Machine Learning and Deep Learning Approaches for Income Tax Fraud Detection: A Comparative Review

Taniya Nath[1], Madhumitha H Athreya[2], and Kushagra Maheshwari[3]
[1,2,3]*Student, Woxsen University*

Abstract—**Income tax fraud poses a persistent and evolving challenge for revenue authorities worldwide, with traditional rule-based detection systems increasingly unable to keep pace with sophisticated evasion techniques. This paper presents a comparative review of machine learning (ML) and deep learning (DL) approaches applied to income tax fraud detection, examining their methodologies, performance characteristics, and practical limitations. The review surveys classification techniques including decision trees, random forests, gradient boosting methods, and neural network-based anomaly detection models, evaluating their reported effectiveness across varying data conditions. Because peer-reviewed, income-tax-specific literature remains comparatively limited, the review further extends, under explicit methodological bounds, to transferable techniques from the larger adjacent literature on financial fraud detection, including graph neural networks for relational fraud modelling and federated learning for privacy-preserving cross-institution training, alongside a methodological case study on evaluation flaws such as data leakage that inflate reported performance in credit-card fraud research. Particular attention is given to challenges such as class imbalance, limited access to public tax datasets, and the trade-off between model accuracy and interpretability, a critical concern for regulatory and compliance applications. The paper concludes by identifying gaps in current research and proposing directions for future work, including hybrid models, explainable AI frameworks, and India-specific opportunities to apply graph-based and federated techniques to the existing GSTN and CBDT analytics infrastructure. This review aims to provide researchers, policymakers, and financial institutions with a consolidated understanding of the current ML/DL landscape in tax fraud detection.**

Index Terms—**Anomaly Detection, Explainable AI, Federated Learning, Fraud Detection, Graph Neural Networks, Income Tax, Machine Learning, Tax Compliance.**

I. INTRODUCTION

Income tax administration depends on voluntary, accurate self-reporting by taxpayers, yet a meaningful share of revenue is lost every year to under-reporting, false deductions, and outright concealment of income. Revenue authorities have historically relied on rule-based flagging and random or risk-weighted audit selection to identify non-compliant filers. These approaches are labour-intensive, slow to adapt to new evasion patterns, and structurally limited by the small fraction of returns that can be manually examined in any given assessment cycle [1].

Machine learning (ML) and deep learning (DL) techniques offer a way to process large volumes of taxpayer data, detect non-obvious patterns of irregularity, and prioritise cases for audit far more efficiently than manual review. Tree-based ensembles, neural networks, and unsupervised anomaly detection methods have each been applied to tax administration problems across multiple countries, with reported results ranging from incremental audit-efficiency gains to substantial revenue recovery [2], [3].

The case for ML/DL adoption in tax administration rests on three practical advantages over legacy systems. First, scale: a trained classifier can score every return in a filing population in minutes, whereas manual risk assessment can only ever cover a sampled subset. Second, pattern discovery: supervised and unsupervised models can surface combinations of features, such as unusual ratios between declared income and asset growth, or anomalous relationships within a taxpayer's transaction network, that static rule sets were never written to catch. Third, adaptability: models can in principle be retrained as new data arrives, narrowing the lag between the emergence of a new evasion technique and the system's ability to detect it, although this advantage is also one of the

field's persistent weak points in practice, discussed further in Section VII.

This review synthesises the existing literature on ML and DL approaches to income tax fraud detection, and, where the tax-specific literature is thin, draws carefully bounded comparisons to the substantially larger body of work on financial fraud detection more broadly, including credit card and corporate financial-statement fraud. It is organised around five central questions: which model families are most commonly applied and why; what performance these models achieve under real or near-real tax data; what evaluation pitfalls commonly inflate reported performance in adjacent fraud domains, and what that implies for tax-specific claims; what structural challenges, principally data scarcity, class imbalance, and interpretability, continue to limit deployment at scale; and what this body of work implies for Indian tax administration specifically, given the relevance of the Goods and Services Tax Network (GSTN) analytics ecosystem and the Central Board of Direct Taxes (CBDT) Project Insight initiative to Indian readers and policymakers.

The remainder of this paper proceeds as follows. Section II positions this review against the most comprehensive existing survey of tax risk detection. Section III surveys model families applied directly to income tax fraud. Section IV presents a comparative performance table. Section V presents a comparative performance table. Section VI examines dataset and data-access constraints. Section VII extends the discussion to transferable techniques from adjacent financial fraud domains, including a methodological cautionary case study. Section VIII discusses evaluation metrics in imbalanced fraud settings. Section IX covers persistent challenges. Section X discusses emerging trends, with a dedicated India-specific subsection. Section XI concludes.

II. RELATED WORK

Zheng et al. [4] provide the most comprehensive existing survey of tax risk detection using data mining techniques, categorising fourteen distinct methods into relationship-based and non-relationship-based approaches. Their survey identifies four recurring bottlenecks across the field: the difficulty of integrating fragmented fiscal knowledge, the unexplainable nature of many model outputs, the high computational cost of advanced algorithms, and a persistent reliance on labelled audit data that is rarely available at scale. This review adopts a similar supervised/unsupervised/hybrid taxonomy but narrows its focus specifically to income tax, rather than tax risk in general, and extends the discussion to research published through 2026.

Earlier domain-adjacent surveys have examined financial fraud detection more broadly, including graph neural network (GNN) approaches to fraud across banking, insurance, and payments [5], [6], and explainable AI and federated learning techniques for privacy-sensitive financial data [7]. These adjacent literatures are treated as methodological background here and are examined in depth, with explicit bounds on what they can and cannot establish about tax fraud specifically, in Section VII.

A further body of related work surveys credit card and corporate financial-statement fraud specifically, rather than financial fraud in general. Systematic reviews of credit card fraud detection techniques, several published in the Journal of Big Data and IEEE Access between 2023 and 2025, catalogue the same supervised, unsupervised, and hybrid taxonomy used throughout this review, but for a domain with far denser, more standardised, and more widely shared benchmark data than income tax fraud has ever had. This density is precisely what makes the methodological critique of Hayat and Magnier [28], discussed in Section VII-C, possible in the first place, since a critique of evaluation rigor requires a large enough body of comparable published results to audit; no equivalent meta-critique yet exists for the income-tax-specific literature, both because the underlying body of work is smaller and because government-partnership datasets are rarely released for independent re-analysis, an absence this review notes as a further limitation in Section XI rather than as a strength of the income-tax literature relative to its adjacent domains.

III. REVIEW METHODOLOGY

This review was conducted as a narrative comparative survey rather than a formal systematic review with predefined PRISMA-style inclusion criteria, reflecting both the breadth of the underlying literature and the comparative, rather than meta-analytic, aim of the paper. Sources were identified through structured

search of major academic databases and publisher platforms, including IEEE Xplore, ScienceDirect/Elsevier, SpringerLink, ACM Digital Library, MDPI journals, PLOS ONE, and arXiv preprints, using combinations of the search terms "income tax fraud," "tax evasion detection," "tax audit machine learning," "VAT fraud detection," "financial fraud graph neural network," and "federated learning fraud detection."

Sources were prioritised for inclusion in Sections III and IV if they reported an application of ML or DL specifically to income tax, value-added tax, or closely related direct-tax fraud detection, with a clear description of the dataset, model, and at least one quantitative performance metric. Sources in Section VII were drawn from the substantially larger adjacent literature on financial fraud detection, specifically graph neural networks and federated learning, and are explicitly bounded to a methodological-transfer role rather than treated as tax-fraud evidence, as stated in Section VII itself. Government and policy sources, such as GSTN technical briefs and CBDT documentation, were included separately and are clearly labelled as non-peer-reviewed throughout, consistent with the distinction drawn in the Recommendations of the underlying research compilation for this review.

Where a single study reported multiple model variants, this review reports the best-performing variant in Table II, with the specific model named, to avoid conflating an author's strongest result with a weaker baseline reported in the same paper. Where reported figures appeared internally inconsistent, for instance an implausibly high AUC alongside a comparatively modest F1-score, this is flagged explicitly in the relevant subsection rather than silently omitted or silently accepted, consistent with the methodological caution raised in Section VII-C.

IV. MACHINE LEARNING AND DEEP LEARNING MODELS APPLIED TO TAX FRAUD DETECTION

Figure 1 summarises the taxonomy used to organise this section: supervised, unsupervised, and hybrid model families, each surveyed in turn below, converging on a shared set of evaluation constraints examined in Section VIII and a field-deployment question examined in Section IX.

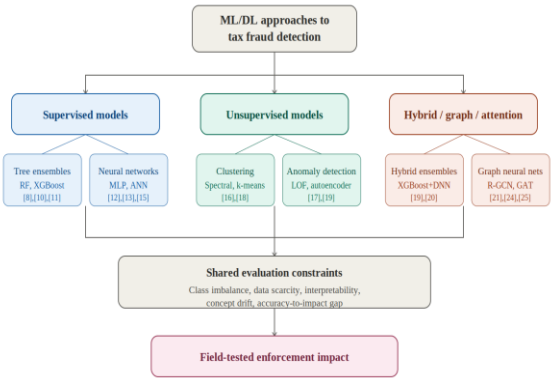

Fig. 1. Taxonomy of ML/DL approaches surveyed in this review.

A. Supervised Tree Ensembles

Random forest and gradient boosting models are the most frequently applied technique in the supervised tax fraud literature, owing to their robustness on tabular data and the interpretability afforded by feature importance scores. Mittal et al. [8] applied random forest classification to Delhi Value Added Tax (VAT) returns to identify fraudulent or "bogus" firms, in what remains the landmark peer-reviewed application of ML to Indian tax data. A subsequent field evaluation of the same model by Barwahwala et al. [9] found that despite strong predictive accuracy, model-driven inspection assignment did not produce a statistically significant increase in enforcement outcomes, an important caution against equating model accuracy with deployment impact.

Baghdasaryan et al. [10] applied XGBoost to the full universe of Armenian business taxpayers and found that supplier-buyer network features improved fraud prediction beyond what historical behavioural data alone could achieve. Zhang et al. [11] compared support vector machines, XGBoost, and random forest on corporate tax compliance data, with random forest achieving the strongest results across both manufacturing and services subsamples.

B. Neural Networks and Deep Learning Models

Neural network approaches have performed strongly in settings where labelled audit outcomes are available. Pérez López et al. [12] applied a multilayer perceptron to Spanish personal income tax (IRPF) data, reporting an efficiency rate of 84.3% and an AUC of 0.918. Murorunkwere et al. [13] applied an artificial neural network to Rwandan income tax data, achieving

92% accuracy, 85% precision, 99% recall, and an AUC of 0.95. A companion study by the same authors [14] compared six algorithms on 15,732 audited Rwandan taxpayers and found that logistic regression performed best after oversampling, while the neural network remained the most robust model overall, underscoring that no single technique dominates across all tax settings.

Rahimikia et al. [15] proposed a hybrid system combining a multilayer perceptron, support vector machine, and logistic regression with Harmony Search optimisation, applied to Iranian corporate tax returns, with the optimised hybrid configuration outperforming any single model on out-of-sample data.

C. Unsupervised and Anomaly Detection Methods

Because labelled fraud data is scarce and subject to sample-selection bias (only audited cases carry reliable labels), unsupervised anomaly detection has emerged as a parallel and in some respects more realistic line of research. De Roux et al. [16] applied spectral clustering and kernel density estimation to under-reported building-tax declarations in Bogotá, Colombia, evaluating results through expert review rather than classification accuracy. Vanhoeyveld et al. [17] applied scalable anomaly detection techniques, including fixed-width anomaly detection and local outlier factor methods, to Belgian VAT declarations across ten sectors, explicitly motivated by both label scarcity and concept drift in evasion behaviour over time.

Savić et al. [18] proposed a hybrid unsupervised outlier detection method (HUNOD) combining domain-weighted k-means clustering with an autoencoder on Serbian personal income tax data, reporting internal validation agreement between 90% and 98%. It is important to note that such figures represent agreement with an internally constructed outlier consensus, not classification accuracy against confirmed fraud labels, and should not be read alongside supervised accuracy figures without this distinction.

D. Hybrid, Graph-Based, and Attention Models

Recent work increasingly combines multiple model families. Alsadhan [19] proposed a four-module framework combining XGBoost, an autoencoder, a behavioural compliance score, and a final ANN/SVM classifier on Saudi VAT filings. AlSobeh et al. [20] combined XGBoost with an attention-based deep neural network and SHAP explainability, achieving 92% accuracy and an AUC of 0.95 on a synthetic dataset of 10,000 tax returns. Xavier et al. [21] incorporated relational graph convolutional network (R-GCN) features into a random forest and neural network pipeline applied to Brazilian corporate tax data spanning approximately 1.6 million companies, reporting that the addition of relational features improved random forest accuracy by approximately eight percentage points over non-relational features alone.

Yang and Chen [29] propose a different style of hybrid architecture, a three-stage stacked ensemble combining a random forest with a deep belief network (DBN), an older but still effective deep generative architecture built from layered restricted Boltzmann machines, applied with class-based limited undersampling to manage imbalance. Their framework introduces a named cost-benefit evaluation metric, RCITA, designed specifically to weigh the administrative cost of an audit against the expected revenue recovered, rather than optimising a generic classification metric such as F1-score in isolation. This is a notable departure from most of the literature surveyed in this section, which optimises statistical performance and leaves the cost-benefit translation to the implementing tax authority; an explicit cost-aware metric of this kind is closer to what a revenue authority such as CBDT or GSTN would actually need to justify a deployment decision, and is a design pattern this review's own recommendations in Section X-B implicitly call for at each proposed implementation stage.

Across the four model families surveyed in this section, a consistent pattern emerges: single-model pipelines, whether a single tree ensemble, a single neural network, or a single clustering method, are increasingly being superseded in the most recent literature by architectures that combine at least two of supervised classification, unsupervised anomaly scoring, relational or graph features, and an explicit explainability or cost-weighting layer. This convergence toward hybrid architectures is itself evidence that no single technique surveyed in Sections IV-A through IV-C is sufficient on its own, a conclusion consistent with the field-evaluation caution raised throughout this review that technique selection

should be driven by the specific operational constraint a tax authority faces, label availability, explainability requirements, or cost sensitivity, rather than by whichever single model reports the highest benchmark accuracy in isolation.

Table IV condenses this section into a practitioner-facing summary, mapping each model family to the operational condition under which the surveyed literature suggests it is the more appropriate starting point, rather than restating the performance figures already given in Table II.

Table I. Model Family Selection Guidance by Operational Condition

If the priority is…	the surveyed literature favours
Labelled audit data exists, interpretability required	Tree ensembles (Sec. IV-A) — built-in feature importance
Labelled data exists, capacity for higher complexity	Neural networks (Sec. IV-B) — strongest raw AUC
Labels scarce or unavailable	Unsupervised anomaly detection (Sec. IV-C)
Entities have known relational structure (e.g. invoices)	Graph-based / R-GCN models (Sec. IV-D, VII-A)
Cost-sensitive deployment decision required	Cost-aware hybrid ensembles, e.g. RCITA-style (Sec. IV-D)
Cross-institution data, strict privacy constraints	Federated learning (Sec. VII-B)

V. COMPARATIVE PERFORMANCE ANALYSIS

Table II summarises reported performance metrics across the principal studies discussed above. Supervised studies that report classification accuracy against confirmed fraud labels are presented separately from unsupervised studies that report internal validation or expert-agreement metrics, since the two are not directly comparable.

Table II. Comparative Performance of Ml/Dl Models in Tax Fraud Detection

Study (Country)	Best Acc.	Model
Pérez López [12] (Spain)	84.3%	MLP (AUC 0.918)
Murorunkwere [13] (Rwanda)	92%	ANN (AUC 0.95)
Murorunkwere [14] (Rwanda)	84.4%	Logistic Reg.
Zhang et al. [11] (China)	92–93.4%	Random Forest
AlSobeh [20] (Synthetic)	92%	XGBoost+Attn DNN
Xavier et al. [21] (Brazil)	>98%	RF / NN / R-GCN
Savić et al. [18]* (Serbia)	90–98%*	HUNOD (k-means+AE)
De Roux [16]* (Colombia)	Expert review*	Spectral Clustering

Unsupervised studies report internal validation or expert-agreement metrics, not classification accuracy against confirmed fraud labels; these are not directly comparable to supervised accuracy figures.

Across the studies surveyed, tree ensembles and neural networks achieve broadly comparable accuracy in the low-to-mid 90% range when sufficient labelled data exists, while unsupervised methods remain the only viable option in jurisdictions where audit labels are unavailable or unreliable. The addition of relational or network-derived features, as in Baghdasaryan et al. [10] and Xavier et al. [21], consistently improves performance over models using individual taxpayer attributes alone, suggesting that supplier-buyer and transactional network data represent an under-exploited signal in many existing pipelines.

VI. DATASETS AND DATA CONSTRAINTS

Public, labelled income tax fraud datasets are essentially nonexistent. Every study reviewed here relies on one of three data sources: confidential government partnerships, synthetically generated data designed to approximate real filing patterns, or proxy datasets from adjacent fraud domains such as credit card transactions, used only to validate modelling methodology rather than tax-specific findings. Table III summarises the government-partnership datasets underlying the studies discussed in Section IV, illustrating both the geographic spread of this research and the typical scale of access researchers have been granted.

Table III. Government-Partnership Datasets in the Surveyed Literature

Jurisdiction	Data Holder	Approx. Scale
Colombia [16]	Bogotá city authority	1,367 declarations
India [8]	Delhi VAT department	Firm-level VAT returns
Armenia [10]	State Revenue Committee	Full taxpayer universe
Rwanda [13],[14]	Rwanda Revenue Authority	15,732 audited taxpayers
Spain [12]	Inst. of Fiscal Studies	Personal income tax sample
Belgium [17]	Federal Public Svc. Finance	VAT, 10 sectors
Serbia [18]	Tax Administration	179,489 entities, 224 features
Saudi Arabia [19]	ZATCA	VAT filings, 2018–2022
Brazil [21]	Receita Federal / Goiás	~1.6 million companies

Synthetic data generation has become an increasingly common workaround for this scarcity. AlSobeh et al. [20] constructed a synthetic dataset of 10,000 tax returns with a fixed 10% fraud rate to evaluate their hybrid XGBoost-attention model, a design choice that allows controlled experimentation but cannot capture the irregular, evolving fraud rates and feature correlations present in real filing populations. Yang and Chen [29] go further, releasing a named synthetic benchmark dataset, TXFCN, alongside their stacked random-forest/deep-belief-network ensemble, explicitly intended to give other researchers a shared, reusable benchmark in the absence of any public real-world equivalent. Generative approaches, including generative adversarial networks (GANs) combined with SMOTE-style oversampling, have also been explored in adjacent e-commerce tax-evasion research as a way to expand minority-class representation without compromising the privacy of real taxpayer records.

This scarcity has two consequences for the field. First, it explains the prevalence of unsupervised and semi-supervised methods, since acquiring confirmed fraud labels typically requires a completed audit, a slow and resource-constrained process, and the resulting labelled sample is itself biased toward whichever taxpayers were selected for audit under the prior detection regime, a subtle but consequential form of sample-selection bias that affects every supervised study in Section IV to some degree. Second, it limits the generalisability of any single study's results, since tax law, filing behaviour, and enforcement capacity differ substantially across jurisdictions, a point emphasised explicitly by Mittal et al. [8] in the Indian context. A model trained on Rwandan income tax data, for instance, encodes assumptions about filing categories, deduction structures, and audit selection criteria that may not transfer cleanly even to a neighbouring East African jurisdiction, let alone to India's substantially larger and more heterogeneous taxpayer base.

VII. TRANSFERABLE TECHNIQUES FROM ADJACENT FINANCIAL FRAUD DOMAINS

Because peer-reviewed, income-tax-specific literature remains comparatively sparse, this section draws bounded comparisons to two larger and more mature adjacent literatures, graph neural network (GNN) fraud detection and credit-card fraud detection, on the explicit understanding that these are methodological parallels rather than tax-fraud evidence. Findings here are presented as technique transfer potential, not as claims about income tax fraud performance.

A. Graph Neural Networks for Relational Fraud Detection

Tax evasion frequently involves coordinated behaviour among related entities, shell companies, falsified supplier chains, and circular invoicing rings, that purely tabular, per-taxpayer models struggle to represent. Graph neural networks address this by modelling taxpayers, transactions, and counterparties as nodes and edges in a relational graph, allowing fraud signals to propagate across connected entities rather than being evaluated in isolation. Pourhabibi et al. [24] provide an early systematic review of graph-based anomaly detection approaches to fraud, establishing the foundational case for relational modelling over purely feature-based classification.

More recent benchmarking work by Cheng et al. [25] compares graph convolutional networks (GCN), graph attention networks (GAT), and specialised architectures such as PC-GNN and GTAN on public fraud benchmark datasets. Reported results show a consistent pattern: graph-based methods outperform feature-based baselines by approximately ten to fifteen AUC points on review-fraud and product-fraud

benchmark datasets, with a simple multilayer perceptron baseline achieving an AUC near 0.70–0.73 against 0.85–0.89 for graph attention and heterophily-aware variants. While these benchmarks are drawn from review and e-commerce fraud rather than tax data, the magnitude of improvement is consistent with, and lends external support to, the relational feature gains already documented directly in the tax literature by Baghdasaryan et al. [10] and Xavier et al. [21].

Architecturally, the field has moved beyond static graph convolution toward representations that capture how fraud networks change over time. Temporal graph neural network variants combine graph convolution at each time step with a recurrent component, typically a gated recurrent unit, so that the model's representation of a taxpayer or entity updates as new transactions and relationships are observed, rather than treating the network as a single fixed snapshot. Transformer-based graph architectures extend this further by allowing attention across the entire graph rather than only a node's immediate neighbours, which directly addresses a structural weakness of standard message-passing GNNs: a fraud scheme that routes transactions through several intermediary shell entities specifically to place several relational "hops" between the beneficiary and the fraudulent activity can evade a neighbourhood-limited model even when the underlying network signal is strong. Edge features, such as transaction amount, timestamp, and type, carry much of the discriminative signal in financial graphs and are incorporated directly in these newer architectures rather than discarded, as they often were in earlier node-only formulations.

Graph construction choices matter as much as model architecture. The financial fraud GNN literature typically represents accounts, entities, and sometimes devices or shared infrastructure as nodes, with edges representing transactions, ownership relationships, or shared attributes, and edge types distinguishing, for instance, an ownership relationship from a transactional one. For a tax-specific application, the natural analogue is to represent individual or corporate taxpayers as nodes and invoice, payment, or shareholding relationships as typed edges, an approach directly supported by the relational feature gains already documented in the tax-specific literature by Baghdasaryan et al. [10] and Xavier et al. [21], and a natural fit for the invoice-level data already collected by GSTN, discussed further in Section X-A.

B. Federated Learning for Privacy-Preserving Fraud Detection

Section VI established that tax data is held in fragmented, jurisdiction-specific silos, and that privacy and confidentiality constraints prevent centralising taxpayer records for model training. Federated learning offers a direct technical response to this constraint: rather than pooling raw data in one location, multiple institutions train a shared model collaboratively, exchanging only model updates rather than underlying records, with each institution's data remaining local throughout [7]. Khan [26], in a framework combining federated learning with explainable AI components, argues that this decentralised approach allows institutions to benefit from collective fraud intelligence while remaining compliant with data protection regulation, since sensitive transaction-level data is never aggregated on a central server.

Applied work in the credit card domain provides a useful proof of concept for the tax setting. A federated learning framework combining class-balancing techniques with a distributed training architecture across simulated bank clients found that the federated approach offered an effective balance between privacy preservation and detection performance, particularly for moderately sized institutional datasets, without requiring any raw data sharing [27]. Subsequent work has extended federated fraud detection to explicitly address heterogeneous data distributions across institutions, a problem directly analogous to the differing taxpayer populations, filing behaviours, and data formats that would arise if Indian direct-tax (CBDT) and indirect-tax (GSTN) data systems were ever to pursue joint federated modelling, as discussed further in Section X.

Federated learning alone exchanges model parameters rather than raw data, but parameter updates can still leak information about the underlying training data under certain attack conditions, which has motivated two complementary privacy-hardening techniques in the wider literature. Homomorphic encryption allows computation to be performed directly on encrypted data without requiring decryption at any point in the training process, so that even the central coordinating server never observes plaintext financial information, at the cost of substantially higher computational overhead that complicates real-time deployment. Differential privacy instead adds carefully calibrated

statistical noise to model updates before they are shared, providing a mathematical guarantee that no individual taxpayer's record can be reliably reconstructed from the aggregated model, at the cost of some reduction in model accuracy proportional to the strength of the privacy guarantee chosen. For a tax administration context, where regulatory compliance and taxpayer confidentiality obligations are typically stricter than in commercial banking, the choice between these techniques, or a combination of both, represents a concrete design decision that would need to precede any federated deployment across CBDT and GSTN systems.

C. A Cautionary Case Study: Methodological Rigor Versus Reported Performance

A critical examination by Hayat and Magnier [28] of the credit-card fraud detection literature offers an important methodological caution that applies directly to how this review interprets reported performance figures throughout. Surveying a substantial sample of published credit-card fraud studies on the widely used European cardholder benchmark dataset, the authors identify four recurring evaluation flaws: data leakage from applying class-balancing techniques such as SMOTE before, rather than after, the train-test split; vagueness in reporting preprocessing order and parameters, which prevents replication; absence of temporal validation despite the inherently time-ordered nature of transaction data; and selective emphasis on recall at the expense of precision.

To demonstrate the scale of the problem, the authors constructed a deliberately minimal classifier, a single-neuron network with no hidden layer, trained on data that had been balanced before splitting, and found it achieved 93.9% recall and 97.6% precision, comparable to or better than many sophisticated published architectures. They conclude that proper evaluation methodology matters more than model complexity, and that a meaningful share of the field's most impressive reported numbers, including several results exceeding 99% accuracy on the same benchmark dataset, are artefacts of data leakage rather than genuine model performance [28].

This finding has direct implications for the present review. It is a further reason, alongside the field-deployment finding of Barwahwala et al. [9] discussed in Section IX, to treat very high reported accuracy figures with appropriate scepticism, whether they originate from the tax literature or from adjacent fraud domains, and to weight studies that report clear preprocessing methodology, temporal validation, and balanced precision-recall trade-offs more heavily than studies reporting only a single, very high headline number.

VIII. EVALUATION METRICS IN IMBALANCED FRAUD DETECTION

Because confirmed fraud cases are a minority class in essentially every dataset discussed in this review, the choice of evaluation metric materially affects how a model's performance should be interpreted, and is worth treating explicitly rather than assuming familiarity. Accuracy, the proportion of all predictions that are correct, is the least informative metric in this setting: in a population where fraud prevalence is below one percent, a trivial classifier that predicts “no fraud” for every case can exceed 99% accuracy while detecting zero fraud [28].

Precision measures the proportion of cases flagged as fraudulent that are genuinely fraudulent, and is the metric most directly tied to wasted audit resources: low precision means auditors spend time investigating cases that turn out to be legitimate. Recall measures the proportion of genuinely fraudulent cases that the model successfully flags, and is the metric most directly tied to missed revenue: low recall means fraudulent filers escape detection entirely. The F1-score, the harmonic mean of precision and recall, is commonly reported as a single balanced figure, but it can obscure a severe imbalance between the two-underlying metrics, as seen in several entries of Table II where precision and recall diverge substantially even though the headline accuracy or F1 figure looks acceptable.

A worked example illustrates why accuracy is misleading in this setting. Consider a filing population of 100,000 taxpayers in which 500 (0.5%) are genuinely fraudulent, broadly consistent with the low fraud-prevalence figures reported across the jurisdictions surveyed in Table III. Table IV compares two hypothetical classifiers evaluated on this population: a trivial classifier that flags nobody, and a realistic classifier modelled loosely on the mid-range performance reported in Section IV, with 80% recall and 70% precision.

Table IV. Accuracy Vs. Precision/Recall on A 0.5% Fraud-Prevalence Population (N=100,000)

Metric	Trivial classifier	Realistic classifier
Fraud cases flagged	0	400 of 500 (80% recall)
False positives	0	~171 (70% precision)
Accuracy	99.5%	~99.8%
Precision	undefined (0/0)	70%
Recall	0%	80%
Fraud cases missed	500 (all)	100

The trivial classifier achieves higher raw accuracy than the realistic one, 99.5% versus approximately 99.8%, the comparison is close because true negatives dominate the denominator either way, yet the trivial classifier misses every single fraud case while the realistic classifier catches 400 of 500. This is the precise failure mode that the abundance-of-true-negatives problem described above produces, and it is why this review treats accuracy figures reported in isolation, without an accompanying precision or recall figure, with caution throughout Table II.

The area under the receiver operating characteristic curve (AUC-ROC) is widely reported across the studies surveyed in Section IV and Section VII, and is useful for comparing the ranking quality of different models independent of a specific decision threshold. However, AUC-ROC can be overly optimistic under extreme class imbalance, since it accounts for true negatives, which are overwhelmingly easy to classify correctly when fraud is rare. The precision-recall curve is increasingly recommended as the more informative alternative in this setting [28], since it focuses entirely on the trade-off between precision and recall without crediting the model for the trivially large number of correctly identified legitimate cases. For tax administrations deciding how to weight these metrics in practice, the appropriate balance depends on institutional capacity: a revenue authority with limited audit staff should weight precision more heavily to avoid wasting scarce investigative capacity, while one prioritising maximum revenue recovery may reasonably accept lower precision in exchange for higher recall.

IX. CHALLENGES AND LIMITATIONS

Six challenges recur throughout the literature surveyed in Sections IV and VII, each examined briefly below.

A. Class Imbalance and Sample-Selection Bias

Class imbalance is near-universal, since confirmed fraud cases typically represent a small minority of filings; researchers commonly address this through SMOTE, ADASYN, cost-sensitive learning, or focal loss, as seen in the oversampling approach of Murorunkwere et al. [14]. As established in Section VIII, accuracy alone is a poor metric under severe imbalance, and the cautionary findings of Hayat and Magnier [28] regarding SMOTE applied before train-test splitting in the credit-card literature are an equally relevant warning for any future tax-specific study using similar resampling techniques. A related but distinct issue is sample-selection bias: because labelled fraud cases originate from audited returns, and audit selection itself was historically guided by the very rule-based or risk-weighted criteria that ML models are meant to improve upon, the labelled training data available to researchers is systematically skewed toward whatever fraud patterns the prior detection regime was already capable of finding, potentially under-representing genuinely novel evasion techniques.

B. Data Scarcity and Privacy

Data scarcity and privacy constraints, discussed in Section VI, remain the field's most binding structural limitation, and are the direct motivation for the federated learning techniques discussed in Section VII-B.

C. Interpretability

Interpretability is a recurring concern given the legally reviewable nature of tax enforcement decisions, with Zheng et al. [4] explicitly naming unexplainable model output as one of four core bottlenecks in the field. This concern is not merely technical: an audit decision triggered by a model is typically subject to administrative or judicial review in most jurisdictions, and a tax authority that cannot explain why a particular taxpayer was flagged risks both successful taxpayer appeals and broader erosion of public trust in algorithmic enforcement. This is the underlying motivation for the SHAP- and LIME-based explainability techniques discussed in Section X, and

explains why several of the more recent hybrid models surveyed in Section IV-D build explainability components directly into the architecture rather than treating them as an optional add-on.

D. Concept Drift

Concept drift, the gradual evolution of evasion tactics that degrades static model performance over time, is highlighted explicitly by Vanhoeyveld et al. [17]. A model trained on one assessment year's fraud patterns will, by construction, lag behind taxpayers who adapt their behaviour in response to known detection criteria, which argues for periodic retraining cycles and, as discussed in Section VII-A, for temporal graph architectures that can represent evolving relational structure rather than a single static snapshot of the taxpayer population.

A concrete illustration is instructive. If a tax authority's model relies heavily on a single feature, for instance, an unusually large input tax credit claims relative to declared turnover, evading taxpayers who observe enforcement patterns over successive assessment cycles can learn to split claims across multiple smaller transactions or shell entities specifically to stay under the feature's effective threshold, a behavioural adaptation the original model was never trained to recognise. This is precisely the dynamic that motivates the relational and graph-based features discussed in Section VII-A: a feature built on transaction network structure, rather than any single transaction's magnitude, is harder for an evader to route around without disrupting the underlying business relationships the scheme depends on.

E. Adversarial Taxpayer Behaviour

Adversarial taxpayer behaviour, the active gaming of known detection criteria, parallels adversarial dynamics documented in the broader financial fraud literature [6]. Where a detection model's decision logic becomes known or inferable, whether through published research, leaked criteria, or simple trial and error, sophisticated evaders can structure transactions specifically to fall just outside the model's learned decision boundary, a dynamic that is structurally similar to adversarial example generation in other applied machine learning domains. This risk is a further argument for ensemble and hybrid approaches over single-model pipelines, since an evader would need to simultaneously defeat multiple, structurally different detection mechanisms, and for periodic, undisclosed updates to detection criteria rather than a single static, publicly documented rule set.

F. The Accuracy-to-Impact Gap

A sixth and more fundamental limitation, raised by the field evaluation of Barwahwala et al. [9], is that predictive accuracy does not automatically translate into enforcement impact. Their randomised evaluation of a high-accuracy random forest model found no statistically significant increase in enforcement outcomes when the model was used to guide inspection assignment in practice, a finding with direct implications for how academic performance metrics should be interpreted by tax administrators considering deployment. Plausible explanations the authors and subsequent commentary have offered include downstream bottlenecks in investigative capacity that a better-targeted case list cannot resolve, resistance or inconsistent uptake among the auditors using the model's recommendations, and the possibility that flagged firms simply cease operating or become unreachable before enforcement action can be completed, none of which would be visible in an offline accuracy metric computed purely against historical labels.

The cost-aware RCITA-style metric introduced by Yang and Chen [29] and discussed in Section IV-D is one direct response to this gap, since it forces the evaluation to incorporate the two real-world quantities a revenue authority actually cares about, the administrative cost of conducting an audit and the revenue expected to be recovered if the flagged case is genuinely fraudulent, rather than a generic classification score divorced from either. A model that achieves marginally lower precision but flags cases with substantially higher average recoverable revenue, for instance, larger corporate filers over numerous small individual ones, may represent a better resource allocation for a capacity-constrained tax authority even though it scores lower on F1. None of the purely statistical metrics discussed in Section VIII can surface this distinction on their own, which is precisely why Barwahwala et al.'s [9] field-evaluation discipline and Yang and Chen's [29] cost-aware metric design represent complementary, not competing, responses to the same underlying problem identified throughout this section.

X. EMERGING TRENDS AND FUTURE DIRECTIONS

Four directions characterise the 2022–2026 literature. Explainable AI (XAI) techniques, particularly SHAP and LIME, are increasingly integrated directly into model pipelines rather than applied post hoc, as demonstrated by the attention-heatmap and SHAP integration in AlSobeh et al. [20] and the combined explainable federated learning framework of Khan [26]; this reflects a growing recognition that tax enforcement decisions require legally defensible justification, not just predictive accuracy. Graph neural networks are being applied to model supplier-buyer and invoice networks for detecting shell-company and fake-invoice fraud rings [24], [25], building on the relational feature gains already demonstrated by Xavier et al. [21] and discussed in detail in Section VII-A. Federated learning is emerging as a response to the privacy constraints inherent in cross-institution tax data, allowing model training without centralising raw taxpayer records [7], [26], as discussed in Section VII-B. Finally, hybrid and attention-based architectures that combine tree ensembles with deep learning components, as in Alsadhan [19] and AlSobeh et al. [20], are increasingly favoured over single-model pipelines.

A. Implications for Indian Tax Administration

India's tax administration already operates a substantial AI-driven analytics infrastructure on both the indirect and direct tax sides, predating much of the academic literature surveyed in this review, which makes the connection between research and practice especially relevant for Indian readers and policymakers.

On the indirect tax side, the Goods and Services Tax Network (GSTN) operate the Business Intelligence and Fraud Analytics (BIFA) platform, developed in partnership with Infosys and deployed from 2019, which applies data analytics, ratio analysis, trend analysis, and network analysis across the full national base of GST returns, invoices, and e-way bills to help officers identify potential evasion [22]. A complementary system, the Advanced Analytics in Indirect Taxation (ADVAIT) framework, profiles taxpayers using invoice-level data to flag anomalous filing patterns. The mandatory e-invoicing rollout has progressively expanded the granularity of transaction-level data available to these systems through a steady series of threshold reductions: from companies with turnover above ₹500 crore in October 2020, to above ₹100 crore in January 2021, above ₹20 crore in April 2021, above ₹10 crore in October 2022, and above ₹5 crore from August 2023, with reporting-window rules tightening in parallel, a 30-day invoice-reporting requirement extended down to the ₹10 crore turnover band from April 2025. Each threshold reduction brings a further tier of small and mid-sized businesses, exactly the segment in which shell-company and circular-trading schemes are most common, into the structured, IRN-stamped invoice network. This is precisely the kind of dense, structured, invoice-level data that the graph neural network literature discussed in Section VII-A is best suited to exploit, an opportunity that does not yet appear to have a corresponding peer-reviewed Indian academic study at the time of writing.

On the direct tax side, the Central Board of Direct Taxes (CBDT) operates Project Insight, contracted to L&T Infotech in 2016 and live in phases from 2017, which builds a consolidated profile of each taxpayer through two linked engines: the Income Tax Transaction Analysis Centre (INTRAC), which integrates data from banks, property registries, credit card records, GST payments, and more recently cryptocurrency holdings, and the Compliance Management Centralized Processing Centre (CMCPC), which flags under-reporting and administers the Non-intrusive Usage of Data to Guide and Enable (NUDGE) strategy of automated compliance reminders [23]. The Annual Information Statement (AIS), introduced as a taxpayer-facing extension of this infrastructure, surfaces much of this consolidated financial data directly to filers before submission, which functions as a deterrence mechanism distinct from, but complementary to, the post-filing fraud-detection models surveyed in this review.

Two research directions appear particularly promising and currently under-explored in the peer-reviewed literature specific to India. First, applying graph neural network architectures to the GSTN supplier-buyer invoice network, building on both the existing BIFA/ADVAIT infrastructure and the international relational-modelling evidence reviewed in Section VII-A, to detect circular trading and fake input tax credit claims more systematically than current rule-

based flagging allows. Second, federated learning approaches that could allow CBDT's INTRAC and CMCPC systems to incorporate cross-institution financial signals, for instance from banks or GSTN, without centralising sensitive taxpayer records in a single system, directly addressing the data-fragmentation and privacy constraints described in Section VI using the techniques surveyed in Section VII-B. Both directions would benefit from the field-evaluation discipline modelled by Barwahwala et al. [9], measuring actual enforcement outcomes rather than offline classification accuracy alone.

B. An Indicative Implementation Sequence

Translating the two research directions above into a deployable system is not a single step, and the literature surveyed in this review suggests a natural staged sequence, broadly mirroring how the international studies in Section IV themselves progressed from simple to more sophisticated models. A first stage would establish a supervised baseline, a random forest or gradient-boosted tree model of the kind shown most effective in Mittal et al. [8] and Zhang et al. [11], trained on existing GSTN invoice-level and CBDT INTRAC features, since tree ensembles require comparatively little infrastructure investment and provide built-in feature-importance interpretability that satisfies at least part of the explainability concern raised in Section IX-C. A second stage would layer relational features from the GST supplier-buyer network on top of this baseline, following the demonstrated accuracy gains of Baghdasaryan et al. [10] and Xavier et al. [21], before committing to a full graph neural network architecture. A third stage would pilot federated coordination between GSTN and CBDT on a limited feature set, with the privacy-hardening techniques discussed in Section VII-B chosen based on the specific regulatory and computational constraints applicable to each system, rather than adopted wholesale from a banking-sector implementation. At every stage, the field-evaluation discipline of Barwahwala et al. [9] argues for measuring actual enforcement outcomes, audit conversion rates, revenue recovered, and case-processing time, rather than offline accuracy alone, before progressing to the next stage.

XI. LIMITATIONS OF THIS REVIEW

This review is itself subject to limitations that should inform how its conclusions are used. It is a narrative comparative review rather than a formal systematic review, and, as stated in Section III, does not apply predefined PRISMA-style inclusion and exclusion criteria, meaning the selection of studies, while structure, retains an element of researcher judgment. The income-tax-specific literature surveyed in Section IV draws on studies from nine jurisdictions, none of which is India beyond the single Delhi VAT study of Mittal et al. [8], so the comparative performance figures in Table II should be read as illustrative of methodological approach rather than as a benchmark for what any specific model would achieve on Indian taxpayer data. The transferable-techniques material in Section VII is explicitly bounded to adjacent domains and, per the caution raised in Section VII-C, may itself be subject to the same evaluation flaws it describes in sources this review was unable to independently re-verify.

A further limitation concerns reproducibility. None of the government-partnership datasets summarised in Table III are publicly available for independent re-analysis, which means the comparative figures in Table II cannot be re-verified by a third party in the way that, for instance, the credit-card fraud literature surveyed by Hayat and Magnier [28] could be audited against a shared public benchmark. This is a structural feature of the field rather than a flaw specific to any individual study, since the underlying taxpayer data is properly confidential, but it does mean that this review's comparative table, like the field itself, rests on a degree of trust in each study's self-reported methodology that a more mature, benchmark-driven field would not require. Finally, given the pace of publication in this area, evidenced by several of the sources cited here dating from 2025 and 2026, this review reflects the literature available at the time of writing and should be expected to be supplemented by subsequent work within a relatively short period.

XII. CONCLUSION

This review has surveyed the principal ML and DL approaches applied to income tax fraud detection across supervised, unsupervised, and hybrid model families, and extended the discussion to transferable

techniques, graph neural networks and federated learning, from the larger adjacent literature on financial fraud detection. Tree ensembles and neural networks achieve strong reported accuracy where labelled tax data exists, while unsupervised anomaly detection remains essential in the more common setting where labels are scarce or unavailable. Graph-based relational modelling and federated learning each address a specific structural weakness identified elsewhere in this review, the under-use of network data and the fragmentation of sensitive financial data across institutions, respectively, and both appear under-applied in the income-tax-specific literature relative to their demonstrated value in adjacent fraud domains.

Two cautionary findings recur throughout this review and deserve particular emphasis for practitioners. First, the field evaluation of Barwahwala et al. [9] shows that high predictive accuracy does not guarantee improved enforcement outcomes once a model is deployed in a real administrative setting. Second, the methodological critique of Hayat and Magnier [28] shows that a meaningful share of the most impressive reported performance figures in adjacent fraud domains are artefacts of evaluation flaws, particularly data leakage from improper preprocessing order, rather than genuine model capability. Both findings counsel a degree of scepticism toward headline accuracy figures, in this review's own comparative table as much as in the literature it surveys, and a corresponding emphasis on methodological transparency, temporal validation, and field-tested enforcement impact over offline benchmark performance alone.

Future research, particularly in the Indian context, would benefit from closer integration between academic model development and the analytics infrastructure already operated by GSTN and CBDT, specifically through graph-based modelling of GST invoice networks and federated learning across direct- and indirect-tax data silos, alongside more rigorous field evaluation of deployed systems rather than offline accuracy alone. Taken together, the literature surveyed here suggests that the technical question of which model performs best in isolation is, at this stage of the field's development, less consequential than the institutional question of how a chosen model is integrated into an existing audit workflow, evaluated against real enforcement outcomes, and adapted as both taxpayer behaviour and the underlying data infrastructure continue to evolve.

ACKNOWLEDGMENT

The authors thank Woxsen University for the academic context within which this review was prepared.

Appendix: Glossary Of Acronyms

Acronym	Meaning
ADVAIT	Advanced Analytics in Indirect Taxation (GSTN)
AIS	Annual Information Statement (CBDT)
ANN	Artificial neural network
AUC-ROC	Area under the receiver operating characteristic curve
BIFA	Business Intelligence and Fraud Analytics (GSTN)
CBDT	Central Board of Direct Taxes
CMCPC	Compliance Management Centralized Processing Centre
DBN	Deep belief network
DNN	Deep neural network
GAT	Graph attention network
GCN	Graph convolutional network
GNN	Graph neural network
GSTN	Goods and Services Tax Network
HUNOD	Hybrid Unsupervised Outlier Detection (method name)
INTRAC	Income Tax Transaction Analysis Centre
IRN	Invoice reference number
IRP	Invoice Registration Portal
LIME	Local interpretable model-agnostic explanations
LOF	Local outlier factor
MLP	Multilayer perceptron
NUDGE	Non-intrusive Usage of Data to Guide and Enable
R-GCN	Relational graph convolutional network
SHAP	SHapley Additive exPlanations
SMOTE	Synthetic Minority Over-sampling Technique
SVM	Support vector machine
VAT	Value-added tax
XAI	Explainable artificial intelligence
XGBoost	Extreme Gradient Boosting
ZATCA	Zakat, Tax and Customs Authority (Saudi Arabia)

REFERENCES

T. Hashimzade, G. D. Myles, and B. Tran-Nam, "Tax evasion and the economics of compliance," in *Handbook of Behavioural Economics and Smart Decision-Making*, Cheltenham, U.K.: Edward Elgar Publishing, 2017, ch. 4.

B. F. Murorunkwere, O. Tuyishimire, D. Haughton, and J. Nzabanita, "Fraud detection using neural networks: A case study of income tax," *Future Internet*, vol. 14, no. 6, p. 168, 2022.

S. Mittal, O. Reich, and A. Mahajan, "Who is bogus? Using one-sided labels to identify fraudulent firms from tax returns," in *Proc. 1st ACM SIGCAS Conf. Computing and Sustainable Societies (COMPASS)*, 2018.

Q. Zheng, Y. Xu, H. Liu, B. Shi, J. Wang, and B. Dong, "A survey of tax risk detection using data mining techniques," *Engineering*, vol. 34, no. 3, pp. 43–59, 2024.

S. Motie and B. Raahemi, "Financial fraud detection using graph neural networks: A systematic review," *Expert Systems with Applications*, vol. 240, p. 122156, 2024.

D. Cheng, Y. Zou, S. Xiang, *et al*., "Graph neural networks for financial fraud detection: A review," *Frontiers of Computer Science*, 2024.

T. Awosika, R. M. Shukla, and B. Pranggono, "Transparency and privacy: The role of explainable AI and federated learning in financial fraud detection," *IEEE Access*, vol. 12, pp. 64551–64560, 2024.

S. Mittal, O. Reich, and A. Mahajan, "Who is bogus? Using one-sided labels to identify fraudulent firms from tax returns," in *Proc. ACM SIGCAS Conf. Computing and Sustainable Societies*, 2018.

T. Barwahwala, A. Mahajan, S. Mittal, and O. Reich, "Is model accuracy enough? A field evaluation of a machine learning model to catch bogus firms," *NBER Working Paper*, no. w32705, 2024.

[10] V. Baghdasaryan, H. Davtyan, A. Sarikyan, and Z. Navasardyan, "Improving tax audit efficiency using machine learning: The role of taxpayer's network data in fraud detection," *Applied Artificial Intelligence*, vol. 36, no. 1, 2022.

R. Zhang *et al*., "Predictive modeling of tax compliance risks: A comparative study of machine learning approaches," *PLOS ONE*, vol. 20, 2025.

C. Pérez López, M. J. Delgado Rodríguez, and S. de Lucas Santos, "Tax fraud detection through neural networks: An application using a sample of personal income taxpayers," *Future Internet*, vol. 11, no. 4, p. 86, 2019.

B. F. Murorunkwere, D. Haughton, J. Nzabanita, F. Kipkogei, and I. Kabano, "Predicting tax fraud using supervised machine learning approach," *African Journal of Science, Technology, Innovation and Development*, 2023.

B. F. Murorunkwere, O. Tuyishimire, D. Haughton, and J. Nzabanita, "Fraud detection using neural networks: A case study of income tax," *Future Internet*, vol. 14, no. 6, p. 168, 2022.

E. Rahimikia, S. Mohammadi, T. Rahmani, and M. Ghazanfari, "Detecting corporate tax evasion using a hybrid intelligent system: A case study of Iran," *International Journal of Accounting Information Systems*, vol. 25, pp. 1–17, 2017.

D. De Roux, B. Pérez, A. Moreno, M. del P. Villamil, and C. Figueroa, "Tax fraud detection for under-reporting declarations using an unsupervised machine learning approach," in *Proc. 24th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD)*, 2018, pp. 215–222.

J. Vanhoeyveld, D. Martens, and B. Peeters, "Value-added tax fraud detection with scalable anomaly detection techniques," *Applied Soft Computing*, vol. 86, p. 105895, 2020.

M. Savić, J. Atanasijević, D. Jakovetić, and N. Krejćić, "Tax evasion risk management using a hybrid unsupervised outlier detection method," *Expert Systems with Applications*, vol. 193, p. 116409, 2022.

N. Alsadhan, "A multi-module machine learning approach to detect tax fraud," *Computer Systems Science and Engineering*, vol. 46, no. 1, pp. 241–253, 2023.

A. R. AlSobeh, M. F. A. El Rob, K. Rouibah, and A. Shatnawi, "Proactive detection of tax fraud using explainable AI techniques: A hybrid approach," *Issues in Information Systems*, vol. 26, no. 3, pp. 255–274, 2025.

A. R. Xavier *et al*., "Tax evasion identification using open data and artificial intelligence," *Revista de Administração Pública*, vol. 56, no. 3, pp. 426–440, 2022.

Goods and Services Tax Network, "Business Intelligence and Fraud Analytics (BIFA) and ADVAIT Analytics Framework," *GSTN Technical Brief*, 2021.

Central Board of Direct Taxes, "Project Insight: INTRAC and CMCPC Compliance Management Framework," Ministry of Finance, Government of India, 2019.

T. Pourhabibi, K.-L. Ong, B. H. Kam, and Y. L. Boo, "Fraud detection: A systematic literature review of graph-based anomaly detection approaches," *Decision Support Systems*, vol. 133, p. 113303, 2020.

D. Cheng, Y. Zou, S. Xiang, *et al*., "Graph neural networks for financial fraud detection: A review," *Frontiers of Computer Science*, 2024.

M. A. Khan, "Secure and transparent banking: Explainable AI-driven federated learning model for financial fraud detection," *Journal of Risk and Financial Management*, vol. 18, no. 4, p. 179, 2025.

M. A. Salam, K. M. Fouad, D. L. Elbably, and S. M. Elsayed, "Federated learning model for credit card fraud detection with data balancing techniques," *Neural Computing and Applications*, vol. 36, no. 11, pp. 6231–6256, 2024.

K. Hayat and B. Magnier, "Data leakage and deceptive performance: A critical examination of credit card fraud detection methodologies," *arXiv preprint* arXiv:2506.02703, 2025.

K. Yang and X. Chen, "Random Forest deep ensemble for tax audit case selection," *Applied Sciences*, vol. 16, no. 10, p. 4682, 2026