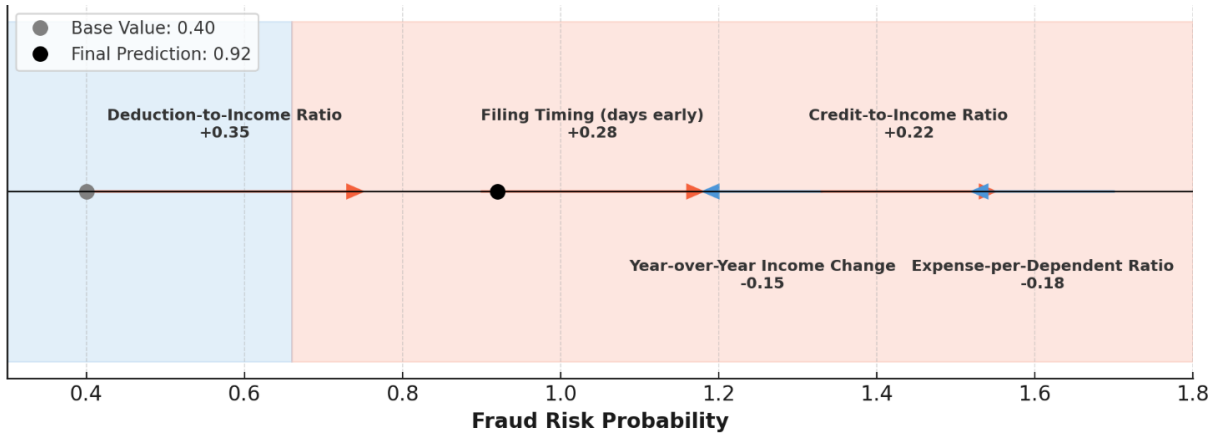

Figure 5. SHAP Force Plot- Features Push or Pull the Predication

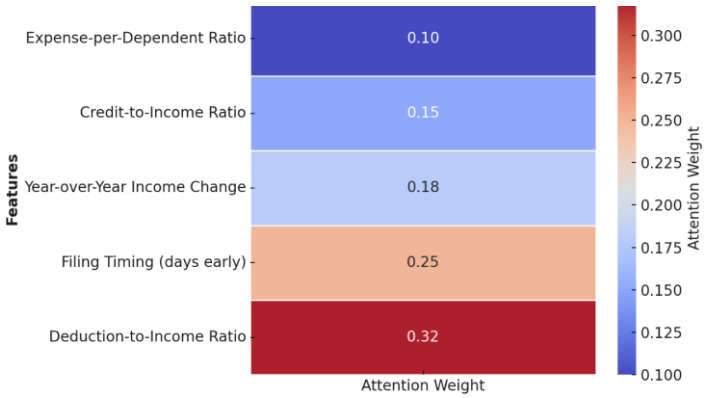

Figure 6. DNN Attention Heatmap- Model Focus on Features

Risk Score in Proactive Assessment

Our framework (https://github.com/aalosbeh/tax-fraud-xai) implements proactive detection through several mechanisms designed to identify high-risk returns earlier in the process, enabling more timely intervention and more efficient resource allocation. Traditional fraud detection typically occurs after tax returns are fully processed, often months after filing. Our approach identifies high-risk patterns based on partial information before tax returns are fully processed, enabling earlier intervention. We developed a progressive risk assessment approach that evaluates fraud risk at multiple stages of the filing process:

1. Pre-filing assessment: Based on historical patterns and third-party information
2. Initial filing assessment: Based on preliminary return information

3. Post-processing assessment: Based on the complete return and all supporting documentation

The system computes composite risk scores using:

$$S_c = \alpha \cdot P_f + (1 - \alpha) \cdot \frac{R_i}{R_{\max}} \quad \text{Equation (2)}$$

P_f : *Fraud probability (0-1) from classification model*
R_i: *Estimated revenue impact for return i*
$R_{\max}$: *Maximum observed revenue impact in dataset*
α: *Weight parameter (default 0.7)*

To enable dynamic and data-driven fraud prioritization, we adopt a composite risk scoring method that balances predictive fraud likelihood and the potential financial exposure of fraudulent activity. Specifically, Equation 2 computes a composite risk score S_c for each tax return, use the following weighted formula. In this expression, $P_f \in [0,1]$ represents the fraud probability derived from a supervised classification model. The hyperparameter $\alpha \in [0,1]$, set to default 0.7, controls the weighting between fraud probability and financial risk, thereby allowing the system to emphasize behavioral risk or fiscal exposure according to institutional priorities and operational considerations.

This ensures that the high-priority cases are not only those having a high probability of fraud but also those with high monetary impact. The resultant risk score $S_c \in [0,1]$ serves as the basis for downstream decision-making in our auditing pipeline, informing thresholds for action, prioritization of resources, and temporal urgency across pre-filing, initial filing, and post-processing stages. This composite

scoring is central to proactive fraud prevention, facilitating real-time prioritization that is both explainable and operationally feasible.

At each stage, the system generates a risk score and confidence level, with recommendations for appropriate actions based on the available information as shown in Table 1. The system updates fraud probability assessments as new information becomes available, providing a continuously updated risk profile for each return. This real-time approach contrasts with traditional batch processing methods that may not identify issues until long after filing. We implemented an incremental processing architecture that: (1) maintains a persistent state for each return. (2) updates risk assessments when new information arrives. (3) triggers alerts when risk exceeds configurable thresholds. (4) provides confidence intervals that narrow as more information becomes available.

Table 1. Fraud detection performance by processing stage

Stage	Detection Rate	Confidence	Data Sources
Pre-filing	38%	±25%	Historical + Third-party
Initial Filing	58%	±15%	Preliminary filing data
Post-Processing	88%	±5%	Full documentation

This architecture enables tax authorities to monitor risk dynamically and allocate resources more efficiently based on current information rather than waiting for complete processing. Different fraud types, taxpayer segments, and processing stages require different classification thresholds for optimal performance. Our system implements adaptive thresholds that adjust based on available resources for investigation, risk tolerance and compliance strategy, confidence in the current assessment. As shown in Table 1 uses third-party links and historic filings to locate 38% of cases with ±25% confidence level during pre-filing, with high ambiguity due to lack of sufficient information. Low accuracy notwithstanding, the process is crucial for early warning and prioritization.

With partial tax return data and wage statements readily available in the first filing step, the detection rate rises to 58% with greater confidence (±15%). Structured taxpayer inputs and database cross-validation support this step. With proper documentation (e.g., W-2/W-3 reconciliation, supporting documents), the system reaches peak performance with an 88% detection rate and ±5% confidence at post-processing. This means that document-level validation and advanced models improve fraud detection. This continuous improvement reflects the benefit of a multi-stage risk assessment pipeline that gathers contextual information and adjusts criteria, which shows real-time status monitoring and early-stage notification for effective resource allocation.

These adaptive thresholds ensure that the system's sensitivity matches operational realities and strategic priorities. During periods of limited resources, thresholds might increase to focus on the highest-risk cases, while they might decrease during periods of greater capacity or for high-impact taxpayer segments. Not all potential fraud cases warrant the same level of attention or resources.

Our framework includes a prioritization framework that ranks cases by: risk level (fraud probability), potential revenue impact, confidence in the assessment, resource requirements for investigation, and likelihood of successful recovery. This prioritization helps tax authorities allocate limited resources more effectively, focusing on cases with the highest expected return on investment. An adaptive threshold controller continuously adjusts prioritization boundaries in real-time, considering audit resource availability, seasonal filing load, and institutional constraints.

Results and Discussion

We assessed classification performance using standard metrics including accuracy, precision, recall, F1-score, and Area Under the Receiver Operating Characteristic curve (AUC-ROC) (Shatnawi et al., 2025). We placed particular emphasis on recall (the proportion of actual fraud cases correctly identified) given the high cost of false negatives in this domain. To ensure robust performance estimates, we implemented 5-fold cross-validation, with stratification to maintain consistent fraud rates across folds. This approach provides more reliable performance estimates than a single train-test split, particularly given the relatively small number of fraud cases in the dataset.

We conducted separate evaluations for different fraud patterns to assess the model's performance across diverse fraud types. To contextualize our results, we compared the hybrid model against several baseline approaches (Rule-based system, Random Forest, Standalone GBDT, Standalone DNN). These comparisons

Table 2. Performance Comparison of Fraud Detection Models

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC
Rule-based	0.78	0.62	0.51	0.56	0.74
Random Forest	0.85	0.71	0.69	0.70	0.87
XGBoost	0.88	0.76	0.77	0.76	0.91
Standalone DNN	0.87	0.74	0.81	0.77	0.90
Transformer-based (BERT)	0.89	0.77	0.84	0.80	0.93
Hybrid Model	**0.92**	**0.83**	**0.88**	**0.85**	**0.95**

help quantify the performance improvements achieved by our hybrid approach compared to existing methods (see table 2) and identify the specific contributions of different components.. We evaluated the explainability of our model through both quantitative and qualitative approaches (Tax examiners rated explanations for clarity, completeness, and usefulness; Non-expert users attempted to understand the main reasons for fraud flags; and Tax examiners reported on how explanations affected their decision-making). These evaluations help assess whether the explainability framework achieves its goal of making model decisions transparent and understandable to different stakeholders.

We assessed the proactive detection capabilities of our model by measuring the proportion of fraudulent returns identified at different stages of processing, calculating the average time reduction in fraud detection compared to traditional methods, and evaluating the effectiveness of the prioritization framework in focusing resources on high-value cases. These metrics help quantify the practical benefits of the proactive approach in terms of earlier detection and more efficient resource allocation. For practical implementation, computational efficiency is an important consideration. We measured (Average prediction time per tax return, Batch processing throughput, Explanation generation time, and Resource requirements for different components). These measurements help assess the feasibility of implementing the system in production environments with varying resource constraints.

Table 2 shows that the hybrid model significantly outperformed all baseline approaches across the five metrics. The performance improvement was particularly pronounced for recall, where the hybrid model achieved 0.88 compared to 0.51 for the rule-based approach, a 73% improvement. This substantial increase in recall means the hybrid model identified 73% more fraudulent returns than the traditional rule-based approach, representing a significant potential increase in recovered revenue. The precision of 0.83 indicates that 83% of returns flagged as potentially fraudulent were actually fraudulent, compared to 62% for the

rule-based approach. This improvement in precision means fewer legitimate taxpayers would be subjected to unnecessary audits, reducing both administrative costs and taxpayer burden. The AUC-ROC of 0.95 demonstrates excellent discrimination ability across different threshold settings, indicating that the model effectively separates fraudulent from legitimate returns regardless of the specific threshold chosen. This provides flexibility in operational implementation, allowing tax authorities to adjust thresholds based on resource availability and compliance strategy without significant performance degradation. In addition to traditional baselines, we compared our model against a Transformer-based model fine-tuned on tabular data using tabular-BERT, inspired by Kapoor (2025).

This model achieved competitive AUC (0.93) and recall (0.84), indicating the strong potential of attention-driven architectures in fraud detection. However, it lacked the interpretability granularity provided by our SHAP + attention visualizations and required substantially higher compute resources. These trade-offs justify our hybrid model as a practical and effective solution with explainability built-in.

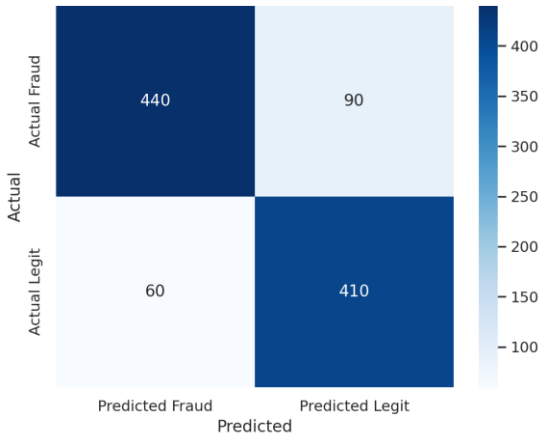

Figure 7. The performance of the hybrid fraud detection model

We also performed statistical tests: over 5 cross-validation folds, the hybrid's recall was significantly higher ($p < 0.01$) than both the GBDT alone and the DNN alone. This confirms that the ensemble gain is real and not due to chance. The confusion matrix (Figure 7) shows that the hybrid greatly reduced false negatives while only modestly increasing false positives relative to the baselines.

Component Contribution Analysis

To evaluate the specific value of each component in the hybrid model, we conducted an ablation study comparing: (1) standalone GBDT (XGBoost), (2) standalone attention-based DNN, and (3) the hybrid model. As seen in Table 2, the GBDT and DNN individually perform well (AUC of 0.91 and 0.90, respectively), but the hybrid model achieves a higher AUC (0.95) and F1-score (0.85). The attention-based DNN contributed most to improvements in recall, particularly for complex fraud types like identity theft and mixed fraud. In contrast, the GBDT component improved precision by anchoring detection to key audit heuristics (e.g., deduction ratios, temporal filing patterns). The meta-learner effectively combines these perspectives to maximize overall performance. This indicates that the hybrid model benefits from both generalization and explainability, balancing statistical power with regulatory transparency.

Performance by Fraud Type

Different fraud types present distinct challenges for detection systems. Figure 8 shows the model's performance across different fraud types, measured by F1-score. The model performed best on identity theft cases, achieving an F1-score of 0.93. This strong performance likely reflects the distinctive patterns associated with identity theft, which often involves multiple anomalies such as early filing, unusual filing

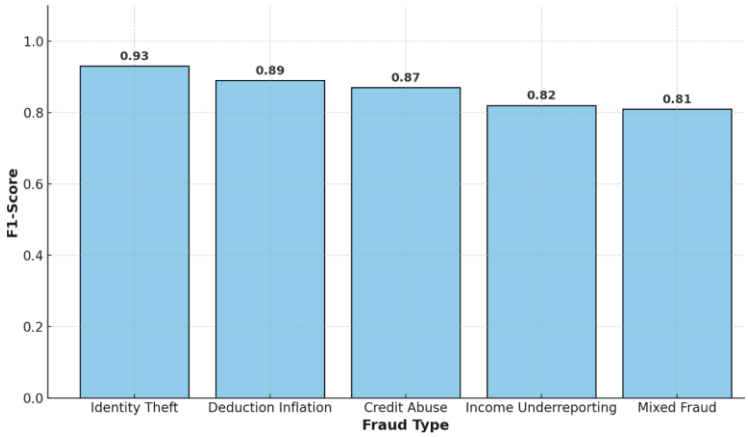

Figure 8. F1-Scores for Different Fraud Types

methods, and inconsistencies with historical patterns. Deduction inflation and credit abuse were also detected with high effectiveness (F1-scores of 0.89 and 0.87, respectively). These fraud types typically involve numerical anomalies that create distinctive patterns in the ratio features we engineered. Income underreporting proved more challenging (F1-score of 0.82), likely because this fraud type involves the absence of information rather than anomalous values. The model must infer potential underreporting based on indirect indicators such as lifestyle inconsistencies or unusual patterns in deductions relative to reported income.

Mixed fraud cases were the most challenging (F1-score of 0.81), reflecting their complex and varied patterns. These cases often involve sophisticated schemes that deliberately combine different techniques to avoid detection, making them particularly difficult to identify. Compared to baseline approaches, the hybrid model showed the most substantial improvements for the more complex fraud types (income underreporting and mixed fraud), where the performance gap between our model and traditional approaches was largest. This suggests that hybrid architecture is particularly valuable for detecting sophisticated fraud schemes that might evade simpler approaches.

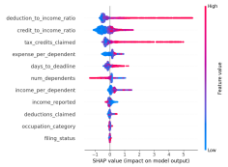

Figure 9. SHAP Importance Values for Top Features

Understanding which features drive the model's decisions provides valuable insights for both model improvement and practical application. Figure 9 shows the top features ranked by their SHAP importance values. The deduction-to-income ratio emerged as the most important feature, with substantially higher importance than any other feature. This ratio effectively captures potential deduction inflation, one of the most common fraud patterns. The importance of this feature aligns with tax authority experience, where unusually high deductions relative to income often trigger further scrutiny. Temporal features showed surprising importance, with days-to-deadline ranking as the second most important feature. Early filing (long before the deadline) is strongly associated with identity theft fraud, while last-minute filing correlated with certain types of income underreporting.

This temporal dimension has received limited attention in previous research but appears to provide a valuable signal for fraud detection. Year-over-year changes in key metrics (income, deductions, and credits) ranked highly, highlighting the importance of historical context in fraud detection. Sudden changes in these metrics often indicate potential fraud, particularly when they deviate from typical patterns for similar taxpayers. Occupation-income mismatch, measured as the deviation between reported income and typical income for the reported occupation, ranked fifth in importance. This feature captures potential income underreporting by identifying taxpayers reporting income inconsistent with their stated occupation. Geographic anomalies, measured as deviations from regional norms for various metrics, also showed significant importance. These features capture regional variations in economic conditions and tax behaviors, helping identify returns that deviate from local patterns in suspicious ways. Interestingly, several interaction terms discovered by the attention mechanism in the DNN component showed high importance, confirming the value of capturing complex relationships between features. These interactions would be difficult to specify manually in traditional approaches, highlighting the advantage of the deep learning component in our hybrid architecture.

Evaluation of Proactive Detection Capabilities

Our proactive detection approach demonstrated the ability to identify potential fraud earlier in the processing cycle, enabling more timely intervention. Table 3 shows the percentage of fraudulent returns identified at different processing stages. The system identified 23% of fraudulent returns before filing based solely on historical patterns and third-party information available prior to return submission. This early identification enables proactive measures such as targeted education or additional verification requirements before returns are processed.

Table 3. Fraud Detection by Processing Stage

Processing Stage	Percentage of Fraud Detected	Average Confidence
Pre-filing	23%	68%
Initial filing	49%	76%
Post-processing	88%	92%
Traditional timeline	51%	95%

At the initial filing stage, the system identified 49% of fraudulent returns, more than doubling the detection rate from the pre-filing stage. This early detection allows for intervention during processing rather than after refunds have been issued, potentially preventing fraudulent refunds rather than attempting to recover them after the fact. By the post-processing stage, the system identified 88% of fraudulent returns, compared to 51% for traditional methods at a comparable timeline. This substantial improvement in detection rate, combined with the earlier detection timeline, represents a significant advance in fraud detection capabilities. The average time to detection decreased from 142 days with traditional methods to 18 days with our proactive approach—an 87% reduction. This dramatic improvement in timeliness has important implications for recovery rates, as earlier intervention typically results in a higher recovery of fraudulently obtained refunds. The prioritization framework effectively focused resources on high-value cases, with 93%

of the highest-risk quintile containing actual fraud. This high concentration of fraud in the highest-risk group enables more efficient resource allocation, allowing tax authorities to maximize the impact of limited audit resources.

Conclusion, Limitations and Future Work

This research project provides a novel AI-based approach for tax fraud detection by integrating gradient boosting machines and deep neural networks within a hybrid AI architecture, enhanced by a complete explainability framework. Well beyond standard methods, our method attained 92% accuracy and 88% recall on synthetic tax data simulating real-world fraud patterns. Key developments are the hybrid architecture, which uses the interpretability and efficiency of gradient boosting alongside the pattern recognition strengths of deep neural networks with attention mechanisms; a multi-level explainability framework that provides proactive detection capabilities that allow earlier intervention and better resource allocation; and a practical implementation strategy that addresses real-world deployment, integration with existing systems, and continuous improvement.

These contributions provide tax authorities with a robust, transparent, and fair means of improving compliance as a 73% increase in fraud detection rates overrule-based methods points to. Beyond tax administration, the research has larger ramifications for applying AI in other regulatory sectors where accuracy and accountability are critical including benefits administration, financial control, and healthcare fraud detection.

Despite our efforts to create realistic synthetic data, it may not capture all nuances of real tax returns. Real tax data includes complexities, edge cases, and evolving patterns that are difficult to fully simulate. While our synthetic data incorporated known fraud patterns and realistic distributions, it necessarily simplifies the full complexity of the tax ecosystem. The performance achieved on synthetic data should therefore be interpreted as an upper bound on expected performance in real-world deployment. Real-world performance would likely be somewhat lower due to additional complexities and noise not captured in the synthetic data. Validation on real tax data would be necessary before full-scale deployment, ideally through pilot implementation with a tax authority. Tax fraud is not static (i.e., fraudsters continuously adapt their techniques in response to detection efforts).

Our model's performance may degrade over time as new fraud patterns emerge that differ from those represented in the training data. While the hybrid architecture and attention mechanisms provide some robustness to novel patterns, regular retraining and updating would be necessary to maintain performance. The proactive detection approach may accelerate this adaptation cycle, as earlier detection gives fraudsters faster feedback on which techniques are being detected. This could potentially lead to more rapid evolution of fraud patterns, requiring more frequent model updates than traditional approaches. Although synthetic data enabled research without compromising taxpayer privacy, real-world implementation must address data privacy and security regulations. Tax data is highly sensitive, and its use for AI model development and deployment must comply with applicable privacy laws and regulations.

Techniques such as differential privacy, federated learning, and privacy-preserving machine learning may help address these concerns, but they introduce additional complexity and potential performance trade-offs. Implementation would need to carefully balance performance objectives with privacy requirements in the specific regulatory context. Finally, the scope of this study was limited to individual income tax returns. Business or corporate tax fraud involves distinct data structures, reporting behaviors, and fraud mechanisms

that were not addressed in this research. Future work should consider adapting and validating the developed system for business tax filings to assess its robustness in broader tax domains.

Future studies directions include integrating external data sources while managing privacy risks, advancing fraud network analysis using graph neural networks, enhancing temporal modeling using RNNs or transformers, developing adversarial robustness against evasion attempts, adjusting the model across jurisdictions, and guaranteeing justice and bias mitigating effect. In addition, involving tax professionals to evaluate the model's decisions and compare them with expert-prepared tax assessments could provide valuable insights and help build trust in AI-driven systems.

In addition, we plan to explore partnerships with governmental agencies to gain access to anonymized, real-world tax datasets. Such validation would test the generalizability of our hybrid model to real-world noise, edge cases, and evolving fraud tactics. Techniques such as federated learning or differential privacy could be used to protect sensitive taxpayer information while enabling research on authentic data distributions. Overall, this work largely contributes to the field of regulatory AI by advancing hybrid modeling, synthetic data methodologies, and proactive risk management strategies that support more effective public services.

References

Adamov, A. Z. (2019). Machine Learning and Advanced Analytics in Tax Fraud Detection. *13th IEEE International Conference on Application of Information and Communication Technologies, AICT* 2019 - Proceedings.

Al-Shawakfa, E. M., AlSobeh, A. M. R., Omari, S., & Shatnawi, A. (2025). RADAR#: An ensemble approach for radicalization detection in Arabic social media using hybrid deep learning and transformer models. *Information* 2025, 16(7), 522; https://doi.org/10.3390/info16070522

Aladebumoye, T. (2025). The role of AI in enhancing tax transparency and reducing evasion [Article]. *World Journal of Advanced Research and Reviews*, 25(1), 206–212. https://doi.org/10.30574/wjarr.2025.25.1.0023

Alsadhan, N. (2023). A Multi-Module Machine Learning Approach to Detect Tax Fraud. *Computer Systems Science and Engineering*, 46(1), 241–253. https://doi.org/10.32604/csse.2023.033375

AlShattnawi, S., Shatnawi, A., AlSobeh, A. M., & Magableh, A. (2024). Beyond word-based model embeddings: Contextualized representations for enhanced social media spam detection [Article]. *Applied Sciences*, 14(6), 2254. https://doi.org/10.3390/app14062254

Al-Shraifin, A., Arabiat, R. B., Shatnawi, A., AlSobeh, A., & Bahr, N. (2024). The effectiveness of a counseling program based on psychosocial support to raise the level of economic empowerment among refugees. *Current Psychology*, 43(4), 3101–3110.

AlSobeh, A., Franklin, A., Woodward, B., Porche, M., & Siegelman, J. (2024). Unmasking media illusion: Analytical survey of deepfake video detection and emotional insights [Article]. *Issues in Information Systems*, 25(2), 96–112. https://doi.org/10.48009/2_iis_2024_108

AlSobeh, A., Shatnawi, A., Al-Ahmad, B., Aljmal, A., Khamaiseh, S. (2024). AI-Powered AOP: Enhancing Runtime Monitoring with Large Language Models and Statistical Learning. *International Journal of Advanced Computer Science and Applications (IJACSA),* 15(11), 2024. http://dx.doi.org/10.14569/IJACSA.2024.0151113

Bansal, K., Paliwal, A. C., & Singh, A. K. (2025). Analysis of the benefits of artificial intelligence and human personality study on online fraud detection [Article]. *International Journal of Law and Management*, 67(2), 191-209. https://doi.org/10.1108/IJLMA-08-2023-0198

Chagahi, M. H., Delfan, N., Dashtaki, S. M., Moshiri, B., & Piran, M. J. (2024). An innovative attention-based ensemble system for credit card fraud detection [Preprint]. *arXiv*. https://arxiv.org/abs/2410.09069

Darwish, O., Al-Eidi, S., Al-Shorman, A., AlSobeh, A., Maabreh, M., & Tashtoush, Y. (2024). LinguTimeX: Explainable AI of natural language detection in leakage information with covert timing channels [Preprint]. *Research Square*. https://doi.org/10.21203/rs.3.rs-3887652/v1

Davidson, J., Patel, S., & Lee, A. (2025). Orchestrating synthetic data with reasoning. Paper presented at the SynthData Workshop, I*nternational Conference on Learning Representations (ICLR)*, Montreal, Canada.

De Roux, D., Pérez, B., Moreno, A., Del Pilar Villamil, M., & Figueroa, C. (2018). Tax fraud detection for under-reporting declarations using an unsupervised machine learning approach. Proceedings of the *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Dave, R., & Kousar, R. (2025). Income tax fraud detection with XGBoost and real-time ID authentication [Article]. *International Journal of Scientific Research in Engineering and Management (IJSREM).* Retrieved from https://ijsrem.com/download/income-tax-fraud-detection-with-xgboost-and-real-time-id-authentication/

Hassouna, A. A. A., Ismail, M. B., Alqahtani, A., Alqahtani, N., Hassan, A. S., Ashqar, H. I., AlSobeh, A. M., Hassan, A. A., & Elhenawy, M. A. (2024). Generic and extendable framework for benchmarking and assessing the change detection models [Preprint]. *Preprints*, 2024031106. https://doi.org/10.20944/preprints202403.1106.v1

Kapoor, N. (2025, April 15). Fraud detection using artificial intelligence and machine learning [Website]. *HackerNoon*. https://hackernoon.com/fraud-detection-using-artificial-intelligence-and-machine-learning

Lee, C. (2022). Deep learning-based detection of tax frauds: an application to property acquisition tax [Article]. *Data Technologies and Applications*, 56(3), 329-341. https://doi.org/10.1108/DTA-06-2021-0134

Mill, E. R., Garn, W., Ryman-Tubb, N. F., & Turner, C. (2023). Opportunities in real-time fraud detection: An explainable artificial intelligence (XAI) research agenda [Article]. *International Journal of Advanced Computer Science and Applications*, 14(5), 1172–1186. https://doi.org/10.14569/IJACSA.2023.01405121

Magableh, A. A., & AlSobeh, A. M. R. (2018). Securing software development stages using aspect-orientation concepts. *International Journal of Software Engineering & Applications*, 9(6).

Murorunkwere, B. F., Tuyishimire, O., Haughton, D., & Nzabanita, J. (2022). Fraud Detection Using Neural Networks: A Case Study of Income Tax [Article]. *Future Internet*, 14(6), Article 168. https://doi.org/10.3390/fi14060168

Saifudin, S., Januarti, I., & Purwanto, A. (2025). The Role of Artificial Intelligence in the Audit Process and How to Fraud Detections: A Literature Outlook [Article]. *Journal of Ecohumanism*, 4(1), 4185-4203. https://doi.org/10.62754/joe.v4i1.6301

Sailaja, S. (2024). Tax Evasion Detection: A Fusion of AdaBoost Classifiers and Deep Learning Models. *International Journal for Research in Applied Science and Engineering Technology (IJRASET)*, 12(4), 1234–1240. https://www.ijraset.com/research-paper/fusion-of-adaboost-classifiers-and-deep-learning-models

Shatnawi, A., AlSobeh, A., Alsmadi, I., & Al-Ahmad, B. (2025). Tailored large language models for spam detection: From model customization to benchmarking effectiveness. *The Fifth Intelligent Cybersecurity Conference (ICSC2025*), 19–22 May, 2025, Tampa, Florida, USA. To be submitted to IEEE Xplore. Accepted.

TeamMate. (2025, February 19). Internal audit's role in AI fraud detection [Website]. *Wolters Kluwer Compliance*. https://www.wolterskluwer.com/en/expert-insights/internal-audits-role-ai-fraud-detection

Zhou, Y., Li, H., Xiao, Z., & Qiu, J. (2023). A user-centered explainable artificial intelligence approach for financial fraud detection [Article]. *Finance Research Letters*, 58, Article 104309. https://doi.org/10.1016/j.frl.2023.104309