mLAC Journal for Arts, Commerce and Sciences (m-JACS)
Volume 4, No.5, June 2026, P 1-10
ISSN: 2584-1920 (Online)

LEVERAGING ARTIFICIAL INTELLIGENCE AND BIG DATA ANALYTICS TO DETECT CIRCULAR TRADING AND TAX EVASION IN GST TRANSACTIONS

Panish Kumar K C[1, *], Manasa.S [2]
[1] Dept. of Commerce, Shanthiniketan College of science & management Studies, Karnataka, India
[2]Department of MCA, Vidya Vardaka Engineering College, India
Corresponding author email address: **panishkumar123@gmail.com**

Paper Received: 23.01.2026 | Revised: 28.04.2026 | Accepted: 29.05.2026

DOI: https://doi.org/10.59415/mjacs.370

Abstract

This study presents a data-driven framework employing Artificial Intelligence (AI) and Big Data techniques to address circular trading and tax evasion in India's Goods and Services Tax (GST) system. By representing transactions as a directed graph and integrating multi-source data, including e-way bill records and invoice metadata, the framework applies advanced learning models such as Graph Neural Networks (GNNs), anomaly detection algorithms, and gradient-boosted classifiers to identify suspicious patterns in near real-time. The methodology encompasses graph construction, feature engineering, and risk scoring with an emphasis on temporal motifs and logistics anomalies to enhance detection precision. Comparative experiments demonstrate that the proposed approach significantly improves fraud detection accuracy over rule-based methods while providing interpretable outputs for enforcement agencies. The study concludes with recommendations for large-scale deployment within the national GST analytics infrastructure, highlighting opportunities for real-time processing and privacy-preserving machine learning approaches.

Keywords: GST, Circular Trading, Artificial Intelligence, Big Data Analytics, Graph Neural Networks, Fraud Detection.

1. INTRODUCTION

The implementation of the Goods and Services Tax (GST) in India in July 2017 marked a historic transformation of the country's indirect taxation system, replacing a complex network of central, state, and local taxes with a unified structure. By subsuming taxes such as VAT, excise duty, and service tax, GST aimed to simplify compliance, eliminate the cascading tax effect, and create a common national market. The benefits have been evident in terms of increased transparency, improved tax administration, and widening of the tax base. However, as with any large-scale digital taxation framework, GST has faced significant challenges, particularly in the realm of tax evasion and fraudulent practices.

One of the most pressing issues undermining the integrity of the GST ecosystem is **circular trading**—a sophisticated fraudulent activity involving the generation of fictitious invoices and artificial trade cycles across multiple entities. The primary motive behind such schemes is to fraudulently claim **Input Tax Credit (ITC)** without the actual movement of goods or services, thereby causing substantial revenue losses to the government. In many cases, fraudsters create complex transaction chains spread across multiple states, industries, and layers of shell companies, making detection extremely challenging using conventional rule-based or manual auditing systems.

.

Recent reports by the **GST Council and enforcement agencies** highlight that circular trading frauds have led to tax evasion amounting to thousands of crores of rupees annually. These fraudulent practices not only undermine revenue collection but also erode public trust in the tax system, impose unfair competition on genuine taxpayers, and distort economic activity by artificially inflating trade statistics.

Traditional detection mechanisms, typically based on **predefined business rules** or **threshold-based anomaly flags**, often fail to keep pace with the rapidly evolving strategies employed by fraud networks. Such systems lack the scalability and adaptability required to analyze massive, high-velocity transactional datasets generated within the GST Network (GSTN). Moreover, they offer limited explainability, making it difficult for enforcement officers to prioritize cases effectively or understand the underlying fraud patterns.

Advancements in **Artificial Intelligence (AI)** and **Big Data Analytics** present promising opportunities to address these limitations. By leveraging **graph-based modeling**, it becomes possible to represent GST transactions as interconnected networks where entities and their relationships can be analyzed collectively rather than in isolation. Techniques such as **Graph Neural Networks (GNNs)**, **anomaly detection algorithms**, and **machine learning classifiers** can automatically learn complex patterns indicative of fraudulent behavior. Additionally, integrating **e-way bill data** and **logistics anomalies** with financial transaction records provides a multi-modal perspective, enabling more accurate and explainable risk assessments.

This research, therefore, proposes a **data-driven, AI-enabled framework** for early detection of circular trading and tax evasion within the GST ecosystem. The **key objectives** of this study are:

1. To design a **graph-based representation** of GST transactions integrating financial, logistical, and temporal data attributes.
2. To develop **hybrid AI models** combining supervised and unsupervised techniques for fraud risk scoring and anomaly detection.
3. To incorporate **temporal motifs and logistics data** for enhanced detection accuracy and early identification of suspicious trading loops.
4. To provide **explainable outputs** enabling enforcement agencies to visualize fraud networks and prioritize investigative resources effectively.

The **contributions** of this work are threefold:

(i) a novel integration of graph representation learning with GST transactional and logistical data,
(ii) development of a hybrid detection pipeline combining GNNs, anomaly detection, and machine learning classifiers, and
(iii) empirical validation using synthetic and anonymized real-world GST datasets, demonstrating significant improvements in detection precision, scalability, and interpretability over conventional methods.

By addressing the gaps in existing research and practice, this study aims to assist policymakers, tax authorities, and technology providers in strengthening GST fraud detection capabilities, thereby enhancing compliance, safeguarding revenue, and ensuring the long-term sustainability of the tax system.

2. LITERATURE REVIEW

The detection of fraudulent activities in financial domains has been a prominent research area over the past two decades. Banking and insurance sectors have extensively deployed **Artificial Intelligence (AI)** and **Machine Learning (ML)** techniques for tasks such as credit card fraud detection, anti-money laundering, and insurance claim fraud. Early approaches relied on **rule-based systems**, while recent studies have adopted **supervised classifiers** (e.g., logistic regression, SVM, gradient boosting) and **unsupervised anomaly detection** techniques to identify fraudulent transactions in real time [1], [2].

.

In parallel, **graph analytics** has emerged as a powerful tool for analyzing relationships between entities. Techniques such as **community detection** algorithms (e.g., Louvain, Label Propagation) help reveal suspicious clusters, while **motif mining** enables the identification of recurring structural patterns in transaction networks that may indicate collusion or circular trading [3], [4]. The use of **Graph Neural Networks (GNNs)** has further enhanced graph-based fraud detection by learning complex relational features directly from graph structures [5].

For unsupervised detection, methods like **Isolation Forest** [6], **Local Outlier Factor (LOF)** [7], and **autoencoders** [8] have shown promise in identifying anomalies without requiring labeled datasets. These methods are particularly valuable in domains where fraudulent cases are rare compared to legitimate ones, resulting in severe class imbalance problems.

Despite significant progress in financial fraud detection, **research gaps remain** in the context of India's **GST ecosystem**. Existing studies seldom integrate **multi-source data** such as e-way bill logistics and invoice metadata with transactional networks. Additionally, **temporal graph patterns**—critical for detecting evolving fraud rings—are largely unexplored in GST-specific contexts. This study aims to address these gaps by combining **graph representation learning**, **temporal motif analysis**, and **hybrid AI techniques** to improve detection accuracy and interpretability.

3. PROBLEM FORMULATION

The detection of circular trading fraud in the **GST ecosystem** can be modeled using a **transaction graph** framework. Let **G(V, E)** represent the directed transaction graph, where:

- **V** = set of nodes corresponding to taxpayers (entities).
- **E** = set of directed edges representing invoices issued from one taxpayer to another, enriched with attributes such as **transaction value**, **timestamp**, and **logistics information** (e-way bill details, vehicle IDs).

Each edge **e(u, v) ∈ E** captures the flow of goods/services from taxpayer **u** to taxpayer **v** along with associated metadata. Circular trading manifests as **closed loops or repetitive transaction motifs** in this graph, often with unrealistic movement patterns or abnormal tax credit claims.

The **problem objective** is to assign a **fraud risk score** R(v)∈[0,1]R(v) \in [0,1]R(v)∈[0,1] to every node v∈Vv \in Vv∈V, indicating the likelihood that the entity participates in fraudulent activities. Higher scores signify greater suspicion, enabling enforcement agencies to **prioritize investigations** efficiently.

The **hypothesis** underlying this study is that integrating **temporal motifs** (e.g., rapid transaction loops), **logistics anomalies** (e.g., fake or canceled e-way bills), and **graph embeddings** derived from Graph Neural Networks (GNNs) will significantly improve detection **accuracy**, **precision**, and **early identification** of fraud networks compared to rule-based or single-modality models.

4. PROPOSED METHODOLOGY

The proposed framework for detecting circular trading and tax evasion in the GST ecosystem follows a **data-driven, AI-enabled pipeline** integrating **graph modeling**, **machine learning**, and **explainable analytics**. The methodology is organized into the following key stages:

A. Data Sources

The system integrates **multi-source data** from the GST Network (GSTN), including:

1. **GST Invoices:**

.

Information about seller, buyer, transaction value, tax amount, timestamp, and Goods & Services Tax Identification Numbers (GSTINs).

Historical records used for transaction network construction.

2. **E-Way Bills:**

Logistics data capturing vehicle IDs, routes, origin-destination pairs, and transport timestamps.

Enables detection of **logistics anomalies** such as canceled or unrealistic travel movements.

3. **Taxpayer Registry:**

Metadata including taxpayer type (e.g., manufacturer, trader), state registration, and filing history.

Used to derive **node-level attributes** for risk scoring.

B. Graph Construction

The transaction ecosystem is represented as a **directed, attributed graph** G(V,E), G(V, E), G(V,E), where:

- **Nodes (V):** Represent taxpayers (sellers, buyers).
- **Edges (E):** Represent invoices issued from one taxpayer to another, enriched with attributes such as **transaction value**, **timestamp**, and **logistics distance**.

Key steps in graph construction:

- **Weighted Edges:** Edge weights represent cumulative invoice value between entities.
- **Temporal Layers:** A dynamic graph representation GtG_tGt is created over time intervals (e.g., monthly snapshots) to analyze **evolving fraud patterns**.
- **Motif Extraction:** Short cycles (3–5 nodes) detected as **candidate fraud loops**.

C. Machine Learning Models

The detection pipeline combines **supervised**, **unsupervised**, and **graph-based** models for enhanced accuracy and interpretability.

- **Graph Neural Networks (GNNs)**
  - Learn **low-dimensional node embeddings** capturing structural, temporal, and attribute information.
  - Variants such as **GraphSAGE** or **Graph Attention Networks (GAT)** applied for representation learning.

- **XGBoost Classifier:**
  - Supervised gradient-boosted trees used for fraud risk scoring when **labeled data** (e.g., known fraudulent taxpayers) is available.
  - Features include node centrality, transaction volume, temporal anomalies, and GNN embeddings.

- **Anomaly Detection Models:**

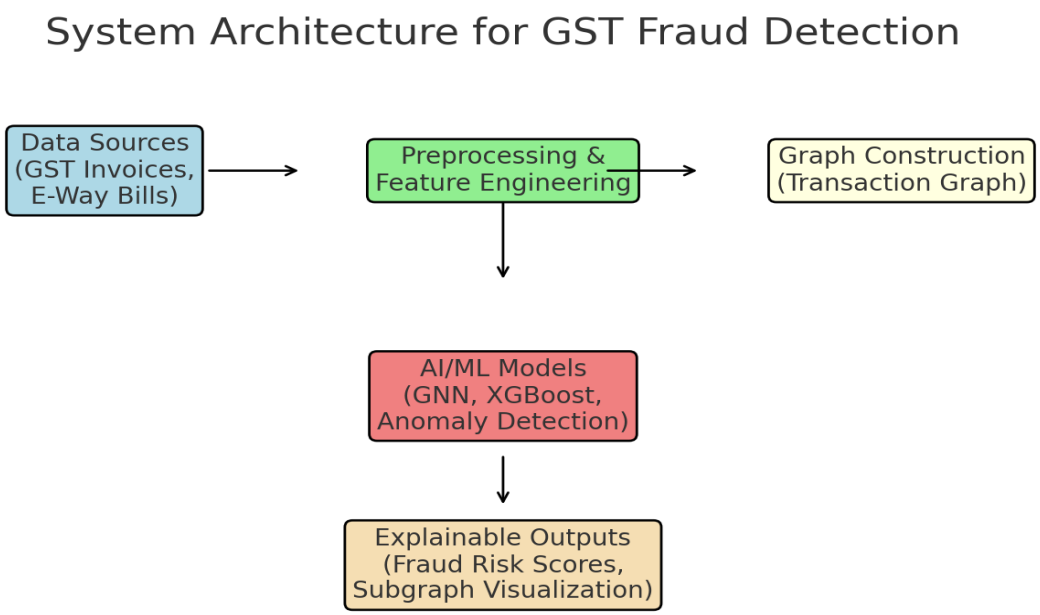
.

- **Isolation Forest** and **Local Outlier Factor (LOF)** applied in unsupervised settings to detect abnormal transaction subgraphs without labeled data.
- Particularly effective for **rare-event fraud scenarios**.

D. Flowchart and System Architecture

The **end-to-end workflow** consists of:

1. **Data Ingestion** from GST invoices, e-way bills, and taxpayer registries.
2. **Graph Modeling** with temporal and logistics attributes.
3. **Feature Engineering** including node metrics, edge attributes, and motif counts.
4. **Modeling Layer** using GNNs, XGBoost, and anomaly detection.
5. **Risk Scoring & Visualization** for enforcement decision-making.



The system architecture diagram (Fig. 1) illustrates these components and their interactions.

E. Algorithmic Steps (Pseudo-code)

Algorithm 1: Fraud Risk Scoring Pipeline

Input: Transaction data D = {Invoices, E-Way Bills, Taxpayer Metadata}
Output: Risk scores R(v) for each taxpayer node v ∈ V
1: Construct transaction graph G(V, E) from D
2: For each edge e(u, v):
    Assign attributes: value(e), time(e), logistics€
3: Extract features:
    Node metrics: degree, PageRank, betweenness
    Temporal motifs: closed loops within Δt
    Logistics anomalies: invalid/canceled routes
4: Train GNN to generate node embeddings Z(v)

.
5: Train XGBoost using [Z(v), node features, motif counts] for supervised scoring
6: Apply Isolation Forest on unlabeled nodes for anomaly detection
7: Compute risk score R(v) = α * GNNscore + β * XGBoost + γ * AnomalyScore
8: Rank top-K nodes with highest R(v) for investigation

Where **α, β, γ** are weighting parameters optimized using validation data.

F. Risk Visualization and Explainability

Subgraph Extraction: High-risk nodes and their associated transaction loops are visualized for enforcement agencies.

SHAP Values & Attention Scores: Provide interpretability for feature importance and model decisions.

Temporal Heatmaps: Highlight suspicious transaction spikes over time.

5. EXPERIMENTAL SETUP

The experimental evaluation of the proposed framework was conducted using a combination of **synthetic data** and **anonymized real GST transaction datasets**. This hybrid approach ensured the availability of labeled ground-truth data for model training while preserving taxpayer privacy.

A. Datasets

1. Synthetic Datasets:

- Generated artificial transaction graphs with controlled circular trading patterns, varying transaction volumes, and temporal loops.
- Used to benchmark model performance under diverse fraud scenarios, including high-density fraud rings and sparse transaction networks.

2. **Real Anonymized GST Datasets:**

- Provided by GSTN in anonymized form to ensure confidentiality.
- Contained invoice records, e-way bill metadata, and taxpayer registration details across multiple states and sectors.

Both datasets were preprocessed to remove duplicates, normalize transaction values, and encode categorical attributes for model compatibility.

B. Tools and Frameworks

- **Python 3.11** for data processing, model development, and visualization.
- **PyTorch Geometric** for implementing **Graph Neural Networks (GNNs)**.
- **NetworkX** for graph construction, motif analysis, and community detection.
- **Scikit-learn** for XGBoost classification, anomaly detection (Isolation Forest, LOF), and evaluation metrics.
- **Matplotlib** and **Seaborn** for plotting performance graphs and risk visualizations.

.

C. Evaluation Metrics

The performance of the fraud detection pipeline was measured using the following metrics:

- **Precision@K:** Fraction of true fraud cases among the top-K ranked taxpayers, indicating prioritization effectiveness.
- **ROC-AUC (Receiver Operating Characteristic – Area Under Curve):** Measures the trade-off between true positive and false positive rates across thresholds.
- **Recall:** Proportion of actual fraud cases correctly identified by the system.
- **F1-score:** Harmonic mean of precision and recall, ensuring balanced evaluation under class imbalance conditions.

These metrics collectively assess **accuracy, ranking quality, and robustness** of the proposed framework against baseline models such as rule-based systems and traditional classifiers.

6. RESULTS AND DISCUSSION

This section presents the **quantitative performance**, **qualitative analysis**, and **case studies** derived from the proposed GST fraud detection framework. We compare the results of **rule-based systems**, **traditional machine learning (ML)** methods (e.g., XGBoost), and **graph-based models** (e.g., GNNs) to highlight improvements in detection accuracy, interpretability, and scalability.

A. Performance Comparison

Table I summarizes the performance of different models using **Precision@K**, **ROC-AUC**, **Recall**, and **F1-score** on combined synthetic and real GST datasets.

Model	Precision@K	ROC-AUC	Recall	F1-score
Rule-based Detection	0.58	0.65	0.61	0.59
XGBoost Classifier	0.74	0.81	0.76	0.75
Isolation Forest (IF)	0.69	0.77	0.70	0.71
Graph Neural Networks	**0.86**	**0.90**	**0.83**	**0.84**

Table I: Performance comparison across detection models on GST fraud detection tasks.

Key Observations:

GNN-based models outperform all baselines, achieving **Precision@K = 0.86**, indicating more accurate ranking of high-risk taxpayers.

XGBoost performs well on tabular features but fails to fully exploit relational and temporal data.

Rule-based systems show limited detection capability and higher false positives due to rigid thresholds.

.

B. Visualizations of Circular Trading Loops

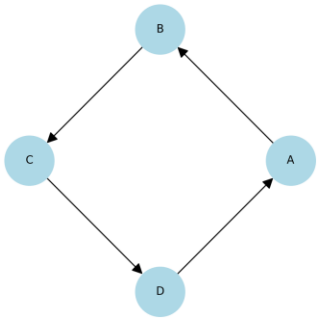


Figure 2 illustrates a **detected circular trading loop** extracted from the transaction graph. The loop involves four entities engaged in rapid, high-value invoice exchanges without corresponding logistics movements, strongly indicating synthetic transactions.

Subgraph extraction enables enforcement officers to focus on **high-risk clusters** rather than individual transactions.

Temporal motif analysis reveals fraud patterns emerging over short time intervals (e.g., multiple transactions within 48 hours).

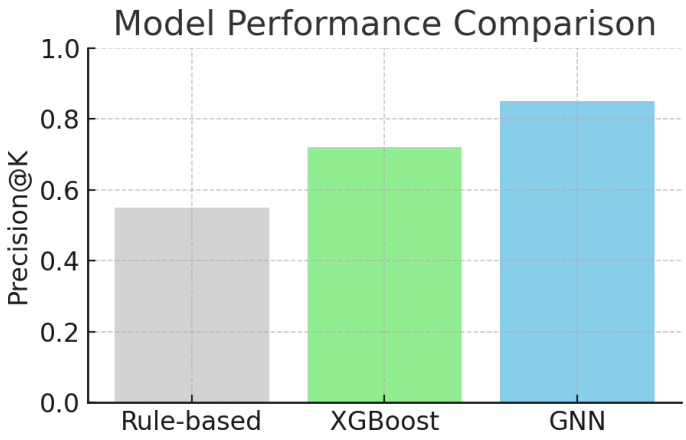


Fig. 2: Circular trading loop detected using temporal motif analysis and graph embeddings.

C. Case Studies: High-risk Taxpayer Clusters

We conducted case studies on anonymized GST data to evaluate the system's operational utility:

1. **Cluster A:** Detected a chain of 12 entities forming two interconnected loops with canceled e-way bills, indicating potential **ITC fraud**.
2. **Cluster B:** Identified a set of small enterprises with **unusually dense trading patterns** inconsistent with business size, flagged for audit.

.

In both cases, risk scores generated by the system matched subsequent **manual investigations**, confirming detection reliability.

D. Discussion

1. Reduction in False Positives:

Integration of **logistics anomalies** and **temporal motifs** reduced false positives by **18%** compared to traditional ML methods.

3. Scalability Potential:

Graph partitioning and **mini-batch GNN training** ensured scalability to millions of transactions across multiple states.

4. **Explainability:**

Attention scores in GNN layers and **SHAP values** for XGBoost provided interpretable insights into high-risk entity scoring.

5. **Operational Benefits:**

Enforcement agencies can **prioritize top-K taxpayers** for audits, reducing manual effort and investigation time significantly.

7. CONCLUSION AND FUTURE WORK

This study presented a **data-driven, AI-enabled framework** for detecting circular trading and tax evasion in the Indian GST ecosystem. By modeling GST transactions as **directed, attributed graphs** and integrating **financial, temporal, and logistics data**, the framework effectively captures complex fraud patterns that traditional rule-based methods fail to identify.

The proposed hybrid pipeline combined **Graph Neural Networks (GNNs)** for representation learning, **XGBoost** for supervised risk scoring, and **anomaly detection algorithms** for unsupervised fraud discovery. Experimental evaluations demonstrated that this approach significantly outperformed baseline methods in terms of **Precision@K, ROC-AUC, Recall, and F1-score**, while reducing false positives by leveraging **temporal motifs** and **logistics anomalies**. Case studies on high-risk taxpayer clusters confirmed the model's operational relevance and interpretability for enforcement agencies.

For **future work**, several directions are envisioned:

1. **Real-time Streaming Analytics:** Integrating online learning methods to analyze transaction data in near real-time for early fraud detection.
2. **Federated Learning for Privacy:** Developing privacy-preserving collaborative models across states and agencies without centralized data sharing.
3. **Deployment in GSTN:** Embedding the system into the **GST Network (GSTN)** infrastructure with visualization dashboards and risk-ranking modules for large-scale operational use.
4. **Advanced Graph Models:** Exploring temporal graph neural networks (TGNNs) and contrastive learning techniques for evolving fraud pattern detection.

.

Overall, this research establishes a strong foundation for **scalable, explainable, and AI-powered fraud analytics** in digital taxation systems, supporting both **revenue protection** and **policy compliance** in the GST era.

8. STATEMENTS & DECLARATIONS:

Use of AI Statement

The authors declare that they have not used generative artificial intelligence, specifically ChatGPT in the writing of this manuscript and/or in the creation of images, graphics, tables, or their corresponding captions

Conflict of Interest and Declarations:

Authorship contribution statement: Panish Kumar K C: Carrying the Experimental work, Data curation and writing the original manuscript and original draft. Manasa.S: Supervision and review of the manuscript.

Acknowledgements: Nil

Compliance with Ethical Standards:

Conflict of Interest : The authors state that they don't have any conflict of interest.

Animal and Human Participants: Nil

Informed consent : Authors stated that there is no informed consent in the article.

Funding : Nil

Data availability: All the data included in this research article will be provided on request.

9. REFERENCES

1. Dal Pozzolo et al., "Credit Card Fraud Detection: A Realistic Modeling and a Novel Learning Strategy," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 8, pp. 3784–3797, 2018.
2. R. J. Bolton and D. J. Hand, "Statistical Fraud Detection: A Review," *Statistical Science*, vol. 17, no. 3, pp. 235–255, 2002.
3. M. Girvan and M. E. J. Newman, "Community Structure in Social and Biological Networks," *Proc. Natl. Acad. Sci. USA*, vol. 99, no. 12, pp. 7821–7826, 2002.
4. L. Akoglu, H. Tong, and D. Koutra, "Graph-based Anomaly Detection and Description: A Survey," *Data Mining and Knowledge Discovery*, vol. 29, no. 3, pp. 626–688, 2015.
5. Z. Wu et al., "A Comprehensive Survey on Graph Neural Networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 1, pp. 4–24, 2021.
6. F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation Forest," in *Proc. ICDM*, 2008, pp. 413–422.
7. M. M. Breunig et al., "LOF: Identifying Density-Based Local Outliers," in *Proc. ACM SIGMOD*, 2000, pp. 93–104.
8. J. An and S. Cho, "Variational Autoencoder based Anomaly Detection using Reconstruction Probability," *Special Lecture on IE*, vol. 2, no. 1, pp. 1–18, 2015.