Benchmarking TabPFN Against Traditional and Relational Models for Provider Fraud Detection

Dipak Argade* and Gandeeban K

Allianz Services India

argadedipak@gmail.com, gandeebank@gmail.com

Abstract

Healthcare fraud detection is inherently difficult to model because the available data combine several challenges at once. The number of confirmed fraudulent providers is typically very small compared to legitimate ones. The features are diverse and providers are indirectly connected through shared patients and treatment patterns. Although gradient boosted decision trees remain widely used in operational systems due to their reliability on structured data, recent developments in pretrained tabular models and graph based learning suggest that alternative modelling strategies may capture these complexities more effectively. [1]

In this study, we evaluate a various modelling approaches including Logistic Regression, LightGBM, CatBoost, TabPFN, and a GraphSAGE based provider classifier on the Healthcare Provider Fraud Detection Analysis dataset. To ensure a fair comparison, we construct a provider-level representation by aggregating inpatient, outpatient, and beneficiary information into a single feature set. We further assess model performance using a cost-sensitive evaluation framework that jointly considers discrimination ability, probability calibration, and expected investigation utility. [2]

Our experiments showed TabPFN consistently provides the strongest overall ranking performance without requiring task specific tuning. At the same time, the lightweight GraphSAGE model remains competitive by incorporating provider beneficiary relationships, which proves particularly helpful for identifying borderline cases where tabular signals alone are less decisive. To encourage reproducibility and follow up work, we make our full Python pipeline publicly available for future research on relational and foundation models in insurance fraud detection. [3]

Keywords: Healthcare fraud detection, TabPFN, Graph neural networks, Provider-level modeling, Cost-sensitive learning

1. Introduction

Healthcare fraud continues to impose substantial financial pressure on both public and private insurance systems worldwide. Estimates suggest that a non-trivial share of healthcare expenditure often cited in the range of a few percent of total spending may be associated with fraudulent or abusive billing practices. At the same time, the rapid growth of healthcare data has made fraud detection increasingly complex. Modern datasets combine heterogeneous claim attributes, beneficiary characteristics, and provider behaviour, which limits the effectiveness of traditional rule-based screening and purely statistical approaches. [4]

As a result, machine learning methods have become a central component of contemporary fraud detection workflows. In practice, models such as logistic regression and gradient boosted decision trees remain widely used because they perform reliably on structured tabular data and provide interpretable signals for investigators. However, recent studies suggest that these models are not always well suited to capturing the relational nature of healthcare systems, where suspicious activity may emerge from interactions among providers, patients, and treatment patterns rather than from isolated claims alone. [5]

Graph-based learning methods have therefore attracted growing interest. Graph neural networks (GNNs) allow entities to be represented within a network structure and can model dependencies between providers linked

* Corresponding author: argadedipak@gmail.com

R. Vasanth Kumar Mehta et al. (eds.), *Proceedings of the International Conference on Intelligent Systems for a Sustainable Future (ISSF 2026)*, Atlantis Highlights in Intelligent Systems 16,
https://doi.org/10.2991/978-94-6239-693-7_77

through shared beneficiaries or similar service profiles. Early evidence indicates that such relational representations can help surface coordinated or anomalous behaviour that might be difficult to detect using independent tabular features. [6]

In parallel, a separate research direction has focused on foundation models for tabular data. These pretrained models attempt to approximate general learning strategies through large scale synthetic training and have recently demonstrated strong performance across diverse tabular classification tasks with limited task specific tuning. Despite this progress, their applicability to healthcare fraud detection remains only lightly explored.

The present study aims to bring these strands together. We evaluate traditional tabular models, a tabular foundation model, and a relational graph neural network within a unified provider level fraud detection framework, allowing their strengths and limitations to be assessed under comparable conditions.

Our contributions are threefold:

1. We construct a provider level dataset derived from the Kaggle Healthcare Provider Fraud Detection Analysis data, ensuring aggregation at the provider level to avoid claim level leakage.
2. We benchmark gradient boosted trees, TabPFN, and a relational GraphSAGE classifier within a consistent cost sensitive evaluation framework.
3. We release a reproducible implementation pipeline to support further research on combining tabular foundation models and relational learning in insurance analytics. [3]

2. Related Work

Most existing research on healthcare fraud detection has focused on supervised learning applied to aggregated claims data. In operational settings, gradient boosted decision trees are frequently preferred because they handle heterogeneous tabular inputs well and can capture nonlinear interactions between billing features. However, these approaches generally treat providers as independent observations and therefore may overlook relational signals that arise from shared patients or coordinated treatment patterns.

To address this limitation, more recent studies have begun exploring graph based representations of healthcare systems. In such approaches, entities such as providers, beneficiaries, or claims are connected through network structures, allowing models to learn from patterns of interaction rather than from isolated records. Many of these studies, however, model relationships at the claim or patient level, which can introduce scalability challenges or label leakage when applied in real world fraud detection pipelines. [7]

In parallel, a separate research stream has examined foundation models for tabular data. Models such as TabPFN aim to approximate general learning behaviour through large scale pretraining and have shown strong performance across a range of small and medium tabular benchmarks. Despite these encouraging results, their use in healthcare fraud detection particularly at the provider level remains only lightly investigated in the literature. [8]

Our work builds on both directions by evaluating relational and foundation model approaches within a single provider level framework. By combining these modelling strategies under consistent cost sensitive evaluation metrics, we aim to better understand how relational structure and pretrained tabular representations contribute to fraud detection performance in practical settings.

3. Dataset and Feature Construction

3.1 Dataset

Our experiments use the Healthcare Provider Fraud Detection Analysis dataset available on Kaggle. This dataset contains inpatient and outpatient claims linked to beneficiary records, together with provider level fraud labels derived from prior investigation outcomes. [2]

To reduce the risk of label leakage and to better reflect the level at which fraud investigations are typically conducted, we aggregate all information to the provider level and perform modelling exclusively on this representation. We further restrict the analysis to the training split provided with the dataset, ensuring that all feature construction and evaluation steps are applied consistently across models.

3.2 Provider Level Features

To represent each provider in a consistent and compact way, we aggregate claims and beneficiary information into a provider level feature vector. The goal is to summarize billing behaviour, patient reach, and case mix while keeping the representation interpretable and suitable for both tabular and graph-based models.

Claim Utilization Features: For both inpatient and outpatient claims, we compute summary statistics of reimbursed amounts at the provider level. Specifically, we record the mean reimbursed amount, the total reimbursed amount, and the number of submitted claims for each claim type. These statistics yield six utilization features in total and are intended to capture both the intensity of billing and the overall service volume associated with each provider.

Patient Volume: To approximate the breadth of a provider's activity, we also compute the number of unique beneficiaries treated across all claim types. Providers associated with unusually large patient pools may exhibit atypical service patterns, making this measure useful for highlighting potential over utilization behaviour.

Beneficiary Demographics: We further incorporate coarse demographic indicators derived from linked beneficiary records. At the provider level, we include the modal gender and the mean encoded race value of treated beneficiaries. While these features are not intended as causal indicators of fraud, they provide rough signals of patient mix and treatment complexity that may influence billing patterns.

Missing Data Handling: Aggregation occasionally results in missing values, particularly for providers with activity in only one claim category. We impute such values with zero to maintain a consistent feature space across providers and to avoid introducing additional modelling assumptions that could differ between methods.

3.3 Labels and Class Imbalance

Provider labels are mapped to a binary outcome variable indicating whether a provider was historically flagged as potentially fraudulent or not. As is common in fraud detection problems, the number of confirmed fraudulent providers is substantially smaller than the number of legitimate ones. This imbalance motivates our emphasis on precision–recall evaluation, calibration analysis, and cost sensitive performance measures rather than relying solely on accuracy.

4. Models

4.1 Tabular Baselines

We evaluate a set of commonly used tabular classifiers alongside a pretrained foundation model. Logistic Regression serves as a simple linear baseline with class balanced loss. LightGBM and CatBoost represent modern gradient boosted tree implementations designed to handle heterogeneous tabular inputs and nonlinear feature interactions.

We also include TabPFN, a pretrained tabular foundation model built on a transformer architecture. TabPFN is trained on large collections of synthetic tabular tasks to approximate a general learning procedure. At inference time, it processes the dataset in a single forward pass without conventional hyperparameter tuning, which makes it particularly appealing for small to medium tabular problems where rapid deployment is desirable. [9]

4.2 GraphSAGE Provider Classifier

To capture relational structure between providers, we additionally train a GraphSAGE based classifier. GraphSAGE is an inductive graph neural network that learns node representations by sampling and aggregating information from neighbouring nodes. This allows the model to generalize to unseen providers while incorporating relational signals derived from shared beneficiaries.

We construct a provider similarity graph in which each node represents a provider and an undirected edge connects two providers if they share at least one beneficiary. Node attributes correspond to the same provider level features used by the tabular models, ensuring comparability across approaches.

The GraphSAGE architecture used in our experiments consists of two aggregation layers with hidden dimensions of 64 and 16 units, followed by a linear classification head. This relatively lightweight design was chosen to keep

the model interpretable and computationally efficient while still enabling information exchange between providers linked through overlapping patient populations.

5. Experimental Setup

5.1 Train–Test Split

All experiments are conducted using a single stratified 80/20 split at the provider level so that the proportion of fraudulent providers is preserved in both subsets. Using a common split across models ensures that performance differences reflect modelling behaviour rather than variations in the data partition.

For methods that are sensitive to feature scale namely logistic regression, TabPFN, and the GraphSAGE model input variables are standardized using z score normalization. The scaling parameters are computed from the training set only and then applied to the test set, preventing any information from the evaluation data from leaking into the training process.

5.2 Evaluation Metrics

Because fraud detection involves identifying a rare positive class, we report multiple complementary metrics rather than relying on a single performance measure.

We include ROC-AUC as a measure of overall discrimination and PR-AUC to better capture performance on the minority class. To assess behaviour at a concrete operating point, we also report F1 score and balanced accuracy at a fixed decision threshold. Finally, we compute the Brier score to evaluate the calibration of predicted probabilities.

5.3 Cost Sensitive Evaluation

In practical fraud investigation settings, the usefulness of a model depends not only on classification accuracy but also on the trade-off between correctly identifying fraudulent providers and the cost of investigating legitimate ones. To approximate this operational perspective, we introduce a simple expected net gain metric defined as

$$\text{Net Gain} = R \cdot \text{TP} - C \cdot \text{FP}$$

where R denotes the relative benefit of correctly flagging a fraudulent provider and C represents the cost associated with investigating a non-fraudulent one. Rather than fixing a single threshold, we compute this quantity across all possible thresholds and report the maximum achievable net gain in relative utility units.

5.4 Visualization

To complement the numerical metrics, we provide several standard diagnostic plots. ROC curves illustrate overall separability, precision–recall curves highlight minority class performance, and calibration plots show how well predicted probabilities align with observed outcomes. To complement the numerical metrics, we provide several standard diagnostic plots. As shown in Fig. 1, ROC curves illustrate overall separability across models. Precision-recall behaviour is presented in Fig. 2, highlighting differences in minority class performance. Calibration characteristics are shown in Fig. 3. Finally, a comparison of average precision scores across models is summarized in Fig. 4.

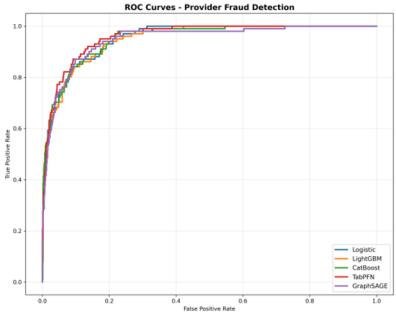

Fig. 1. ROC Curves: ROC curves for all evaluated models. Due to high overall separability at the provider level, curves largely overlap, motivating the use of PR-AUC and cost sensitive metrics for finer comparison.

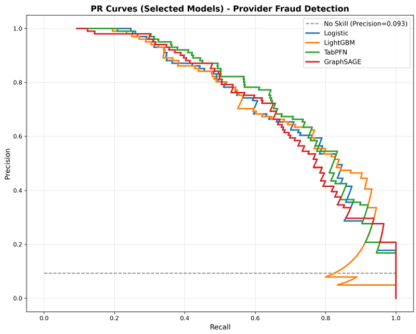

Fig. 2. PR Curves: Precision–recall curves for selected models. Differences are more pronounced in mid recall regions, which are most relevant under constrained investigation budgets.

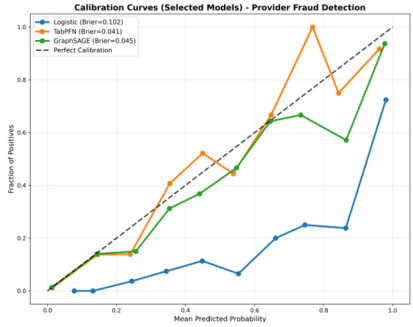

Fig. 3. Calibration Curves: Calibration curves are noisy due to class imbalance; Brier scores provide complementary scalar assessment.

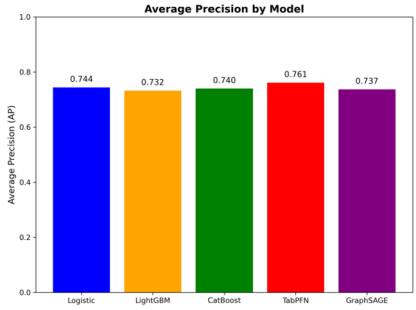

Fig. 4. Average Precision by Model: This confirms TabPFN's ranking superiority: highest PR-AUC (0.761) enables investigators to prioritize the most suspicious providers most effectively, even without hyperparameter tuning.

6. Results

6.1 Quantitative Results

The quantitative performance comparison of all evaluated models across multiple metrics is presented in Table 1.

Table. 1. Model Performance Comparison on Provider Fraud Detection Test Set

Model	ROC-AUC	PR-AUC	F1	BalAcc	Brier	Max Net Gain
Logistic	0.953	0.742	0.535	0.864	0.102	759
LightGBM	0.950	0.730	0.639	0.845	0.056	757
CatBoost	0.953	0.736	0.605	0.725	0.048	758
TabPFN	**0.960**	**0.760**	**0.640**	0.764	**0.041**	**791**
GraphSAGE	0.954	0.740	0.629	**0.767**	0.045	781

6.2 Discrimination and Precision–Recall Analysis

As shown in Table. 1, all models achieve very high ROC-AUC values above 0.95, suggesting that, once information is aggregated at the provider level, fraudulent and non-fraudulent providers are broadly separable in feature space. This result indicates that even relatively simple models can learn useful signals from utilization patterns alone.

More informative differences appear in the precision–recall analysis, which better reflects performance under strong class imbalance. Here, TabPFN attains the highest PR-AUC, indicating that it produces the most effective ranking of suspicious providers when investigation resources are limited. The GraphSAGE model performs similarly across much of the recall range and appears particularly competitive in the mid recall region, where investigators often operate when prioritizing borderline cases.

6.3 Cost Sensitive Performance and Calibration

Calibration curves reveal distinct behaviours across model families. Logistic regression tends to underestimate risk for providers assigned high predicted probabilities, whereas the tree-based models generally provide stronger calibration but still show some deviation from perfect reliability.

Both TabPFN and GraphSAGE exhibit comparatively smoother calibration profiles, with TabPFN achieving the lowest Brier score among the evaluated models. This suggests that its probability estimates align more closely with observed outcomes, which is valuable in settings where predicted risk is used to guide investigation decisions.

From a cost sensitive perspective, TabPFN achieves the highest maximum net gain across thresholds, reflecting its ability to rank fraudulent providers effectively while limiting unnecessary investigations. GraphSAGE remains competitive, largely because the relational information it incorporates improves recall for difficult to classify providers that might otherwise be missed.

7. Discussion

The results reinforce several trends observed in recent fraud detection research. First, gradient boosted trees remain strong baselines, as they are well suited to capturing nonlinear relationships in structured claims data and continue to perform reliably across different datasets. [4]

At the same time, our experiments support the view that fraud in healthcare systems often manifests through relationships between entities rather than through isolated claims. Graph based models have been proposed as a way to capture these interaction patterns more explicitly [6], and our findings are consistent with this perspective. While GraphSAGE does not dramatically outperform tabular models in overall discrimination, it improves recall and balanced accuracy for ambiguous providers, suggesting that relational context can provide useful complementary signals.

The performance of TabPFN highlights another emerging direction: pretrained tabular models that can generalize across tasks without extensive hyperparameter tuning. Although such models have mostly been studied on

benchmark datasets, our results suggest that they can transfer effectively to applied fraud detection scenarios, offering strong ranking performance with relatively little modelling effort.

Taken together, these observations point toward a complementary modelling strategy for operational fraud detection. Tree ensembles provide dependable baselines, foundation models offer rapid deployment with strong ranking ability, and graph models help capture relational effects that are difficult to express in tabular form alone. This combination aligns with recent work advocating hybrid pipelines that integrate multiple sources of signal for high stakes detection problems.

8. Conclusion

This study presents a comparative evaluation of traditional tabular models, a pretrained tabular foundation model, and a relational graph neural network for healthcare provider fraud detection within a unified cost sensitive framework.

Our results confirm that gradient boosted trees remain reliable and competitive baselines. At the same time, we find that TabPFN can deliver superior ranking performance without task specific tuning, making it attractive for scenarios where rapid deployment or limited modelling resources are considerations. The GraphSAGE model further demonstrates that incorporating provider relationships can improve recall for difficult cases, supporting the broader view that fraud detection benefits from explicitly modeling interactions between entities.

Overall, the findings suggest that future fraud detection systems may benefit from combining pretrained tabular models with relational graph representations rather than relying on a single modelling approach.

Several directions remain open for further study, including the use of richer multi entity graphs, incorporation of temporal dynamics in billing behaviour, development of hybrid tabular graph ensembles, and exploration of calibrated decision policies tailored to investigation budgets.

To facilitate continued work in this area, we release a fully reproducible implementation pipeline alongside this study. [3]

Code Availability

The complete experimental pipeline, including data preparation, model training, and evaluation scripts, is publicly available at: https://github.com/dipakml/healthcare-provider-fraud-benchmark/tree/main

References

1. Healthcare insurance fraud detection using data mining https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-024-02512-4
2. HEALTHCARE PROVIDER FRAUD DETECTION ANALYSIS https://www.kaggle.com/datasets/rohitrox/healthcare-provider-fraud-detection-analysis
3. https://github.com/dipakml/healthcare-provider-fraud-benchmark/tree/main
4. Fraud detection in healthcare claims using machine learning: A systematic review https://www.sciencedirect.com/science/article/pii/S0933365724003038
5. Health insurance fraud detection based on multi-channel heterogeneous graph structure learning https://www.sciencedirect.com/science/article/pii/S2405844024060766
6. Fraud detection and explanation in medical claims using GNN architectures https://www.nature.com/articles/s41598-025-22910-6
7. Fraud Detection in Healthcare Insurance Claims Using Machine Learning https://www.mdpi.com/2227-9091/11/9/160
8. TABPFN: A TRANSFORMER THAT SOLVES SMALL TABULAR CLASSIFICATION PROBLEMS IN A SECOND https://arxiv.org/pdf/2207.01848
9. https://github.com/PriorLabs/TabPFN