Hierarchical Graph Neural Network for Patient Treatment Preference Prediction with External Knowledge

Quan Li[1], Lingwei Chen[2], Yong Cai[3,4] *, and Dinghao Wu[1]

[1] Pennsylvania State University, University Park, PA, USA
{qbl5082, dwu12}@psu.edu
[2] Wright State University, Dayton OH, USA
lingwei.chen@wright.edu
[3] California State University, Monterey Bay, CA, USA
[4] IQVIA Inc., Wayne PA, USA
yong.cai@iqvia.com

Abstract. The healthcare industry has a wealth of data that can be used by researchers and medical professionals to infer a patient's condition and intention to receive treatment using machine learning models. However, this line of research generally suffers from some limitations: (1) struggling to leverage structural interactions among patients; (2) attending to learn patient representations from electronic medical records (EMRs) but rarely considering supplementary contexts; and (3) overlooking EMR data imbalance issue. To address these limitations, in this paper, we propose a hierarchical graph neural network for patient treatment preference prediction. Doctors' information and their viewing activities are first integrated as external knowledge with EMRs to construct the hierarchical graph, where a dual message passing paradigm is then devised to perform intra- and inter-subgraph aggregation to enrich patient representations and advance label propagation. To mitigate patient data imbalance issue, a community detection method is further designed to better prediction. Our experimental results demonstrate the state-of-the-art performance on patient treatment preference prediction.

Keywords: Hierarchical graph neural network · Oncology treatments · Preference prediction · Healthcare · Community detection.

1 Introduction

In many oncology treatments, doctors and patients generally adopt watch-and-wait strategy [12]. After confirmed diagnoses, patients can wait for a long time to take aggressive treatment. For example, it takes 5 to 10 years on average for a Chronic Lymphocytic Leukemia (CLL) patient before taking treatment. But the treatment decision is highly dependent on patient condition and doctors'

* Corresponding author.

scrutiny. Some patients may only take a short period of time on watch-and-wait. Estimating and predicting if a patient has been ready to take a treatment can serve as reminders and assist doctors and patients to make the right decisions. However, due to the high variation of treatment patterns in oncology area, predicting the likelihood for a patient to take treatment is a challenging problem.

With the rapid development in machine learning and deep learning [19], the healthcare industry has started to exploit these data-driven concepts and theories into practical products and applications to predict patient conditions and propensity for treatment and medication [10, 29, 22], which in turn facilitate doctors' analyses and decisions to plan treatments for patients. One of the widely used data for such tasks is electronic medical record (EMR), which maintains rich and important patient information, and keeps growing in its volume and diversity. This has thus attracted researchers in the healthcare industry to take EMRs as inputs to train machine learning models and make patient-specific predictions through them [32, 31, 37, 35]. Though with the promising performance, these models trained on EMRs provide the successful principles to solve the high variation issues in patients, their inputs are inherently self-contained, and struggle to leverage structural interactions with other patients.

Graph neural networks (GNNs) have recently emerged as one of the most powerful techniques for graph mining [16, 18, 5]. These GNNs perform information aggregation to extract high-level features from the nodes and their neighborhoods [4], which have boosted the performances for various tasks over graphs. Therefore, a surge of effective research works build GNNs to learn structural semantics from EMRs and advance patient-specific models [8, 9, 26, 24, 3]. For example, GRAM [8] and KAME [26] constructed the knowledge graph over EMRs to depict the hierarchy of medical concepts in the form of a parent-child relationship and utilized GNNs to embed medical code to characterize each patient. Liu et al. [24] analyzed EMR using heterogeneous GNN to capture more diverse patient information (e.g., profile, symptoms, and visit history). However, these structured approaches still suffer from two limitations. (1) While attending to depict patients and learn higher-level patient representations from EMRs, this line of research rarely utilizes any supplementary contexts. As indicated by some surveys and case studies [7], the doctor-patient relationship may essentially impact on patients' treatment preferences; in other words, the external knowledge (e.g., doctor information) can be extracted to further assist in predicting if a patient would like to take treatment or not. (2) EMR data imbalance issue has been completely overlooked by current researches as well. Due to laborious process and delay effect on data annotation, the imbalance issue exists across common and rare diseases, and the downstream patient treatment distribution, which naturally enforces data-driven models to favor the majority class over the minority class and degrade their prediction performances. With this in mind, our goal here is to investigate how much patient treatment preference prediction can benefit from a structured imbalanced learning model with external knowledge.

To this end, in this paper, we propose a novel hierarchical graph neural network model with external knowledge for patient treatment preference predic-

tion that can effectively mitigate the impact of data imbalance as well. More specifically, we introduce the doctors' information and their viewing activities (captured by website topics) as external knowledge to be integrated with EMRs to enrich patient representations and advance the structured learning model for better prediction performance. The hierarchical graph is first constructed to abstract the interactions of patients, doctors, and topics, where a dual message passing paradigm is devised to perform intra-subgraph and inter-subgraph neighborhood aggregation for node representation refinement and label propagation. To cope with imbalanced patient data, a community detection method is further designed to cluster the higher-level embeddings of negative and unlabeled patients to derive community-preserving patient graph, where the treatment preference predictions for patients are produced through communities.

2 Problem Statement

EMRs contain the medical and treatment history of the patients in different practices, which allow us to predict the propensity of patients to take treatments for different diseases. In this paper, we focus on the prediction of patient oncology treatments. Without loss of generality, we represent our data as $\mathcal{X} = \{(x_{pi}, y_i)\}_{i=1}^{l} \cup \{x_{di}\}_{i=1}^{m} \cup \{x_{ti}\}_{i=1}^{n}$ consisting $l + m + n$ samples, where l is the number of patient records, m is the number of doctor records, and n represents the number of topics retrieved from website data. Unlike existing works [3, 11, 25] that merely use EMRs to train the models for performing patient-specific tasks, we constructively consider doctor information out of EMRs to interact with patients and facilitate our prediction. Each patient record x_p is annotated with a ground truth $y \in \{0, 1\}$ for a specific treatment preference, where $y = 1$ indicates that the patient prefers to take the treatment and $y = 0$ denotes that the patient has no such intention. Note that, positives are much smaller than negatives in our data and also in the real-world scenario. We initially map $\mathcal{X}$ including patients, doctors, and topics into k-dimensional feature vectors and learn a patient representation function ϕ through hierarchical graph neural network to aggregate information from patients, doctors, and topics to obtain higher-level $\mathbf{X}_p = \phi(\mathcal{X}_p, \mathcal{X}_d, \mathcal{X}_t)$, $\mathbf{X}_p \subseteq \mathbb{R}^{l \times k}$. Resting on patient representations, we aim to learn a classification model $f : \mathbf{X}_p \rightarrow \mathbf{Y}$ to perform our prediction task. Thus, the treatment preference label for a given patient data x can be predicted as:

$$y^* = \underset{y \in \{0,1\}}{\operatorname{argmax}} f_y(\mathbf{x}_p) \tag{1}$$

where $f_y(\mathbf{x}_p)$ is the confidence score of predicting patient $\mathbf{x}_p$ as treatment preference label y using the classification model f. From Eq. (1), we can see that the final label assigned to the input is the one with the highest confidence score.

3 Proposed Model

In this section, we present the technical details of our proposed model as follows, the overview of which is illustrated in Figure 1.

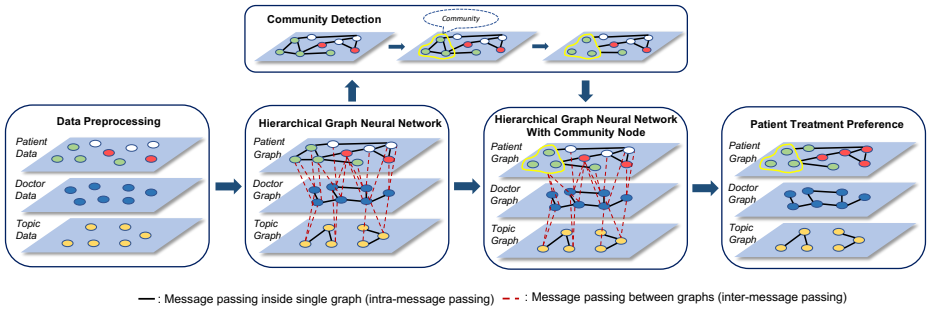


Fig. 1. The overview of our proposed model.

3.1 Hierarchical Graph Construction

To proceed with patient representation learning using GNNs, the first step is to construct the graph. As we introduce doctors' information and the website topics they have viewed as external knowledge, here we design a hierarchical graph to integrate patients, doctors, and topics.

Hierarchical graph notations. This hierarchical graph can be formalized as $G = (V, E, \mathbf{X})$, where V is the node set (i.e., patients, doctors, and topics), E is the edge set to connect the node pairs, and $\mathbf{X}$ is the initial feature matrix. More specifically, G can be further refined into three subgraphs: $G = \{G_p, G_d, G_t\}$. G_p is the patient graph with nodes V_p and edges E_p, G_d is the doctor graph with nodes V_d and edges E_d, and G_t is the topic graph with nodes V_t and edges E_t. In addition, E_{pd} connects patient graph and doctor graph when patients and doctors are associated with national patient identifiers (NPIs), and E_{dt} connects doctor graph and topic graph when doctors view the website topics.

Node representations. The node feature matrix $\mathbf{X}$ is composed of three matrices $\mathbf{X}_p$, $\mathbf{X}_d$, and $\mathbf{X}_t$ such that $\mathbf{X} = \{\mathbf{X}_p, \mathbf{X}_d, \mathbf{X}_t\}$, where $\mathbf{X}_p$, $\mathbf{X}_d$, and $\mathbf{X}_t$ embed the feature spaces for patients, doctors, and topics respectively. Each patient feature vector $\mathbf{x}_p$ is initialized as $\mathbf{x}_p = \langle x_{p1}, x_{p2}, x_{p3}, \cdots, x_{pk} \rangle$, where $x_{pi} \in \{0, 1\}$ is a binary value indicting the absence or presence of a disease symptom i in patient $\mathbf{x}_p$. Each doctor $\mathbf{x}_d$ is represented as a set of profile attributes, where each attribute is directly converted into numerical feature values using one-hot encoding. Each topic $\mathbf{x}_t$ is represented as either a word or a phrase; in this regard, we leverage SBERT [28] to derive a fixed-size embedding for each topic. In order to keep the dimensionality of all nodes consistent for message passing yet the dimension of $\mathbf{x}_d$ and $\mathbf{x}_t$ is smaller than $\mathbf{x}_p$, we zero-pad $\mathbf{x}_d$ and $\mathbf{x}_t$ to be k-dimensional, and hence the node feature matrix $\mathbf{X} \subseteq \mathbb{R}^{(l+m+n)\times k}$.

Patient graph. Given a set of patient records $\mathcal{X}_p$, we construct a fully-connected graph $G_p = (V_p, E_p, \mathbf{X}_p)$ to associate patients (both labeled and unlabeled). Manifold learning [23] is non-linear dimensionality reduction process which reveals the low-dimensional manifold embedded in the high-dimensional space, which can be feasibly exploited to build up the intrinsic neighborhood among patient representations. Thus, we formulate each edge $e_p \in E_p$ between v_{pi} and

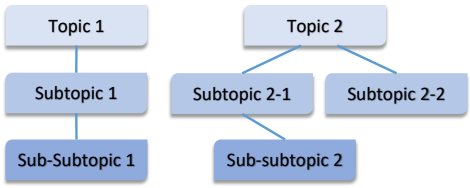


Fig. 2. Topic dependency in the topic graph.

v_{pj} in G_p by a layerwise non-linear combination of distance between $\mathbf{x}_{pi}$ and $\mathbf{x}_{pj}$:

$$e_p = g_{\boldsymbol{\Theta}}(\mathbf{x}_{pi}, \mathbf{x}_{pj}) = \sigma(\cdots\sigma(|\mathbf{x}_{pi} - \mathbf{x}_{pj}|\boldsymbol{\Theta}^{(0)})\cdots\boldsymbol{\Theta}^{(L-1)})\boldsymbol{\Theta}^{(L)} \tag{2}$$

where $\sigma(\cdot)$ is a non-linear activation function (e.g., ReLU), and $\boldsymbol{\Theta}$ is learnable weight matrix for each layer. As the constructed structure behaves differently regarding different patient representations, the learned edges do not specify a fixed patient graph, suggesting the graph can be refined when the embedding space across patient nodes is updated.

Doctor graph. In addition to doctors' profile attributes, our collected doctor data $\mathcal{X}_d$ also record the numbers of patients shared with other doctors, which can be directly exploited to build the doctor graph. To be specific, if two doctors v_{di} and v_{dj} share greater than or equal to one patient in common, we create an edge $e_d \in E_d$ between v_{di} and v_{dj} in G_d, such that the doctor graph $G_d = (V_d, E_d, \mathbf{X}_d)$ can be easily derived with fixed structure. Afterwards, the doctor nodes V_d are further associated with the patient nodes V_p through $e_{pd} \in E_{dp}$ when the doctor v_d is the patient v_p's primary care doctor identified by NPI.

Topic graph. To better characterize doctors, we integrate doctors' viewing activities into their profile attributes for doctor presentation learning, where these activities are captured by the website topics viewed by doctors. To this end, we build a topic graph $G_t = (V_t, E_t, \mathbf{X}_t)$ to model this data. As demonstrated in Figure 2, all the topics are organized through layer-wise dependency; for example, a topic may contain another one or more subtopics, where some other topics may be listed under a subtopic. An edge $e_t \in E_t$ between v_{ti} and v_{tj} in G_t can be thus formulated when v_{tj} is v_{ti}'s subtopic. Naturally, the topic nodes V_t can be associated with the doctor nodes V_d through their viewing records.

3.2 Hierarchical Graph Neural Network with Dual Message Passing

Considering the constructed hierarchical graph with intra-subgraph and inter-subgraph neighborhood structures, we propose a hierarchical graph neural network to perform the dual message passing for node representation refinement and label propagation, including intra-message passing and inter-message passing.

Intra-message passing. Intra-message passing is the propagation mechanism that aggregates the information from neighbors inside the patient graph, doctor graph, and topic graph, respectively, the data flow paths of which are specified

as black lines in Figure 1. A regular graph convolutional network (GCN) [16] is implemented for a single subgraph. Specifically, given a subgraph (i.e., patient graph, doctor graph, or topic graph), we build the adjacency matrix $\mathbf{A}^{(h)}$ using its edge information (edge matrix needs to be normalized first for patient graph). The message passing can be then formalized as multi-layer neighborhood information aggregation, which receives an input $\mathbf{X}^{(h)}$ and produces $\mathbf{X}^{(h+1)}$:

$$\mathbf{X}^{(h+1)} = \sigma(\widetilde{\mathbf{A}}^{(h)}\mathbf{X}^{(h)}\mathbf{W}^{(h)}_{intra}) \tag{3}$$

where at layer h, $\mathbf{W}_{intra}$ is weight matrix, $\widetilde{\mathbf{A}} = \mathbf{D}^{-\frac{1}{2}}\hat{\mathbf{A}}\mathbf{D}^{-\frac{1}{2}}$, $\hat{\mathbf{A}} = \mathbf{A} + \mathbf{I}$, and $\mathbf{D}$ is the diagonal degree matrix defined on $\hat{\mathbf{A}}$, i.e., $\mathbf{D}_{ii} = \sum_{j=1}^{n} \hat{\mathbf{A}}_{ij}$.

Inter-message passing. Inter-message passing mechanism is used to propagate the information between two subgraphs, including patient-doctor and doctor-topic neighborhoods in our hierarchical graph, the data flow paths of which are specified as red lines in Figure 1. Similarly, a GCN is implemented for a single inter-subgraph neighborhood, where an adjacency matrix $\mathbf{A}^{(h)}$ is first constructed based on the node set from both subgraphs and the edge set (i.e., E_{pd} or E_{dt}) connecting subgraphs, and then the message passing is performed:

$$\mathbf{X}^{(h+1)} = \sigma(\widetilde{\mathbf{A}}^{(h)}\mathbf{X}^{(h)}\mathbf{W}^{(h)}_{inter}) \tag{4}$$

where at layer h, $\mathbf{W}_{inter}$ is weight matrix for inter message passing. Different from intra-message passing, we do not add self-loops to the adjacency matrix in Eq. (4) to allow better aggregation of heterogeneous information.

Optimization. With dual message passing from topic graph to doctor graph, and then from doctor graph to patient graph, the output of the final GCN layer for intra-message passing over the patient graph can be defined as:

$$\mathbf{Z} = f_{\mathbf{W}}(\mathbf{A}, \mathbf{X}_p) = \text{softmax}(\mathbf{X}_p^{(H)}) \tag{5}$$

where $\mathbf{W}$ refers to the complete trainable weights raised by intra- and inter-message passing. Therefore, the optimization of hierarchical GNN model can be formulated to minimize the training loss as follows:

$$\mathbf{W}^* == \underset{\mathbf{W}}{\text{argmin}}\ \mathcal{L}(\mathbf{Z}, \mathbf{y}) + \lambda\|\mathbf{W}\|_2^2 \tag{6}$$

where $\mathcal{L}$ is the cross-entropy loss, and λ is the regularization parameter. This model can be applied under inductive and transductive settings. In this paper, we focus on transductive patient treatment preference prediction where all node connections and features are accessible during training.

3.3 Community Detection for Data Imbalance

As discussed in Section 1, another significant challenge for patient treatment preference prediction is the EMR data imbalance issue. This enforces GNN models to aggregate information from majority-class nodes and become less sensitive

to under-represented positive samples, which leads to less-accurate prediction performance. Accordingly, different paradigms have been presented to address this issue, such as oversampling [1, 14], undersampling [2, 17], and cost-sensitive learning [21, 13]. Due to the fact that sampling techniques tend to generate models with relatively low generalizability that either overfit on oversampled data or underperform for discarding potentially useful data, and cost-sensitive learning is easily impacted by weights, making it hard to select optimal cost values, these methods are still limited for our task.

In this paper, we explore a community detection method to cope with imbalanced patient data. The motivations behind this choice are that: (1) community detection [34, 27] is one of the widely used approaches to analyze complex networks involving social interactions, which is perfectly applicable to the patient graph; (2) individuals are known by the community they keep, while patients in EMRs are natural individuals whose treatment behaviors and preferences can be represented by a group of others with very similar symptoms; and (3) community detection works as undersampling but can effectively mitigate the impact of information loss [20]. More specifically, the proposed community detection method to address data imbalance consists two steps:

- **Detecting communities**. If we start community detection over the graph using the initial patient representations, we need to traverse the graph to reveal the community structure using algorithms such as infomap [30]. Instead, here we follow the strategy to first learn the higher-level patient representations using hierarchical GNN with dual message passing to embed semantics from patients and doctors, and abstract graph structure, such that we can then simply apply standard clustering algorithm such as k-means to cluster the embeddings of negative patients into K distinct communities, where K is equal to the number of positive patients, and cluster the embeddings of unlabeled patients into N communities, where N is dependent on test data size (we evaluate the impact of N on prediction performance in Section 4.4). Afterwards, all the edges ending with the patient nodes in a community are adjusted to be connected with this community as one node.
- **Training GNN using community-preserving patient graph**. With the new community-preserving patient graph, we continue performing dual message passing over hierarchical graph and train the hierarchical GNN by minimizing the cross-entropy loss in Eq. (6). During testing, the prediction label of a community node will be assigned to all patients in this community.

4 Experiments and Results

4.1 Experiment Setup

Datasets. Our experiments are tested on EMRs for CLL patients with doctor data and website topics provided by IQVIA. The patient data retain patients' records including their different symptom features and NPIs for their primary care doctors. The doctor data include doctors' profiles (e.g., age, gender, location,

Table 1. Statics of datasets

Dataset	#Distinct data	#Features	#Positives	#Negatives
Patients	93,474	2,016	773	92,701
Doctors	2,134	112	-	-
Website Topics	300	-	-	-

etc.) and the number of patients shared with other doctors. The topic data contain topics of websites viewed by doctors. The data statistics are shown in Table 1, illustrating that there are 93,474 patients (773 positives and 92,701 negatives), 2,134 valid doctors, and 300 website topics, respectively.

Baselines. To the best of our knowledge, we are the first to predict patient treatment preference; no previous work can thus be used as baselines. We select rare disease prediction models, traditional classification models, GNN models, and imbalanced learning models as our baselines. Note that, for GCN, GRAM, and RA-GCN designed for single graph, we only use the patient graph as input.

- **Support Vector Machine (SVM)**: This is one of the supervised learning methods which can be used to find a hyperplane for classification.
- **Random Forest (RF)**: This is an ensemble learning method for classification by constructing a number of decision trees.
- **Multi-Layer Perceptron (MLP)**: This is a fully connected class of artificial neural network, which is a traditional supervised classification model.
- **Graph Convolutional Network (GCN)** [16]: This is a semi-supervised learning model on graph-structured data with graph convolutional layers.
- **GRAM** [8]: GRAM is a graph-based attention model for healthcare representation learning, which leverages graph attention network [33] to get the information from neighbors with different importance. We use their attention mechanism to build the graph neural network and set it as our baseline.
- **HSGNN** [24]: Heterogeneous similarity GNN is designed for heterogeneous graphs with healthcare data. We reconstruct the graph with our patient and doctor data and set their model as our baseline.
- **Oversampling** [14]: This is an approach to deal with imbalanced data by increasing the minority class in the dataset. In our experiments, we simply add the minority class repeatedly for oversampling.
- **Undersampling** [17]: This is an approach to deal with imbalanced data by randomly removing the data from the majority class in the dataset.
- **XGBoost** [6]: It is one of the state-of-the-art and widely used machine learning models. It becomes the powerful machine learning model of many data scientists and can deal with irregularities of data, which has been justified as one of the most popular methods for dealing with imbalanced data.
- **Pseudo-labeling** [36]: This approach generates pseudo-labels from unlabeled data, which are injected into training data to address data imbalance.
- **RA-GCN** [15]: It sets different weights to different classes to address the data imbalance problem with GCN for disease prediction.

Table 2. Evaluation results with different baselines

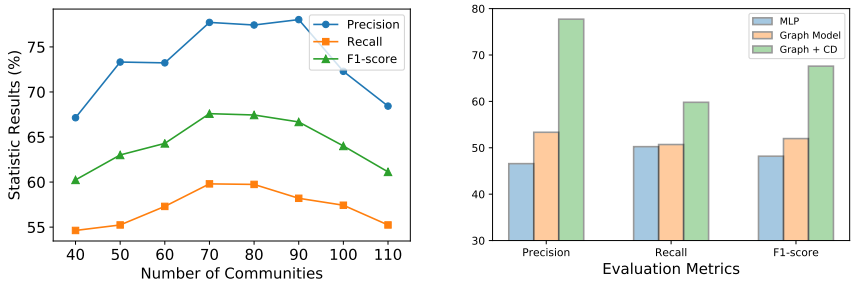
Model	Precision (%)	Recall (%)	F1-score
Support Vector Machine	45.60	50.00	0.4769
Random Forest	45.73	50.10	0.4781
Multi-Layer Perceptron	46.30	50.24	0.4818
GCN	45.64	49.88	0.4744
GRAM	46.70	50.00	0.4952
HSGNN	46.58	48.20	0.4738
Oversampling	75.28	50.21	0.6024
Undersampling	23.40	50.03	0.3188
XGBoost	52.28	59.24	0.5554
Pseudo-labeling	53.41	51.86	0.5262
RA-GCN	47.22	42.46	0.4471
Our Model	**77.72**	**59.80**	**0.6759**

4.2 Data Preprocessing and Setting

We preprocess the data by filtering out those patient records whose primary doctors cannot be found in the doctor data and removing doctors whose profiles are missing, which leads to 19,176 patients (351 positives) and 2,134 doctors left, respectively. Due to resource limitation, we select all positives and 3,510 negatives as experimental data and randomly split it by 8:2 for training and testing. We use precision, recall, and F1-score as evaluation metrics, which are typically used for healthcare data. All the GCNs used for intra- and inter-message passing are set as a two-layer structure with 16 hidden units.

4.3 Comparisons with Baselines

In this section, we compare our model with the selected baselines. The competitive result is illustrated in Table 2. We can observe that among baselines, the traditional models (i.e., SVM, RF, and MLP) have the worst performance on the patient treatment preference prediction with single patient data input. GNN models perform slightly better than traditional ones with the precision increases by around 1%, which is limited for data imbalance issue. Most of the models dealing with imbalanced data achieve better performance than others, where oversampling delivers the best results, XGBoost also provides some promising performance boost, but undersampling significantly underperforms for information loss. Obviously, our model completely outperforms baselines with a large improvement margin of precision (2% - 30%), recall (0.6% - 11%), and F1-score (0.07 - 0.36). This confirms that (1) doctor information and viewing activities can serve as external knowledge to enrich patient representations; (2) community detection performed on negatives and unlabeled patients can effectively mitigate the data imbalance issue and better prediction performance.



(a) **(b)**

Fig. 3. Evaluation on model: (a)Number of Communities N (b) Different components

4.4 Impact of Community Number over Test Data

The community number K over negatives is decided by the number of positives, but the community number N over test data is adjustable. In this section, we analyze the impact of N on prediction performance. The results are shown in Figure 3(a): the prediction results increase when N falls in the range of [40,70], then keep stable at [70,90], and decrease drastically when N rises to 110. The reason behind this could be that when $N \in [70, 90]$, the positive data and negative data are in a relative equilibrium, which alleviates the data imbalance impact and prevents the model from favoring any majority class; when N is too small, some unique patients tent to get misrepresented by communities; when N is too large, the imbalance issue may emerge to degrade the predictions.

4.5 Ablation Study

In this section, we conduct the ablation study to evaluate the performance contributed by different design parts. As our model proceeds with (1) hierarchical graph construction and (2) community detection, we construct three alternative models: (1) MLP: feeds patient features directly to MLP without any graph or community detection; (2) Graph Model: applies the hierarchical graph with dual message passing; (3) Graph + Community Detection (Graph + CD): leverages community detection to build community-preserving graph and trains the hierarchical GNN over that. The results are reported in Figure 3(b).

As shown in Figure 3(b), the performance becomes better with the components added to the model. The graph model increases the precision, recall, and F1-score from multi-layer perceptron by about 8%, 1%, and 4% respectively. Our model with hierarchical graph and community detection is able to further improve precision, recall, and F1-score to a higher level, which are around 77%, 60%, and 67% respectively. This reaffirms that the hierarchical graph enables doctor information and activities to be propagated to patients through dual message passing and increases the expressiveness of patient representations, while community detection successfully alleviates the effect of data imbalance and makes the model more effective on patient treatment preference prediction.

5 Conclusion

In this paper, we propose a novel hierarchical GNN for patient treatment preference prediction. We first leverage external knowledge (i.e., doctor information and their viewing activities) in addition to EMR patient records to construct the hierarchical graph, where a dual message passing paradigm is then devised to perform intra- and inter-subgraph neighborhood aggregation to enrich patient representations and advance label propagation. We further introduce community detection to alleviate patient data imbalance issue. The state-of-the-art results validate its effectiveness and superiority to the current widely used baselines.

References

1. Ando, S., Huang, C.Y.: Deep over-sampling framework for classifying imbalanced data. In: ECML-PKDD. pp. 770–785. Springer (2017)
2. Buda, M., Maki, A., Mazurowski, M.A.: A systematic study of the class imbalance problem in convolutional neural networks. Neural networks **106**, 249–259 (2018)
3. Cai, D., Sun, C., Song, M., Zhang, B., Hong, S., Li, H.: Hypergraph contrastive learning for electronic health records. In: SDM. pp. 127–135. SIAM (2022)
4. Chen, J., Ma, T., Xiao, C.: Fastgcn: fast learning with graph convolutional networks via importance sampling. arXiv preprint arXiv:1801.10247 (2018)
5. Chen, L., Li, X., Wu, D.: Enhancing robustness of graph convolutional networks via dropping graph connections. In: ECML-PKDD. pp. 412–428. Springer (2020)
6. Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: SIGKDD. pp. 785–794 (2016)
7. Chipidza, F.E., Wallwork, R.S., Stern, T.A.: Impact of the doctor-patient relationship. The primary care companion for CNS disorders **17**(5), 27354 (2015)
8. Choi, E., Bahadori, M.T., Song, L., Stewart, W.F., Sun, J.: Gram: graph-based attention model for healthcare representation learning. In: SIGKDD (2017)
9. Choi, E., Xu, Z., Li, Y., Dusenberry, M., Flores, G., Xue, E., Dai, A.: Learning the graphical structure of electronic health records with graph convolutional transformer. In: AAAI. vol. 34, pp. 606–613 (2020)
10. Chu, J., Dong, W., Wang, J., He, K., Huang, Z.: Treatment effect prediction with adversarial deep learning using electronic health records. BMC MIDM (2020)
11. Cui, L., Biswal, S., Glass, L.M., Lever, G., Sun, J., Xiao, C.: Conan: complementary pattern augmentation for rare disease detection. In: AAAI (2020)
12. Dossa, F., Chesney, T.R., Acuna, S.A., Baxter, N.N.: A watch-and-wait approach for locally advanced rectal cancer after a clinical complete response following neoadjuvant chemoradiation: a systematic review and meta-analysis. The lancet Gastroenterology & hepatology **2**(7), 501–513 (2017)
13. Elkan, C.: The foundations of cost-sensitive learning. In: IJCAI. vol. 17, pp. 973–978. Lawrence Erlbaum Associates Ltd (2001)
14. Fernández, A., Garcia, S., Herrera, F., Chawla, N.V.: Smote for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. Journal of artificial intelligence research **61**, 863–905 (2018)
15. Ghorbani, M., Kazi, A., Baghshah, M.S., Rabiee, H.R., Navab, N.: Ra-gcn: Graph convolutional network for disease prediction problems with imbalanced data. Medical Image Analysis **75**, 102272 (2022)

16. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907 (2016)
17. Lee, W., Seo, K.: Downsampling for binary classification with a highly imbalanced dataset using active learning. Big Data Research **28**, 100314 (2022)
18. Li, Q., Li, X., Chen, L., Wu, D.: Distilling Knowledge on Text Graph for Social Media Attribute Inference. In: SIGIR. pp. 2024–2028 (2022)
19. Li, X., Chen, L., Wu, D.: Turning attacks into protection: Social media privacy protection using adversarial attacks. In: SDM. pp. 208–216. SIAM (2021)
20. Lin, W.C., Tsai, C.F., Hu, Y.H., Jhang, J.S.: Clustering-based undersampling in class-imbalanced data. Information Sciences **409**, 17–26 (2017)
21. Ling, C.X., Sheng, V.S.: Cost-sensitive learning and the class imbalance problem. Encyclopedia of machine learning **2011**, 231–235 (2008)
22. Liu, R., Wei, L., Zhang, P.: A deep learning framework for drug repurposing via emulating clinical trials on real-world patient data. Nature machine intelligence **3**(1), 68–75 (2021)
23. Liu, Y., Lee, J., Park, M., Kim, S., Yang, E., Hwang, S.J., Yang, Y.: Learning to propagate labels: Transductive propagation network for few-shot learning. arXiv preprint arXiv:1805.10002 (2018)
24. Liu, Z., Li, X., Peng, H., He, L., Philip, S.Y.: Heterogeneous similarity graph neural network on electronic health records. In: IEEE Big Data (2020)
25. Ma, F., Wang, Y., Gao, J., Xiao, H., Zhou, J.: Rare disease prediction by generating quality-assured electronic health records. In: SDM. pp. 514–522. SIAM (2020)
26. Ma, F., You, Q., Xiao, H., Chitta, R., Zhou, J., Gao, J.: Kame: Knowledge-based attention model for diagnosis prediction in healthcare. In: CIKM (2018)
27. Papadopoulos, S., Kompatsiaris, Y., Vakali, A., Spyridonos, P.: Community detection in social media. Data mining and knowledge discovery **24**(3), 515–554 (2012)
28. Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. arXiv preprint arXiv:1908.10084 (2019)
29. Ross, M.K., Yoon, J., van der Schaar, A., van der Schaar, M.: Discovering pediatric asthma phenotypes on the basis of response to controller medication using machine learning. Annals Of The American Thoracic Society **15**(1), 49–58 (2018)
30. Rosvall, M., Bergstrom, C.T.: Maps of random walks on complex networks reveal community structure. Proceedings of the national academy of sciences (2008)
31. Saqib, M., Sha, Y., Wang, M.D.: Early prediction of sepsis in emr records using traditional ml techniques and deep learning lstm networks. In: EMBC (2018)
32. Segura-Bedmar, I., Colón-Ruíz, C., Tejedor-Alonso, M.Á., Moro-Moro, M.: Predicting of anaphylaxis in big data emr by exploring machine learning approaches. Journal of biomedical informatics **87**, 50–59 (2018)
33. Thekumparampil, K.K., Wang, C., Oh, S., Li, L.J.: Attention-based graph neural network for semi-supervised learning. arXiv preprint arXiv:1803.03735 (2018)
34. Yang, J., McAuley, J., Leskovec, J.: Community detection in networks with node attributes. In: ICDM. pp. 1151–1156. IEEE (2013)
35. Yang, J., Liu, Y., Qian, M., Guan, C., Yuan, X.: Information extraction from electronic medical records using multitask recurrent neural network with contextual word embedding. Applied Sciences **9**(18), 3658 (2019)
36. Yang, Y., Xu, Z.: Rethinking the value of labels for improving class-imbalanced learning. NeurIPS **33**, 19290–19301 (2020)
37. Zhao, J., Gu, S., McDermaid, A.: Predicting outcomes of chronic kidney disease from emr data based on random forest regression. Mathematical biosciences (2019)