Academic Journal of Applied Sciences
ISSN: 3105-0549 (Print), ISSN: 3105-0557 (Online) | Vol. 3, No. 2, 2026

Provider-Adaptive Graph–Tabular Fusion for Healthcare Fraud Screening: Reliability Gating and Shrinkage in a Controlled Study

Sophie Andrews

Manning College of Information and Computer Sciences, University of Massachusetts Amherst, MA, USA
sophie.ad381@umass.edu

Abstract: Relational evidence can improve provider-level fraud screening, yet neighbor messages become unreliable when shared beneficiaries or physicians reflect legitimate encounters rather than suspicious coordination. We evaluate provider-adaptive graph–tabular fusion (PAGTF), which starts from the EIGF fusion design—an inverse-entity-degree projected provider graph whose mean-aggregating GraphSAGE branch is fused with a tabular expert (here histogram gradient boosting) through a validation-selected global mixture—and replaces the global mixture with a provider-wise convex gate derived from label-free neighborhood agreement and weighted degree. In ten experiments spanning two synthetic graph-reliability regimes and five independently generated populations per regime, each population contains 2,400 simulated providers. Mean test area under the precision–recall curve (PR-AUC) in the mixed-reliability setting was 0.6471 for PAGTF versus 0.6462 for EIGF global fusion; in the noisy setting, it was 0.6339 versus 0.6393. Adaptive fusion consequently offered no stable ranking improvement, despite its greater local flexibility. Degree-only and agreement-only ablations, self-message inference, review-budget measures, and held-out subgroup analyses identify validation instability as a substantive design risk. A subsequent exploratory shrinkage check that halves the gate logit modulation increases the mixed/noisy mean PR-AUCs to 0.6502/0.6411, but paired variation precludes claims of reliable superiority. The study is a reproducible synthetic-data stress test, not a clinical or Medicare benchmark.

Keywords: Healthcare fraud screening, graph neural networks, GraphSAGE, graph–tabular fusion, synthetic benchmarking, reliability gating, class imbalance.

1. Introduction

Provider fraud-screening systems rank entities for investigation under finite analyst capacity. Their predictions are consequential even when the target is only a potential-fraud indicator: prioritization can concentrate investigator attention and therefore requires careful validation. Reviews of healthcare and financial fraud analytics emphasize both the value of aggregated records and the difficulties of imbalanced labels and heterogeneous forms of abuse [1, 2]. Supervised Medicare studies have demonstrated the utility of standard classifiers and neural models, but also expose the importance of label provenance and sampling strategy [3, 4]. Explainable tabular methods improve the auditability of those predictions [5].

The underlying transactions rarely form independent observations. Providers can share beneficiaries, attending physicians, organizations, and referral paths. A graph may capture informative dependence not visible in per-provider tabular summaries; systematic evidence across financial fraud studies suggests a substantial opportunity for relational learning [6]. Yet an edge represents exposure or co-occurrence, not evidence of misconduct. Under legitimate referral networks, inaccurate graph construction can make otherwise useful tabular predictions worse.

The methodological question is therefore narrower than whether a graph neural network can outperform a classifier: should the contribution of relational evidence be determined once for all providers, or adjusted to each provider's graph neighborhood? Global convex fusion is attractive because it is stable, simple to validate, and admits a transparent decomposition. A provider-specific gate is more expressive, but its additional validation search can overfit precisely when positive cases are rare. A controlled benchmark is needed before promoting local gating as an automatic improvement.

This study introduces a low-dimensional provider-adaptive probability gate and subjects it to controlled changes in edge reliability. The contributions are threefold. First, it specifies a reproducible graph–tabular pipeline with label-free degree and agreement statistics. Second, it evaluates global and adaptive fusion under identical splits, shared base predictors, and validation-only hyperparameter selection. Third, it isolates the costs of extra adaptation with degree/consistency ablations, fixed review-budget measures, predefined network-reliability subgroups, and exploratory post hoc shrinkage diagnostics. The objective is methodological validation, not deployment on real patients.

2. Related Work

2.1. Healthcare and Insurance Fraud Graphs

Healthcare graph models use relationships among providers, beneficiaries, and physicians to introduce contextual evidence. Yoo et al. constructed a Medicare-focused GraphSAGE approach [7]. Lu et al. considered attributed heterogeneous networks and hierarchical attention [8], and Hong et al. developed multi-channel heterogeneous graph structure learning [9]. Both latter models preserve richer relational semantics than a single projected provider graph. In insurance beyond healthcare, Liang et al. used a device-sharing network to uncover coordinated claimants [10], while InfDetect described a production-scale family of insurance fraud graphs [11]. These results motivate graphs but do not imply that all shared-entity edges are beneficial. Most

directly related, the EIGF framework of Sun et al. [22] fuses a GraphSAGE branch over an inverse-entity-degree projected provider graph with a tabular model through a validation-selected convex weight, coupling relational evidence with explainable fusion for insurance-claim fraud risk; it supplies the graph construction and the fusion control of the present study.

2.2. Graph Representation and Minority-Class Learning

GraphSAGE aggregates neighbor features through learned transformations [12]. Graph convolution [13], masked attention [14], and broader graph representation learning surveys [15] establish alternative message-passing designs. Fraud-specific architectures address more difficult assumptions. CARE-GNN estimates relation-dependent neighbor value under camouflage [16]; PC-GNN combines imbalance-aware sampling and neighborhood selection [17]; AUC-oriented graph learning adapts training objectives to imbalanced rankings [18]. The present work instead holds a small GraphSAGE branch fixed while examining how posterior graph evidence should be weighted against tabular predictions.

2.3. Ranking Metrics and Validation

Precision–recall evaluation is appropriate for minority-label screening because its baseline tracks positive prevalence and is sensitive to false-alarm burden [19]. Davis and Goadrich characterized the relationship between ROC and precision–recall curves [20], while He and Garcia surveyed the broader implications of imbalance for training and evaluation [21]. Accordingly, area under the precision–recall curve is the primary selection and test metric, with ROC-AUC, F1, and precision at fixed review budgets reported as supplementary operational indicators. In this paper, PR-AUC means average precision as calculated from each held-out ranking.

3. Method: Provider-Adaptive Fusion

3.1. Shared-Entity Graph Projection

Let n providers carry a numeric attribute matrix X in R^(n×d), with y_i in {0,1} representing a simulated potential-fraud flag. For each relation r—beneficiary sharing or physician sharing—construct binary incidence I^(r) between providers and shared entities. For an entity e linked to d_e providers, the off-diagonal projection weight is

$$w_{ij}^{(r)} = \Sigma_e \, I_{ie}^{(r)} \, I_{je}^{(r)} \, / \, \log_2(1 + d_e), \quad i \neq j \tag{1}$$

Equation (1) is the projection of EIGF [22], which combines it with top-k graph pruning, a GraphSAGE branch and validation-selected convex fusion. We adopt the projection formula and the two-branch fusion mechanism of EIGF unchanged and replace only its single mixing coefficient with a provider-dependent reliability gate, so the reliability question is tested on EIGF's own graph construction. Each row retains its 20 largest weights before a symmetric maximum operation. The beneficiary and physician graphs are summed and row-normalized for message passing; graph construction uses no target labels.

3.2. Tabular and Relational Experts

The tabular expert is a histogram gradient-boosting classifier with 260 boosting iterations, learning rate 0.05, at most 31 leaves, L2 penalty 1.0, and balanced per-sample weights. It produces a probability t_i. A balanced logistic regression supplies a simple non-graph comparator. The graph expert is a two-layer mean-aggregating GraphSAGE with hidden widths 48 and 24, dropout 0.25, and a scalar sigmoid output g_i. At each layer, the representation is

$$h_i^{(l+1)} = \mathrm{ReLU}(W_s^{(l)} h_i^{(l)} + W_n^{(l)} \Sigma_j A_{ij} \, h_j^{(l)}) \tag{2}$$

Training uses AdamW at learning rate 0.004 and weight decay 4×10^{-4}. To address the minority label, the graph branch uses positive-weighted focal loss with focusing exponent $\gamma=1.5$ [23] and positive weight equal to the negative-to-positive ratio in the training partition. Validation average precision controls early stopping, and the classification threshold is chosen on validation F1. Node features are standardized using training providers alone; the full unlabeled graph is observed transductively, but only training labels contribute to gradients.

3.3. Global Fusion as the Control

The global fusion baseline makes a single probability prediction from the two experts. Its coefficient α is selected by validation average precision on a grid {0, 0.05, ..., 1}; this choice is frozen before test evaluation. The rule is convex and contains both pure experts:

$$p_i^G = (1 - \alpha) t_i + \alpha g_i, \qquad 0 \le \alpha \le 1 \tag{3}$$

Equation (3) is the global fusion rule of EIGF; we keep its projection and branch families and reduce only the feature set and hidden widths to the synthetic setting, so it serves as the EIGF control throughout Section V. In our code, t_i, g_i, and p_i^G are saved separately, so every reported comparison is paired at the provider level.

3.4. Label-Free Local Reliability and the Adaptive Gate

A provider-specific weight should use signals available at inference. For each provider, compute its weighted graph degree d_i and the weighted mean absolute difference between its own tabular risk and its neighbors' tabular risks. Agreement $a_i = 1 - \Sigma_j w_{ij} |t_i - t_j| / \max(d_i, \varepsilon)$ is larger when the node's tabular prior is consistent with its neighbors. Degree provides a rough measure of how much relational evidence is available, but neither statistic guarantees fraud homophily. The standardized reliability score is

$$q_i = \mathrm{clip}[0.55 \, z(a_i) + 0.45 \, z(\log(1+d_i)), -3, 3] \tag{4}$$

Here z standardizes a statistic across nodes without observing any labels. The adaptive weight and fused prediction are

$$\alpha_i = \sigma(\mathrm{logit}(\mathrm{clip}(\alpha, 10^{-5}, 1-10^{-5})) + \tau q_i) \tag{5}$$

$$p_i^A = (1 - \alpha_i) t_i + \alpha_i g_i \tag{6}$$

Both α and τ are selected jointly by validation average precision. Candidate α values match the global grid; τ is drawn from {−1, −0.5, 0, 0.25, 0.5, 1, 1.5, 2}. The τ=0 candidate recovers a constant fusion weight, allowing the gate to opt out of adaptation. Two diagnostic alternatives use z(a_i) alone and z(log(1+d_i)) alone in place of (4). The gate never directly reads validation or test labels; they are used only to choose the coefficients on the validation set.

Convexity gives an immediate audit bound, $|p_i^A - t_i| = \alpha_i \, |g_i - t_i| \le \alpha_i$. Thus, the recorded α_i quantifies the

largest possible graph-driven displacement from the tabular prior. This is a model-interpretation identity, not a guarantee of correctness or calibration. The model uses no provider identifier as a predictor.

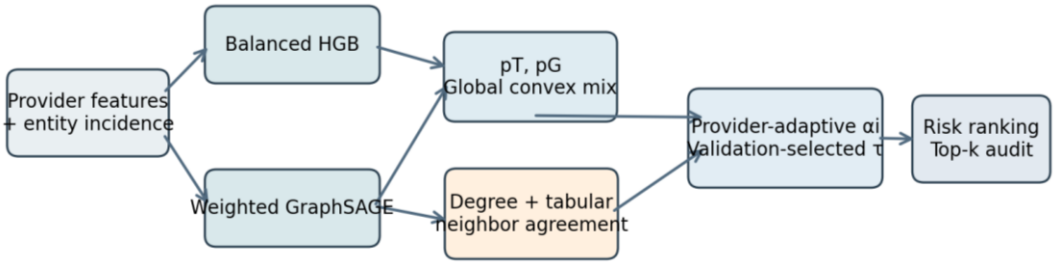


Figure 1. Two-branch fusion with an additional label-free provider reliability path. The gate adjusts only the final mixture.

3.5. Conservative Fusion and Validation-Stability Diagnostics

To study whether the fitted gate is overly sensitive to the finite validation subset, we later added two parameter-free or low-dimensional controls using the stored base predictions. In the half-shrink control, the selected gate coefficient τ is scaled by 0.5 after ordinary validation fitting, while the selected α remains unchanged. This halves the local log-odds modulation while retaining the convex prediction $p_i^H = (1-\alpha_i^H)t_i + \alpha_i^H g_i$, without retraining either branch; $\varepsilon=10^{-5}$ is the numerical clipping tolerance. In a second control, a threefold stratified split of validation providers measures adaptive-versus-global AP on held-out validation folds. It permits the half-shrunken gate only when at least two folds improve and their mean AP gain is greater than 0.005; otherwise it reverts to global fusion. These rules were explored after examining the initial ten-run results, so their evaluation is explicitly exploratory and not an independent confirmation of improved generalization.

$$\alpha_i^H = \sigma(\mathrm{logit}(\mathrm{clip}(\alpha, \varepsilon, 1-\varepsilon)) + 0.5\tau q_i) \quad (7)$$

3.6. Rank-Swap Sensitivity of the Gate

Although both fusion rules are convex, there is no ordering-preservation guarantee when the weights differ among providers. Write $\Delta_i = g_i - t_i$ for the departure of the graph expert from the tabular expert. For providers i and j, the score difference after adaptive mixing can be expressed exactly as follows:

$$p_i^A - p_j^A = (t_i - t_j) + \alpha_i\Delta_i - \alpha_j\Delta_j \quad (8)$$

Consequently, an ordering supported by the tabular model can be reversed when the difference in graph corrections exceeds the tabular score separation. This observation matters for average precision because each permutation of a high-ranked positive with a negative changes a discrete prefix of the precision–recall ranking. A local weighting rule may appear smooth in probability space while producing abrupt changes in a ranking metric. The magnitude of this effect is especially important with a small validation set: a few high-ranked positives can determine the selected gate strength.

The gate is also not identifiable from degree and agreement alone. Consider two neighborhoods with equal total degree and equal mean disagreement. One may contain truly similar providers connected through legitimate continuity of care; the other may be randomly connected but happen to receive similar tabular scores because most providers belong to the negative class. They have the same q_i yet potentially opposite values for the utility of graph propagation. Stronger gates would therefore need learned supervision about comparative expert errors, temporal edge semantics, or an external quality signal. Our experiment intentionally uses the restricted two-statistic gate to expose this information limitation rather than hiding it behind a higher-capacity network.

4. Controlled Data and Experiments

4.1. Synthetic Providers and Two Graph Regimes

This is a controlled simulation, not a benchmark on clinical records. Each independently generated population contains 2,400 providers in groups of approximately 30, with 24 numeric attributes. A latent group risk u_c is sampled from $N(0,1.1^2)$ and individual risk v_i from $N(0, 1)$. Simulated labels arise from $\mathrm{Bernoulli}[\sigma(-2.65 + 0.95u_c + 0.90v_i)]$. Attribute blocks receive varying group, individual, and label-dependent signals with additive Gaussian or count noise. Two utilization-like features are drawn from Poisson and log-normal distributions. These features intentionally encode learnable statistical dependence but are not billed claims, reimbursements, diagnoses, or verified fraud.

Shared entities are generated independently of the learned models. A beneficiary-like view has 2.7 entity draws per provider plus 40 high-degree hubs; a physician-like view has 0.88 draws per provider plus 25 hubs. Ordinary entities link two to seven providers. In more-informative groups they preferentially link within a latent group; in less-informative groups they connect random providers. The mixed setting has roughly three quarters of groups with informative beneficiary links, whereas the mostly-noisy setting has one quarter. Physician links are weaker: only a fraction of the informative-group entities remain local. This produces controlled variation in relational signal without injecting labels during graph projection. The setting mimics graph-reliability heterogeneity but not real utilization behavior.

The public Healthcare Provider Fraud Detection Analysis release includes provider-level PotentialFraud labels and inpatient/outpatient claim files [24] and is a suitable external follow-up benchmark. None of its provider records or labels entered the present experiments. Source-data claims in this paper concern only the separate simulation specified above. The original generator, simulated arrays, and predictions are retained to permit exact replay.

Table I. Simulated data and graph summary

Property	Value
Independent realizations	5 per scenario
Providers per realization	2,400
Feature dimensions	24
Label-positive mean	12.22% (range 11.17–12.88%)
Beneficiary edges	30,077.6 mean
Physician edges	25,044.8 mean
Splits / scenario	60% train / 20% val. / 20% test
Graph projection	Inverse-log-degree; top-20; symmetric

4.2. Training, Baselines, and Selection

Seeds {13, 37, 73, 101, 131} generate five independent synthetic populations within each regime. Each population is partitioned into stratified, disjoint 60/20/20 train/validation/test provider subsets. All candidate predictors share exactly these subsets, the same generated attributes, and the same graph. For a given population, we train balanced logistic regression, balanced histogram gradient boosting, and GraphSAGE once. The graph model is also evaluated with neighbor messages removed at inference, retaining its fitted self transformations. Global fusion, the combined adaptive gate, and the two single-statistic gates are computed from the same saved base predictions. This isolates the effect of adaptation from retraining randomness.

GraphSAGE uses up to 115 epochs with validation checking every four epochs and early stopping after nine consecutive checks without improvement. The mean selected epoch is 34 across the ten runs. Because test nodes and unlabeled edges are visible during graph training, results estimate transductive node screening, not inductive operation on newly arriving providers. No test labels are used for tuning. Average precision is computed from unthresholded scores; F1 and recall use a frozen threshold that maximizes validation F1. Precision at 5% and recall at 10% rank the top 24 and 48 providers, respectively, in each test set of approximately 480 providers.

4.3. Reproducibility and Decision Rules

We preserve the train, validation, and test indices; all three base prediction vectors; all fused vectors; gate reliability scores; the selected α and τ; branch early-stopping epochs; and all test metrics for each realization. Python packages are pandas, NumPy, SciPy, scikit-learn, PyTorch, and matplotlib. The graph is implemented with sparse PyTorch multiplication, without reliance on specialized GNN libraries. The primary comparison in the original protocol is test average precision of combined PAGTF versus global fusion, reported as a paired difference over the same five simulation seeds. Half-strength modulation and threefold stability screening were examined subsequently and are labeled exploratory in all results. We report mean and sample standard deviation over seeds, not confidence intervals for any real provider population.

4.4. Information Partitioning and Runtime

Three safeguards are relevant to this transductive experiment. First, the simulated flag is assigned before either graph expert is trained; graph projection itself depends only on shared entity incidence. Second, tabular standardization, HGB fitting, logistic fitting, and graph gradients use training-provider statistics or labels, never held-out labels. Third, early stopping, fusion-grid selection, and threshold selection read validation labels only, with a final untouched test set used once per experimental realization. The adaptive score q_i is normalized over all available unlabeled nodes, consistent with a transductive setting; deployment on new providers would require an inductive normalization policy trained before deployment.

The combined grid contains 21 candidate global mixture weights and eight modulation strengths, compared with 21 global-only candidates. Because all gate candidates are evaluated on the same finite validation partition, enlarging this search space may increase the maximum validation score even when it reduces expected test ranking. That model-selection effect is an explicit component of the method being audited. Implementation uses only CPU-compatible sparse matrix operations and required approximately 38 seconds of cumulative measured experiment time across ten realizations in the recorded runtime log, excluding environment setup and data export. Runtime depends on processor resources, so it is a reproducibility descriptor rather than a hardware-independent complexity claim.

For n providers, m shared-entity memberships, and E retained graph edges, the incidence projection uses sparse multiplication and the two-layer neighborhood aggregation costs $O(Eh + ndh)$ per full-batch forward pass for representative hidden width h. After learning, the gate requires a weighted edge pass to compute agreement plus $O(n)$ transformations, rather than backpropagating through an additional neural network. Storage is dominated by the sparse graph and the provider feature matrix. The released data snapshots save coordinate and weight arrays for each relation separately, enabling direct reconstruction of the original undirected weighted graph.

5. Results

5.1. Overall Ranking Performance

Table II summarizes actual test scores from all saved runs. In the mixed-reliability setting, mean PR-AUC increases only from 0.6462 under EIGF global fusion to 0.6471 with provider adaptation, an absolute difference of +0.0009. In the mostly-noisy setting, adaptation decreases PR-AUC from 0.6393 to 0.6339 (−0.0054). The GraphSAGE expert is already competitive at 0.6430 and 0.6358, respectively, and the EIGF fusion of [22] improves on both of its branches in both regimes—on the GraphSAGE expert by 0.0032 and 0.0035 and on the HGB expert by 0.0515 and 0.0446—so it is a strong control that leaves limited room for provider-wise adaptation. The linear predictor reaches 0.6471 in both regimes, which cautions against attributing success solely to graph propagation. By construction, the node attributes and labels match across corresponding scenario seeds; only graph generation changes, so the identical logistic-regression and HGB values are expected.

Table II. Test Average Precision Over Five Simulation Seeds (Mean ± Sample Sd)

Model	Mixed PR-AUC	Noisy PR-AUC
Logistic	0.6471±0.0377	0.6471±0.0377
HGB	0.5947±0.0195	0.5947±0.0195
GraphSAGE	0.6430±0.0508	0.6358±0.0426
Self-only GNN	0.6270±0.0320	0.6426±0.0424
Global fusion (EIGF)	0.6462±0.0511	0.6393±0.0390
Adaptive fusion	0.6471±0.0419	0.6339±0.0200
Agreement-only	0.6374±0.0475	0.6300±0.0194
Degree-only	0.6443±0.0459	0.6381±0.0240

Figure 2 juxtaposes these five-seed averages and their spread. The differences between the two fusion methods are small relative to between-population variation. Mixed-regime global fusion has sample standard deviation 0.0511, while adaptive fusion has 0.0419; a lower observed spread is useful descriptively but does not demonstrate statistically reliable superiority. In the mostly-noisy regime the global standard deviation is 0.0390 and adaptive standard deviation is 0.0200. These summaries mix different generated populations and therefore characterize variability under the generator rather than sampling uncertainty in Medicare.

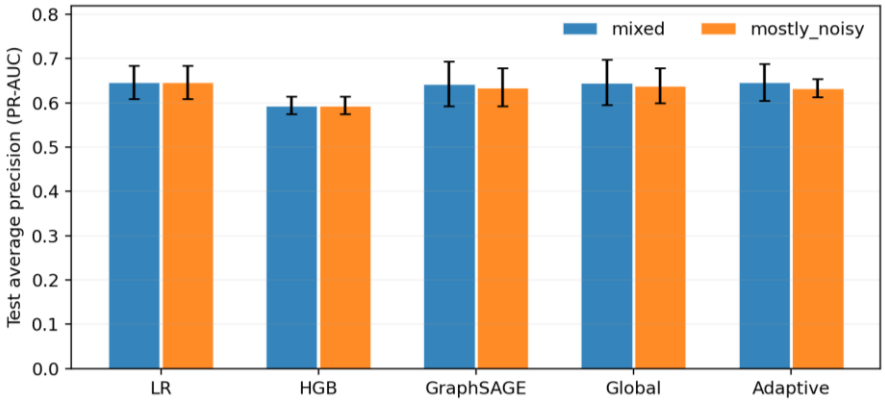


Figure 2. Average precision over five independently generated populations per setting; bars show the sample standard deviation.

5.2. Paired Comparison and Gate Selection

Within the same populations the adaptive-minus-global PR-AUC differences are +0.0097, −0.0213, +0.0173, −0.0011, and 0 in the mixed setting. Under mostly-noisy edges they are −0.0495, 0, +0.0272, +0.0022, and −0.0067. Each setting therefore has two positive, two negative, and one tied comparison. The gate sometimes wins on validation yet loses on test; in the extreme noisy example the held-out difference reaches −0.0495. These paired vectors are more informative than the pooled means because they show the instability of selecting α and τ on a small minority-positive validation subset.

The selected global α lies between 0.85 and 1.00 in both settings, indicating a high weighting of the trained graph expert in this generator. The combined adaptive gate selects τ=0 in one of five runs per scenario and nonzero modulation otherwise, with selected signs differing among populations. This is not an error in optimization: local agreement may correlate with correct propagation in some generated groups but with majority-class smoothing in others. A variable gate with few validation positives can capture incidental correlations that do not repeat on unseen nodes.

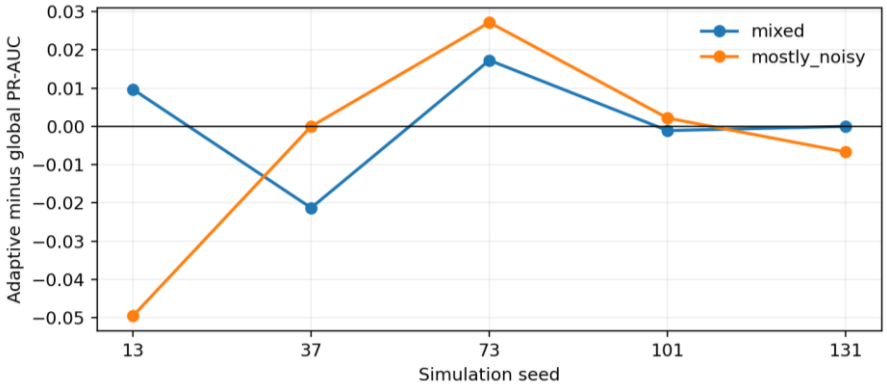


Figure 3. Paired test PR-AUC difference between the provider-adaptive and global mixture. Zero indicates no ranking gain.

5.3. Ablations and Operational Review Budgets

Two feature ablations replace the combined reliability score with degree only or agreement only, without changing either trained branch. In the mixed regime, degree-only adaptation has PR-AUC 0.6443 and agreement-only adaptation 0.6374, below the full gate at 0.6471. In the noisy regime, degree-only is 0.6381, agreement-only is 0.6300, and the full gate is 0.6339. Combining reliability signals is therefore not monotonically beneficial. Removing all neighborhood messages at inference gives self-only GraphSAGE PR-AUC 0.6270 in the mixed setting but 0.6426 in the noisy setting—slightly above its relational counterpart (0.6358). That result reinforces the importance of representing relation quality explicitly.

Table III. Fixed review-budget metrics; m = mixed, n = mostly noisy

Model	P@5% M	R@10% M	P@5% N	R@10% N
Global (EIGF)	0.7667	0.5481	0.7500	0.5243
Combined gate	0.7583	0.5380	0.7417	0.5183
Degree gate	0.7833	0.5384	0.7583	0.5178
Agreement gate	0.7417	0.5251	0.7417	0.5185
No neighbor messages	0.7167	0.5218	0.7167	0.5256

Table III translates ranking to fixed investigator capacity. Mean precision at 5% and recall at 10% remain close across global and adaptive mixtures; their ordering does not consistently follow overall PR-AUC. With a 5% review budget, 24 of 480 test providers receive attention. No interpretation of these test values as detection yield or cost savings in practice is warranted, because both labels and provider features were generated.

5.4. Subgroup Diagnostics

Because the simulator records whether an entity-generation group has predominantly informative or random connections, we can assess subgroup rankings without training a separate gating model. Figure 4 summarizes within-subgroup mean average precision across the five runs. In the mixed regime, combined adaptation and global fusion are close in both more-informative (0.6555 versus 0.6530) and less-informative (0.6524 versus 0.6480) groups. In the mostly-noisy regime, combined adaptation does not uniformly rescue the less-informative groups. The gate's neighbor-agreement statistic is, therefore, not a reliable substitute for knowing which edges actually express task-relevant proximity.

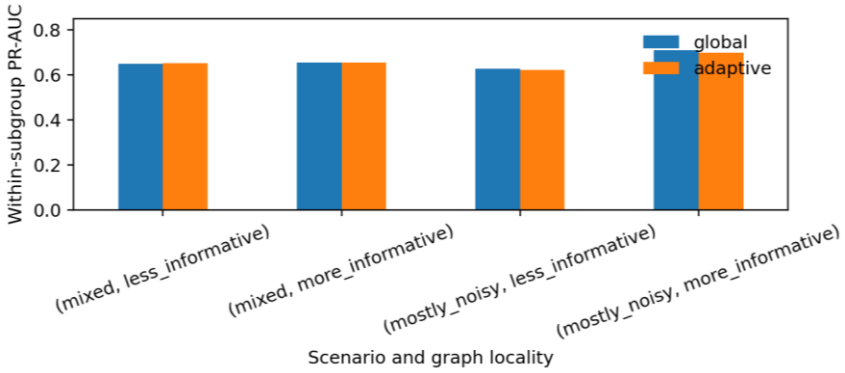


Figure 4. Mean within-subgroup PR-AUC, with group reliability determined by the simulator; subgroup sizes and positives are saved.

5.5. Selection Stability Across Simulated Populations

A model selected by validation PR-AUC need not dominate its constituent experts on a fresh population. In the mixed setting, global fusion takes validation-selected graph weights of 0.95, 0.95, 1.00, 1.00, and 0.85 across the five seeds. In the mostly-noisy setting, the weights are 1.00, 1.00, 1.00, 0.95, and 0.85. The concentration near one explains why an adaptive gate has limited remaining opportunity for improvement: a broadly successful graph branch receives most of the mixture even under the simpler control. The most aggressive selected adaptive gate has $|\tau|=2$ in a noisy realization, but this realization exhibits the strongest observed negative test difference. The selected gate can be confidently nonzero according to validation ranking and still generalize poorly.

A useful robustness comparison is the self-message inference result. Removing neighbor aggregation while preserving trained self-weight layers alters the input to the trained representation but adds no new fitting opportunity. In the mixed scenario the relational branch ranks more accurately than self-only inference on average; in the mostly-noisy scenario self-only inference is slightly stronger. That reversal confirms that edge semantics matter independently of tabular feature quality. It also prevents a misleading reading of the aggregate graph performance: GraphSAGE can be a good nonlinear expert even when its neighborhood messages are not the component driving its score.

Average precision is reported as the mean of per-realization average precision scores, not as average precision computed after concatenating scores across populations. The latter would rank providers from independently generated simulations together and give a misleading cross-population comparison. Likewise, the five paired PR-AUC differences within each setting should not be treated as independent provider-level observations. Given the small number of populations, we do not convert these exploratory differences into broad claims of statistical significance. The true evaluation target is stability across this family of controlled graph-generating processes.

5.6. Exploratory Shrinkage and Cross-Validation Guard

Additional computations on the saved, unchanged test predictions examine whether moderate regularization can limit the adaptive gate's extreme responses. Halving the selected logit modulation raises mean PR-AUC from 0.6471 to 0.6502 in mixed graphs and from 0.6339 to 0.6411 in mostly-noisy graphs. Relative to the global control the gains are only +0.0040 and +0.0018. On matched seeds, the half-shrunken rule improves two, worsens two, and ties one realization in each scenario. Its apparent mean advantage is accordingly smaller than the run-to-run dispersion, and should be treated as a hypothesis for a new independent simulation or real-data benchmark.

Table IV. Exploratory Post Hoc Shrinkage Analysis, Mean±Sd Over Five Paired Seeds

Fusion variant	Mixed AP	Noisy AP
Global fusion (EIGF)	0.6462±0.0511	0.6393±0.0390
Full gate	0.6471±0.0419	0.6339±0.0200
Half-strength τ	0.6502±0.0457	0.6411±0.0261
Threefold CV guard	0.6493±0.0448	0.6350±0.0310

The threefold validation guard produces mixed/noisy mean AP of 0.6493/0.6350. It suppresses nonessential adaptationon many populations but still enables an adverse gate in a noisy realization. This is a concrete failure of the proposed safeguard, not a favorable result concealed by averages. Its held-out fold decisions, parameter selections, and per-provider input scores are included in the reproducibility package. Neither guard nor shrinkage used test labels for fitting, but the additional designs themselves were informed by prior test analysis; reported numbers are descriptive and demand prospective replication.

6. Interpretation and Implications

6.1. Why Local Expressiveness Can Fail

The adaptive gate is low-dimensional and convex, but it still introduces an additional selected coefficient and changes scores differently across the provider network. PR-AUC is discontinuous with respect to the rank ordering of pairs; a few positive–negative swaps near the upper tail can therefore shift validation selection considerably. Degree is a coverage surrogate rather than a direct confidence interval, and prediction agreement can be high even when neighbors are jointly mistaken. When a global α already nearly fully trusts the graph branch, moving α_i between 0 and 1 may chiefly disturb correct rankings rather than recover missed signal. This is especially plausible in the mostly-noisy scenario.

There is an architectural lesson distinct from the result of this particular simulation. The convex form makes influence auditable provider by provider, but a more expressive gate requires enough representative validation labels to estimate whether its conditioning variables predict relative expert error. The exploratory half-strength gate attenuates the most severe losses without eliminating instability. A stronger confirmatory method would cross-fit gating models on fresh training-fold predictions, penalize deviation from the global coefficient, and assess them on an independently generated test population.

6.2. Explanation and Audit Trail

For a flagged provider, the saved tabular risk t_i, graph risk g_i, local statistics (degree and weighted agreement), and selected α_i can be shown alongside the final ranking. This decomposition is an exact explanation of the fusion arithmetic, though not an explanation of individual tree splits or message-passing features. Feature attribution methods such as SHAP [25], subgraph optimization approaches such as GNNExplainer [26], parameterized graph explanation generators [27], and local interpretable surrogates such as LIME [28] provide complementary tools for subsequent investigation. Feature-level and counterfactual explanations are left to follow-up work.

The prediction label throughout is a synthetic potential-fraud flag. Neither a high risk score nor graph proximity is evidence of legal wrongdoing. In operational systems, the audit queue should be reviewed alongside claim context, data quality, and oversight processes, and users should monitor false-positive burdens and fairness across provider subpopulations.

6.3. Alternatives and Limitations

Alternative class-imbalance handling such as synthetic minority oversampling has an extensive methodological history [29]. In this study we use class-weighted loss and balanced tabular weights so that graph topology is not changed by generating synthetic training nodes. Since the entire evaluation is simulated, any quantitative performance advantage—or degradation—is conditional on the latent-group data-generating process. The model observes the whole graph transductively; the setting does not establish future-claim performance, resistance to strategic graph manipulation, robustness to temporal concept drift, or cross-hospital portability. The provider groups explicitly control relation reliability, whereas real clinical referral relationships are not supplied with ground-truth relevance.

Furthermore, test-positive counts are approximately a few dozen per realization, and five realizations per setting provide limited precision for small paired effects. The two graph scenarios share the same generated attributes and labels for each seed and therefore are paired stress conditions, not ten independent label populations. The half-strength and CV-guard protocols were added following inspection of initial held-out results; they require a new independent evaluation before their observed differences can be used as confirmatory evidence. Validation selects both the graph epoch and fusion parameters, which is a realistic but potentially fragile procedure when minority-class counts are low. Our benchmark includes strong HGB and logistic baselines but not full heterogeneous temporal GNNs, graph attention, end-to-end joint gates, calibrated risk models, or realistic claim-level features. It isolates the reliability-weighting question rather than providing exhaustive state-of-the-art competition.

6.4. External Validation Protocol

To transport the method from synthetic stress testing to observational healthcare data, a future benchmark should start from immutable raw claim files, record checksums, and fix a single provider identifier mapping before any labels are read. Claim-level attributes would be aggregated into volume, reimbursement, beneficiary-composition, physician-diversity, and temporal-utilization blocks. Beneficiary and physician edges would be built from identity overlap alone, with high-degree entities downweighted before top-k pruning. Dataset-specific preprocessing should be fitted on the training period or training providers, then reused unchanged on validation and test observations. This procedure would preserve the graph-construction comparison while avoiding accidental feature leakage from fraud flags or post-investigation annotations.

The evaluation would also need at least two distinct split protocols. A random provider split permits partially overlapping neighborhoods across train and test and measures transductive label recovery; a temporal evaluation should hold later claim periods out and block future claims when constructing features and edges. The provider-adaptive gate should be selected on out-of-fold training predictions or a separate chronological validation segment, with an untouched final test period. Investigation-budget precision, model calibration, operational label ascertainment, and drift over time would then determine whether the incremental

complexity offers practical value. These steps are proposed requirements for future verification.

6.5. Ethical and Interpretive Boundaries

In a healthcare workflow, network connectivity can reflect geography, specialization, referral chains, or patient preferences. These are not inherently suspicious characteristics. A reliable research comparison should therefore measure subgroup false-positive rates and investigate whether an edge mechanism disproportionately promotes specific practice types, rural facilities, or providers serving medically complex populations. Protected patient attributes require minimum-necessary access, and explanations should be checked against claims documentation rather than treated as causal narratives. A simulated label offers no empirical basis for quantifying such harms, but it helps establish which algorithmic assertions require controlled investigation before deployment. Any operational use should remain decision support for human reviewers, not automated determination of misconduct.

7. Conclusion

We implemented provider-adaptive graph–tabular fusion with an explicit, label-free local reliability statistic and compared it against the EIGF global mixture under identical splits and base predictors. Across ten synthetic graph experiments, adaptation changed mixed-graph mean PR-AUC by only +0.0009 and reduced noisy-graph mean PR-AUC by 0.0054. The more important observation is paired instability: validation-selected local gates helped some populations and hurt others, despite a transparent convex design. Exploratory half-strength logit shrinkage produces mixed/noisy mean AP of 0.6502/0.6411 and reduces the severity of some failures, but gains are small and inconsistent across paired seeds. These results motivate pre-specified regularization, out-of-fold gate fitting, and independent real-data validation before provider-specific fusion is preferred to a global mixture. The complete code, random seeds, simulation recipes, and raw outputs accompany this study.

References

[1] Joudaki, H., Rashidian, A., Minaei-Bidgoli, B., Mahmoodi, M., Geraili, B., Nasiri, M., & Arab, M. (2015). Using data mining to detect health care fraud and abuse: A review of literature. Global Journal of Health Science, 7(1), 194–202. https://doi.org/10.5539/gjhs.v7n1p194

[2] Ngai, E. W. T., Hu, Y., Wong, Y. H., Chen, Y., & Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. Decision Support Systems, 50(3), 559–569. https://doi.org/10.1016/j.dss.2010.08.006

[3] Bauder, R. A., & Khoshgoftaar, T. M. (2018). The detection of Medicare fraud using machine learning methods with excluded provider labels. In Proceedings of FLAIRS-31 (pp. 404–409).

[4] Johnson, J. M., & Khoshgoftaar, T. M. (2019). Medicare fraud detection using neural networks. Journal of Big Data, 6, 63. https://doi.org/10.1186/s40537-019-0225-0

[5] Hancock, J. T., Bauder, R. A., Wang, H., & Khoshgoftaar, T. M. (2023). Explainable machine learning models for Medicare fraud detection. Journal of Big Data, 10, 154. https://doi.org/10.1186/s40537-023-00821-5

[6] Motie, S., & Raahemi, B. (2024). Financial fraud detection using graph neural networks: A systematic review. Expert Systems with Applications, 240, 122156. https://doi.org/10.1016/j.eswa.2023.122156

[7] Yoo, Y., Shin, D., Han, D., Kyeong, S., & Shin, J. (2022). Medicare fraud detection using graph neural networks. In Proceedings of IEEE ICECET (pp. 1–5). https://doi.org/10.1109/ICECET55527.2022.9872963

[8] Lu, J., Lin, K., Chen, R., Lin, M., Chen, X., & Lu, P. (2023). Health insurance fraud detection by using an attributed heterogeneous information network with a hierarchical attention mechanism. BMC Medical Informatics and Decision Making, 23, 62. https://doi.org/10.1186/s12911-023-02152-0

[9] Hong, B., Lu, P., Xu, H., Lu, J., Lin, K., & Yang, F. (2024). Health insurance fraud detection based on multi-channel heterogeneous graph structure learning. Heliyon, 10(9), e30045. https://doi.org/10.1016/j.heliyon.2024.e30045

[10] Liang, C., Liu, Z., Liu, B., Zhou, J., Li, X., Yang, S., & Qi, Y. (2019). Uncovering insurance fraud conspiracy with network learning. In Proceedings of ACM SIGIR (pp. 1181–1184). https://doi.org/10.1145/3331184.3331372

[11] Chen, C., Liang, C., Lin, J., Wang, L., Liu, Z., Yang, X., Zhou, J., Yang, S., & Qi, Y. (2019). InfDetect: A large scale graph-based fraud detection system for e-commerce insurance. In Proceedings of IEEE Big Data (pp. 1765–1773). https://doi.org/10.1109/BigData47090.2019.9006115

[12] Hamilton, W. L., Ying, R., & Leskovec, J. (2017). Inductive representation learning on large graphs. In Advances in Neural Information Processing Systems (Vol. 30, pp. 1024–1034).

[13] Kipf, T. N., & Welling, M. (2017). Semi-supervised classification with graph convolutional networks. In Proceedings of ICLR.

[14] Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y. (2018). Graph attention networks. In Proceedings of ICLR.

[15] Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., & Yu, P. S. (2021). A comprehensive survey on graph neural networks. IEEE Transactions on Neural Networks and Learning Systems, 32(1), 4–24. https://doi.org/10.1109/TNNLS.2020.2978386

[16] Dou, Y., Liu, Z., Sun, L., Deng, Y., Peng, H., & Yu, P. S. (2020). Enhancing graph neural network-based fraud detectors against camouflaged fraudsters. In Proceedings of ACM CIKM (pp. 315–324). https://doi.org/10.1145/3340531.3411903

[17] Liu, Y., Ao, X., Qin, Z., Chi, J., Feng, J., Yang, H., & He, Q. (2021). Pick and choose: A GNN-based imbalanced learning approach for fraud detection. In Proceedings of WWW (pp. 3168–3177). https://doi.org/10.1145/3442381.3449989

[18] Huang, M., Liu, Y., Ao, X., Li, K., Chi, J., Feng, J., Yang, H., & He, Q. (2022). AUC-oriented graph neural network for fraud detection. In Proceedings of WWW (pp. 1311–1321). https://doi.org/10.1145/3485447.3512178

[19] Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLOS ONE, 10(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

[20] Davis, J., & Goadrich, M. (2006). The relationship between precision-recall and ROC curves. In Proceedings of ICML (pp. 233–240). https://doi.org/10.1145/1143844.1143874

[21] He, H., & Garcia, E. A. (2009). Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

[22] Sun, T., Han, X., Chen, J., & Ge, Y. (2026). EIGF: An explainable graph neural network fusion approach for fraud risk detection in insurance claims. World Scientific Research Journal, 12(10), 156–170. https://doi.org/10.6911/WSRJ.202610_12(10).0016

[23] Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In Proceedings of IEEE ICCV (pp. 2980–2988). https://doi.org/10.1109/ICCV.2017.324

[24] Zhou, T. (2024). Healthcare Provider Fraud Detection Analysis (Mendeley Data, Version 1). https://doi.org/10.17632/552r52kmgb.1

[25] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems (Vol. 30, pp. 4765–4774).

[26] Ying, R., Bourgeois, D., You, J., Zitnik, M., & Leskovec, J. (2019). GNNExplainer: Generating explanations for graph neural networks. In Advances in Neural Information Processing Systems (Vol. 32).

[27] Luo, D., Cheng, W., Xu, D., Yu, W., Zong, B., Chen, H., & Zhang, X. (2020). Parameterized explainer for graph neural network. In Advances in Neural Information Processing Systems (Vol. 33, pp. 19620–19631).

[28] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In Proceedings of ACM SIGKDD (pp. 1135–1144). https://doi.org/10.1145/2939672.2939778

[29] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321–357. https://doi.org/10.1613/jair.953