Graph Neural Networks for Heart Failure Prediction on an EHR-Based Patient Similarity Graph

Heloisa Oss Boll[1,2], Stefan Byttner[2], Mariana Recamonde-Mendoza[1]
[1]Institute of Informatics, Universidade Federal do Rio Grande do Sul, Porto Alegre, Brazil

[2]School of Information Technology, Halmstad University, Halmstad, Sweden

hoboll@inf.ufrgs.br, stefan.byttner@hh.se, mrmendoza@inf.ufrgs.br

Abstract. *Accurate disease prediction is critical in healthcare. This study introduces Graph Neural Networks (GNNs) and Graph Transformer (GT) models to predict heart failure (HF) incidence at the next hospital visit using a patient similarity graph. Using MIMIC-III Electronic Health Records (EHR) data, we constructed a patient similarity graph via K-Nearest Neighbors (KNN) with embeddings from diagnoses, procedures, and medications. We implemented GraphSAGE, GAT, and GT to predict HF, evaluating performance with F1, AUROC, and AUPRC metrics against baseline models. A three-axed interpretability analysis explored model decision-making. GT achieved the highest performance (F1: 0.5361, AUROC: 0.7925, AUPRC: 0.5168). While Random Forest (RF) showed a similar AUPRC, GT provided superior interpretability by leveraging patient relational information. Analyzing attention weights, graph structure, and clinical features offered deeper insights into classification decisions.*

1. Introduction

The integration of electronic health records (EHRs) and artificial intelligence (AI) has transformed healthcare, enabling clinical risk prediction models to assess individualized risks for complex diseases like heart failure (HF) [Sharma et al. 2021]. Deep learning has played a key role in this progress, automatically extracting features from dense datasets like EHRs. However, many models overlook the relational structure in EHR data, treating it as a flat bag of features [Choi et al. 2020]. Graph neural networks (GNNs) address this limitation by modeling patient data as graphs, where nodes represent clinical entities (e.g., patients, diagnoses, treatments) and edges capture co-occurring relationships.

Despite increasing interest in GNNs for EHR analysis, patient similarity graphs remain underexplored. This study extends prior research by using a patient similarity graph constructed from medication, diagnosis, and procedure codes to predict HF using GNNs. Moreover, it also introduces a novel graph-based interpretability analysis to enhance medical decision-making and trust in deep clinical models.

Our key contributions include: (1) developing a new methodology for constructing patient similarity graphs using dense, pretrained EHR representations; (2) benchmarking three GNN architectures—GraphSAGE, GAT, and GT—for HF prediction, addressing gaps in prior model comparisons [Tariq et al. 2023, Tang et al. 2023, Pieroni et al. 2021]; (3) conducting an ablation study to assess the relevance of clinical features; and (4) introducing an interpretability framework leveraging graph statistics, attention weights, and clinical features, improving upon existing graph-based explainability approaches.

2. Materials and Methods

2.1. Data Sources

The study was based on MIMIC-III dataset [Johnson et al. 2015], a benchmark dataset encompassing EHRs from patients of a hospital in the United States. Diagnoses and procedures were encoded using ICD-9 (International Classification of Diseases) codes and medications using NDC (National Drug Code). Patients with at least two hospital visits were selected, resulting in 4,760 patients and 8,891 unique visits. Our dataset was imbalanced, with about 28% of patients with HF.

2.2. Patient Representation

We used pre-trained 300-dimensional embeddings for ICD-9 and NDC medical codes [Choi et al. 2016]. These embeddings were aggregated at the visit level and averaged to create patient-level representations, and were utilized to both construct the patient similarity graph and as input features in the predictive models.

2.3. Patient Similarity Graph

Patient similarity was computed using cosine similarity, and a KNN graph (K=3) was constructed using NetworkX. K=3 was selected based on the distortion metric, ensuring each patient node connected to its three most similar neighbors. We note some nodes had more than three edges if chosen as the closest match by multiple others. The final graph comprised 4,760 nodes and 11,763 edges.

2.4. Model Architectures and Implementation

We evaluated three GNN architectures: GraphSAGE, GAT, and GT, implemented using PyTorch Geometric. For model evaluation, the graph was split into training, validation, and test sets (60-20-20) using the DeepSNAP library.

3. Quantitative results

3.1. GNN architecture performance

We first evaluated GNN architectures for HF prediction using the traditional binary cross-entropy (BCE) loss. GT achieved the highest F1 score (0.5328), while GraphSAGE had the highest AUPRC (0.5476) (Table 1). Confusion matrices and AUROC/AUPRC curves further illustrate GT's superior ability to identify positive cases (Figure 1).

Table 1. Test results from GNN models optimized with the BCE loss, each run thrice. The GT model shows the highest F1 and recall scores.

Metric	SAGE	GAT	GT
F1-Score	0.4758 ± 0.011	0.4832 ± 0.003	**0.5328 ± 0.003**
Accuracy	**0.8032** ± 0.004	0.7356 ± 0.000	0.7377 ± 0.002
Balanced Accuracy	0.6591 ± 0.006	0.6697 ± 0.002	**0.7112 ± 0.002**
Recall	0.3972 ± 0.008	0.5498 ± 0.005	**0.6651 ± 0.002**
Precision	**0.5931 ± 0.017**	0.4310 ± 0.001	0.4443 ± 0.003
AUROC	0.7824 ± 0.000	0.7537 ± 0.001	**0.7918 ± 0.002**
AUPRC	**0.5476 ± 0.00**1	0.4931 ± 0.001	0.5200 ± 0.002

Next, we tested alternative loss functions suitable for class imbalanced problems, replacing binary cross-entropy with weighted BCE (WBCE) and focal loss (FL) with different parameter settings. GT with FL ($\alpha = 0.75$, $\gamma = 1$) achieved the best performance (F1: 0.5531, AUROC: 0.7914, AUPRC: 0.5393) among all combinations and was selected for further experiments.

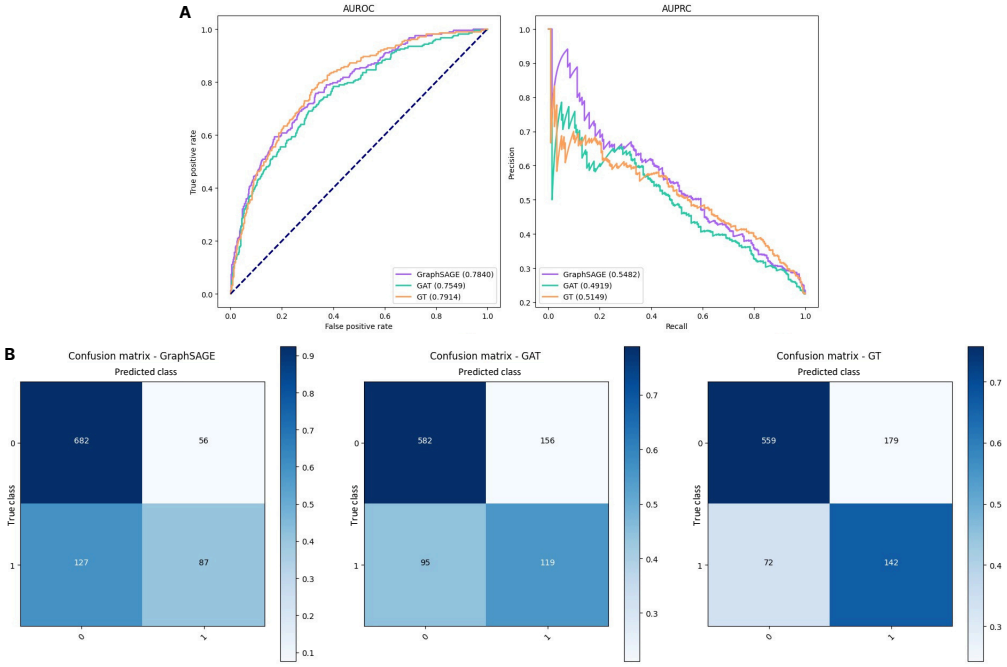

Figure 1. AUROC and AUPRC curves (A) and confusion matrices (B) for GraphSAGE, GAT, and GT models on the test set.

3.2. Impact of clinical data

To assess the contribution of each data type, we retrained the GT with FL model using only medications, procedures, or diagnoses. Medication data yielded the highest recall, followed by diagnoses, while procedures had the least impact. The best performance was achieved when all three data types were combined. An ablation study further confirmed that removing medications caused the greatest performance drop, followed by diagnoses, while excluding procedures had minimal impact. The full model achieved the highest F1 score (0.5361) and AUPRC (0.5227), underscoring the value of integrating multiple data sources (Table 2).

3.3. Benchmarking

We compared the performance of the GT with FL model against five baselines. The GT with FL model demonstrated an increased test AUROC (0.7925) and AUPRC (0.5168) compared to others (Table 3, Figure 2). Although the differences in AUPRC between

Table 2. Test results from the GT models with FL ($\alpha = 0.75$, $\gamma = 1$) for the ablation study, each run thrice.

Metric (Test)	Without diagnosis	Without prescriptions	Without procedures	Combined
F1 score	0.5233 ± 0.001	0.5071 ± 0.002	0.5275 ± 0.008	**0.5361 ± 0.003**
Accuracy	0.7066 ± 0.000	0.6964 ± 0.001	0.7321 ± 0.004	**0.7321 ± 0.002**
Balanced accuracy	0.7101 ± 0.001	0.6958 ± 0.001	0.7083 ± 0.006	**0.7166 ± 0.003**
Recall	**0.7165 ± 0.002**	0.6947 ± 0.002	0.6551 ± 0.010	0.6885 ± 0.005
Precision	0.4122 ± 0.000	0.3993 ± 0.002	0.4370 ± 0.006	**0.4389 ± 0.003**
AUROC	0.7756 ± 0.001	0.7699 ± 0.001	0.7834 ± 0.000	**0.7930 ± 0.001**
AUPRC	0.5058 ± 0.001	0.4793 ± 0.002	0.5162 ± 0.001	**0.5227 ± 0.002**

GT and Random Forest (AUPRC: 0.5132) were modest, the GT model's capacity to use graph-based relationships offers multiple benefits, further investigated in the Discussion.

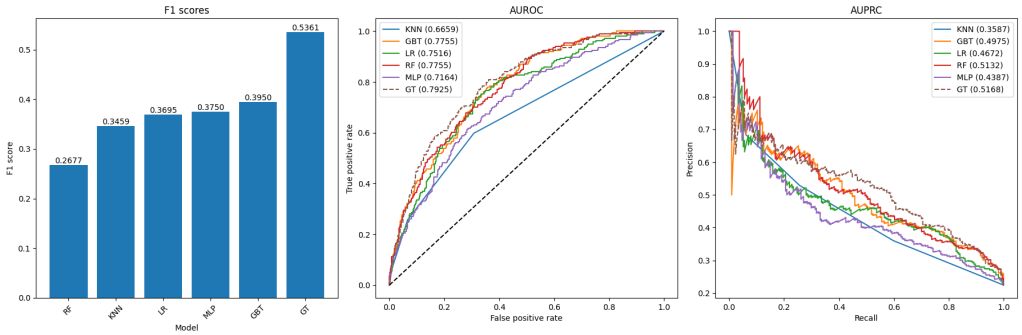

Figure 2. F1 scores, AUROC, and AUPRC curves of baseline algorithms on the test set, compared to the GT with FL, which utilizes relational information from the graph to make predictions.

Table 3. Performance metrics (F1 score, AUROC, AUPRC) of baseline algorithms on the test set, compared to the GT.

Algorithm	F1 Score	AUROC	AUPRC
RF	0.2677	0.7755	0.5132
KNN	0.3459	0.6659	0.3587
LR	0.3695	0.7516	0.4672
MLP	0.3750	0.7164	0.4387
GBT	0.3950	0.7755	0.4975
GT	**0.5361**	**0.7925**	**0.5168**

3.4. Interpretability Results

3.4.1. Graph descriptive statistics

Our analysis focused on node degree and node similarity across true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). TN and FP nodes exhibited the highest average degrees, indicating more diverse connections, while FN nodes had the fewest connections, suggesting that these HF patient profiles are more unique in terms of similarity.

3.4.2. Attention weights

Attention weights, learned during training, highlight the importance of neighboring nodes' features for classifying a target node. We observed a bimodal distribution of weights in the final GT layer, indicating that the model assigned either high or low importance to neighbors.

Furthermore, TP and FP nodes, as well as TN and FN nodes, exhibited similar attention patterns. TN nodes assigned higher attention to other negative neighbors, helping with the correct classification, while TP nodes showed a more balanced attention across neighbor types. FN nodes resemble TN patterns but with slightly more attention to positive neighbors, indicating challenges in correct classification.

3.5. Clinical features

Clinical feature analysis examined diagnosis, procedure, and prescription code patterns from each group. Essential hypertension was the most prevalent diagnosis, particularly in TP and FN, reinforcing its link to HF. Atherosclerotic heart disease frequently appeared in TP and FP, indicating its diagnostic significance but also its potential for misclassification. Atrial fibrillation was common across TP, FN, and FP, reflecting comorbidity, while FN cases showed more respiratory conditions like chronic airway obstruction, suggesting underdiagnosis in complex patients.

Procedural analysis revealed widespread use of critical care interventions like intubation, mechanical ventilation, and venous catheterization. TP and FP cases had more heart-related procedures, such as coronary bypass and arteriography, highlighting both diagnostic relevance and misclassification risks. FN cases exhibited more thoracentesis and parenteral infusion, suggesting complex comorbidity needs that may contribute to underdiagnosis. Prescription patterns emphasized pharmacotherapy's role in HF classification, with sodium chloride and dextrose commonly used. TP cases showed high incidences of heparin sodium and potassium chloride, while TP and FN shared phenylephrine HCl and metoprolol, indicating similar treatment patterns suggestive of HF.

3.6. Integrative analysis

We analyzed the one and two-hop neighborhoods for four randomly selected nodes, one from each group, along with their attention maps. The TN patient had strong similarities with negative patient profiles, with non-cardiac conditions like metabolic imbalances and liver/kidney diseases. Despite a correct classification, a relatively high probability (0.4429) and nearby positive nodes suggested potential HF risk factors. The TP patient, in turn, had clear HF-related features, matching positive neighbors with cardiovascular diseases, diabetes, and comorbidities such as kidney disease. The model relied on individual patient features rather than neighbors, probably due to strong HF-related signals.

The FN patient had only TN neighbors, with conditions like severe infections and cancer, diverging from the typical HF profile. Unique diagnoses (e.g., septicemia, breast cancer) and cancer-related procedures likely contributed to misclassification. The case also suggests a potential rare HF disease pathway. The FP case, in contrast, closely resembled true HF patients, sharing cardiovascular conditions like coronary atherosclerosis and bypass surgery. High attention weights on positive neighbors contributed to their misclassification, highlighting a potential high-risk individual who may require urgent evaluation; or a HF mislabeling case.

4. Discussion

The Graph Transformer (GT) outperformed other GNNs, likely due to its advanced attention mechanism and hence stronger discerning power. While GraphSAGE had the highest AUPRC, GT's superior recall made it more effective for identifying HF-positive cases. All models benefited from loss functions addressing class imbalance. Ablation studies found prescription codes most predictive, likely due to their prevalence in patient data, with raw NDC codes preserving medication granularity better than ATC, which is the most frequently used standard. Considering baselines, GT achieved the highest F1 and AUROC, though its AUPRC (0.5168) was close to RF (0.5132), suggesting RF

could be a resource-efficient alternative and potentially get close to GT's performance with threshold tuning. However, GT's ability to model relational interactions provided unique advantages over RF's bag-of-features-based approach: the interpretability framework leveraging graph statistics, attention weights, and clinical features demonstrated the value of graphs in healthcare, revealing clusters of high-risk individuals and potential novel disease pathways.

5. Conclusion

This study evaluated GNN models (GraphSAGE, GAT, and GT) for HF prediction in an imbalanced patient similarity graph, with GT with focal loss achieving the best performance. Prescription data emerged as the most important data source. While RF performed comparably in AUPRC, GT's ability to analyze relational patterns provided unique insights, exemplified in the extensive interpretability analysis. We open-sourced the complete study in https://github.com/hossboll/patient-gnn. Future research should explore alternative graph representations, dynamic learning, and further interpretability angles to enhance decision-making and trust in clinical outcome deep predictive models.

Acknowledgments

This work was supported by the Swedish Council for Higher Education (Linnaeus-Palme Partnership, 3.3.1.34.16456), CAPES (Finance Code 001), CNPq (308075/2021-8), and FAPERGS (22/2551-0000390-7, 21/2551-0002052-0).

References

Choi, E., Xu, Z., Li, Y., Dusenberry, M., Flores, G., Xue, E., and Dai, A. (2020). Learning the Graphical Structure of Electronic Health Records with Graph Convolutional Transformer. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(01):606–613.

Choi, Y., Chiu, C. Y.-I., and Sontag, D. (2016). Learning low-dimensional representations of medical concepts. *AMIA Summits on Translational Science Proceedings*, 2016:41–50.

Johnson, A., Pollard, T., and Mark, R. (2015). MIMIC-III Clinical Database.

Pieroni, A., Cabroni, A., Fallucchi, F., and Scarpato, N. (2021). Predictive modeling applied to structured clinical data extracted from electronic health records: An architectural hypothesis and a first experiment. *Journal of Computer Science*, 17(9):762–775.

Sharma, V., Davies, A., and Ainsworth, J. (2021). Clinical risk prediction models: The canary in the coalmine for artificial intelligence in healthcare? *BMJ Health & Care Informatics*, 28(1):e100421.

Tang, S., Tariq, A., Dunnmon, J. A., Sharma, U., Elugunti, P., Rubin, D. L., Patel, B. N., and Banerjee, I. (2023). Predicting 30-day all-cause hospital readmission using multimodal spatiotemporal graph neural networks. *IEEE Journal of Biomedical and Health Informatics*, pages 1–12.

Tariq, A., Lancaster, L., Elugunti, P., Siebeneck, E., Noe, K., Borah, B., Moriarty, J., Banerjee, I., and Patel, B. (2023). Graph convolutional network-based fusion model to predict risk of hospital acquired infections. *Journal of the American Medical Informatics Association*, 30(6):1056–1067.