A Federated Graph Neural Network for Integrating Clinical, Imaging, and Genetic Data Across Hospital Systems

Malika *Baizakova,* Kadyralieva *Nuraiym*

INAI, Ankara 1/5, Bishkek,720082, Kyrgyzstan

Abstract

Integrating heterogeneous patient data from multiple hospitals remains a critical challenge in precision medicine. We propose FedGNN, a federated graph neural network framework that jointly models clinical records, medical imaging features, and genomic profiles without centralizing sensitive patient data. Each participating hospital trains a local GNN over a patient-disease-gene heterogeneous graph; only model parameters are shared via a privacy-preserving aggregation protocol. Experiments on simulated multi-site data demonstrate that FedGNN achieves a mean AUC of 0.884 across three target diseases, outperforming locally trained baselines by 6.3% and matching a centralized upper bound within 1.8%. The framework preserves differential privacy (epsilon = 1.0) and reduces communication overhead by 42% compared to FedAvg. Our results suggest that federated GNNs can serve as a viable, privacy-compliant pathway for cross-institutional data fusion.

Keywords

Federated learning, graph neural networks, multi-modal data integration, privacy-preserving machine learning, clinical informatics, heterogeneous graphs

1. Introduction

Modern precision medicine relies on integrating diverse patient data: structured electronic health records (EHR), radiological and pathological imaging, and high-throughput genomic assays. When these data streams are combined, predictive models for complex diseases such as cancer, cardiovascular disorders, and rare genetic conditions achieve substantially higher accuracy than single-modality approaches [1, 2]. However, the most valuable datasets are locked inside individual hospital information systems, separated by regulatory, technical, and institutional barriers.

Federated learning (FL) [3] offers a principled solution: models are trained locally and only parameter updates are transmitted to a central aggregator, so raw patient records never leave the originating institution. Despite this appeal, standard FL methods based on averaging dense weight tensors (FedAvg [4]) struggle when the underlying data are relational in nature. Patient-level graphs encoding gene co-expression, phenotype similarity, and clinical co-occurrence cannot be naturally represented as flat feature vectors.

*IVUS 2026: Information Society and University Studies 2026, May 12, Kaunas, Lithuania

malikabajzakova146@gmail.com (M. Baizakova); nuraiymkadyralieva@gmail.com (K. Nuraiym)

Copyright © 2026 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

published 2026-11-06

CEUR-WS.org/Vol-4317/paper9.pdf

Graph neural networks (GNNs) [5] address this by learning directly over graph-structured data. Recent work has shown that heterogeneous GNNs outperform homogeneous counterparts on biomedical knowledge graphs [6]; yet, no prior study has deployed heterogeneous GNNs inside a federated loop across real hospital environments while satisfying formal differential privacy (DP) guarantees.

In this paper we close this gap and make the following contributions:

1. We design a heterogeneous patient-disease-gene graph schema that unifies clinical, imaging, and genetic modalities into a single relational structure (Section 3).
2. We propose FedGNN, a federated training protocol for heterogeneous GNNs with DP-SGD noise injection and a communication-efficient gradient compression scheme (Section 4).
3. We conduct extensive experiments on multi-site simulated data derived from publicly available cohorts and show that FedGNN consistently outperforms locally trained and standard FL baselines (Section 5).

2. Related Work

2.1. Federated Learning in Healthcare

Since the seminal work of McMahan et al. [4], federated learning has been applied to a wide range of clinical tasks, including diabetic retinopathy detection [7], COVID-19 chest X-ray classification [8], and genomic variant calling [9]. A persistent challenge is statistical heterogeneity: data distributions differ markedly across sites owing to population demographics, local clinical protocols, and device calibration differences. FedProx [10] and SCAFFOLD [11] address this through proximal regularization and variance reduction, respectively, but both assume that each client holds an independent feature vector per sample — an assumption that breaks down for graph-structured biomedical data.

2.2. Graph Neural Networks for Biomedical Data

GNNs have achieved state-of-the-art performance on drug-target interaction prediction [12], protein function annotation [13], and multi-omics disease classification [14]. Heterogeneous GNNs such as HAN [15] and HGT [16] further exploit the rich semantics of multi-relational biomedical graphs. Several studies construct patient similarity networks [17] or phenotype graphs [18] to capture population-level structure, but these approaches require pooling data across sites, which is impractical in privacy-sensitive scenarios.

2.3. Federated Graph Learning

Federated graph learning is an emerging field. FedGCN [19] and SpreadGNN [20] propose graph-specific aggregation strategies but operate on homogeneous, single-relational graphs. FGNN [21] extends this to heterogeneous settings yet does not incorporate formal privacy guarantees. Our work builds on these foundations by combining heterogeneous GNNs with DP-SGD and a communication-efficient protocol tailored to the clinical data fusion problem.

3. Graph Schema and Data Modalities

3.1. Heterogeneous Graph Construction

For each participating hospital h, we construct a heterogeneous graph $G_h = (V_h, E_h, T_V, T_E)$ $where\ T_V = \{Patient, Disease, Gene, Imaging\}$ $and\ T_E = \{has_diagnosis, expresses, similar_to, co_occurs\}$. Node features for each entity type are derived as follows.

Patient nodes encode age, sex, lab results, and vital signs as a 128-dimensional feature vector obtained by linear projection. Disease nodes are initialized with PubMed sentence-level embeddings of clinical descriptions. Gene nodes use pre-trained Gene2Vec [22] embeddings (200 dimensions). Imaging nodes are represented by features extracted from a pre-trained ResNet-50 fine-tuned on multi-modal medical imaging datasets.

3.2. Edge Construction

has_diagnosis edges connect each patient to their recorded ICD-10 diagnoses, weighted by frequency. expresses edges link patients to genes with a measured expression level above the cohort median. similar_to edges are formed between patient pairs whose cosine similarity on clinical features exceeds a threshold theta = 0.75. co_occurs edges connect disease-gene pairs based on literature co-occurrence scores extracted from the DisGeNET database [23].

4. FedGNN Framework

4.1. Local GNN Architecture

Each client trains a two-layer Heterogeneous Graph Transformer (HGT) [16] with 256 hidden dimensions and 8 attention heads per relation type. The node-level representations from the final layer are mean-pooled over each patient subgraph to obtain a patient embedding e_i ∈ R^256. A two-layer MLP classifier head maps e_i to a disease probability score.

4.2. Federated Aggregation with Differential Privacy

We adopt DP-SGD [24] at each client: gradients are clipped to L2 norm C = 1.0 and Gaussian noise $N(0, sigma\^2\ C\^2)$ is added before upload. The global aggregation step uses weighted FedAvg, where the weight of client h is proportional to the number of labeled patient nodes $|V\^P_h|$. We track the privacy budget using the Renyi Differential Privacy accountant [25], targeting $epsilon = 1.0\ at\ delta = 10\^\{-5\}\ over\ T = 100$ communication rounds.

4.3. Communication Efficiency

To reduce bandwidth, we compress gradient updates using top-k sparsification (k = 5% of parameters) combined with error-feedback accumulation [26]. This reduces the per-round communication cost from O(|theta|) to O(0.05|theta|) without measurable degradation in convergence.

5. Experiments

5.1. Experimental Setup

We simulate a four-hospital federation using partitions of three public datasets: MIMIC-IV [27] for clinical records, TCGA [28] for genomic profiles, and RadImageNet [29] for imaging. We target three binary prediction tasks: Type-2 diabetes onset (T2D), colorectal cancer (CRC), and

Alzheimer's disease (AD). Each hospital partition is created by stratified sampling with a Dirichlet concentration alpha = 0.5 to simulate realistic non-IID conditions. We report AUC, F1-score, and accuracy averaged over five random seeds.

5.2. Baselines

We compare FedGNN against four baselines:

- Local GNN: each hospital trains independently, no federation.
- FedAvg-GNN: standard FedAvg applied to GNN weights.
- FedProx-GNN: FedProx with mu = 0.01 regularization.
- Centralized GNN: upper bound assuming full data pooling (no privacy).

5.3. Results

Table 1 summarizes the main results. FedGNN achieves a mean AUC of 0.884 across the three tasks, outperforming Local GNN by 6.3% and FedAvg-GNN by 2.1%. The gap to the centralized upper bound is only 1.8%, which is within the margin expected from DP noise at epsilon = 1.0. FedProx-GNN performs comparably to FedAvg-GNN on homogeneous partitions but degrades on the most heterogeneous split (alpha = 0.1), where FedGNN retains its advantage.

Table 1

AUC scores (mean ± std) across three disease prediction tasks (T2D, CRC, AD)

Method	T2D	CRC	AD	Mean
Local GNN	0.791±0.021	0.803±0.018	0.848±0.015	0.821±0.018
FedAvg-GNN	0.841±0.014	0.855±0.011	0.882±0.010	0.863±0.012
FedProx-GNN	0.843±0.013	0.857±0.012	0.880±0.011	0.860±0.012
FedGNN (ours)	**0.868±0.010**	**0.879±0.009**	**0.906±0.008**	**0.884±0.009**
Centralized GNN	0.882±0.009	0.898±0.008	0.923±0.007	0.901±0.008

5.4. Privacy-Utility Trade-off

Figure 1 illustrates the AUC-epsilon curve for FedGNN on the CRC task. As epsilon increases from 0.1 to 10, AUC rises from 0.821 to 0.887, plateauing near epsilon = 1.0. This confirms that the chosen privacy budget provides a favorable operating point where utility loss is modest relative to the privacy protection afforded.

Communication cost analysis shows that top-k sparsification reduces the total bytes transferred per round by 42.1% (from 148 MB to 86 MB for the full model) while maintaining convergence within five additional rounds compared to full-gradient transmission.

6. Discussion

FedGNN demonstrates that heterogeneous graph learning and federated optimization are complementary, not competing, approaches. The relational structure of the patient-disease-gene graph provides inductive bias that improves local model quality, which in turn reduces the variance amplified by DP noise. This synergy is the primary reason FedGNN closes much of the gap to the centralized baseline.

Several limitations should be acknowledged. First, our evaluation relies on simulated federation from public datasets; a real deployment would require institutional agreements, IRB approvals, and careful handling of rare-event nodes (e.g., patients with ultra-rare variants). Second, the graph schema assumes that all sites use compatible ICD-10 and gene identifiers, which may require ontology alignment in practice. Third, the imaging modality is currently represented by pre-extracted features; end-to-end training over raw DICOM volumes within a federated loop remains computationally prohibitive.

Future work will address cross-lingual and cross-ontology harmonization, validate the framework in real multi-institutional settings with proper IRB approvals and data-sharing agreements, and incorporate secure aggregation alongside defenses against adversarial and membership-inference attacks to strengthen deployment readiness. We also plan to incorporate temporal patient trajectories via temporal GNNs to capture longitudinal disease progression, explore vertical federation where different hospitals hold complementary modalities for the same patient cohort, and investigate end-to-end training over raw DICOM imaging volumes to replace pre-extracted features and improve representation learning

7. Conclusion

We presented FedGNN, a federated heterogeneous graph neural network framework for multi-modal clinical data integration across hospital systems. By combining HGT-based local learning with DP-SGD and communication-efficient gradient compression, FedGNN achieves a mean AUC of 0.884 under epsilon = 1.0 differential privacy — within 1.8% of the centralized upper bound and 6.3% above purely local training. These results establish federated GNNs as a viable path toward privacy-preserving precision medicine at the scale of multi-institutional patient cohorts.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

1. Acosta J.N., Falcone G.J., Rajpurkar P., Topol E.J. Multimodal biomedical AI. Nature Medicine 28 (2022) 1773–1784. doi:10.1038/s41591-022-01981-2.
2. Huang S.-C., Shen L., Lungren M.P., Yeung S. GLoRIA: A multimodal global-local representation learning framework for label-efficient medical image recognition. ICCV (2021) 3942–3951.
3. Khan, A. (2022). Machine learning for chronic disease prediction. CEOS Public Health Research, 1(1), 101.
4. Khan, A., & Usupova, E. (2025). Quantum machine learning algorithms for genome disease diagnosis. DOI: 10.12700/AIS.2025.005
5. Kairouz P., et al. Advances and open problems in federated learning. Foundations and Trends in Machine Learning 14 (2021) 1–210.

6. McMahan H.B., Moore E., Ramage D., Hampson S., Arcas B.A. Communication-efficient learning of deep networks from decentralized data. AISTATS (2017) 1273–1282.
7. Kipf T.N., Welling M. Semi-supervised classification with graph convolutional networks. ICLR (2017).
8. Yue X., et al. Graph embedding on biomedical networks: methods, applications and evaluations. Bioinformatics 36 (2020) 1241–1251.
9. Xu J., et al. Federated learning for healthcare informatics. Journal of Healthcare Informatics Research 5 (2021) 1–19.
10. Yang Q., Liu Y., Chen T., Tong Y. Federated machine learning: Concept and applications. ACM Transactions on Intelligent Systems and Technology 10 (2019) 1–19.
11. Choudhary A., et al. Federated learning for privacy-preserving genomic analysis. npj Genomic Medicine 7 (2022) 23.
12. Li T., et al. Federated optimization in heterogeneous networks. MLSys (2020).
13. Karimireddy S.P., et al. SCAFFOLD: Stochastic controlled averaging for federated learning. ICML (2020) 5132–5143.
14. Nguyen T., et al. Graph convolutional networks for drug response prediction. IEEE/ACM TCBB 18 (2021) 881–891.
15. Gligorijevic V., et al. Structure-based protein function prediction using graph convolutional networks. Nature Communications 12 (2021) 3168.
16. Wang T., et al. MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification. Nature Communications 12 (2021) 3445.
17. Wang X., et al. Heterogeneous graph attention network. WWW (2019) 2022–2032.
18. Hu Z., Dong Y., Wang K., Sun Y. Heterogeneous graph transformer. WWW (2020) 2704–2710.
19. Zitnik M., Agrawal M., Leskovec J. Modeling polypharmacy side effects with graph convolutional networks. Bioinformatics 34 (2018) i457–i466.
20. Pierson E., et al. Inferring multi-dimensional rates of aging from cross-sectional data. Genome Biology 20 (2019) 16.
21. Yao L., et al. Federated graph neural network for cross-silo medical data analysis. IEEE Transactions on Big Data (2023).
22. He C., et al. SpreadGNN: Decentralized multi-task federated learning for graph neural networks on molecular data. AAAI (2022).
23. Zhang K., et al. Subgraph federated learning with missing neighbor generation. NeurIPS (2021).
24. Du M., Li F., Zheng G., Srikumar V. Gene2Vec: distributed representation of genes based on co-expression. BMC Genomics 20 (2019) 82.
25. Pinero J., et al. The DisGeNET knowledge platform for disease genomics. Nucleic Acids Research 48 (2020) D845–D855.
26. Abadi M., et al. Deep learning with differential privacy. CCS (2016) 308–318.
27. Mironov I. Rényi differential privacy. CSF (2017) 263–275.
28. Lin Y., Han S., Mao H., Wang Y., Dally W.J. Deep gradient compression: Reducing the communication bandwidth for distributed training. ICLR (2018).
29. Johnson A.E.W., et al. MIMIC-IV, a freely accessible electronic health record dataset. Scientific Data 10 (2023) 1.
30. TCGA Research Network. The Cancer Genome Atlas Pan-Cancer analysis project. Nature Genetics 45 (2013) 1113–1120.
31. Mei X., et al. RadImageNet: An open radiologic deep learning research dataset for effective transfer learning. Radiology: Artificial Intelligence 4 (2022) e210315.