

Disease Prediction with Heterogeneous Graph of Electronic Health Records and Toxicogenomics Data

Ji-Hyeong Park¹^a, Hyun-Soo Choi^{2,*}^b, Sunhwa Jo³^c and Jinho Kim^{3,*}^d

¹Dept. of Convergence Security, Kangwon Nat'l Univ., Chuncheon, Republic of Korea

²Dept. of Computer Science and Engineering, Seoul Nat'l Univ. of Science and Technology, Seoul, Republic of Korea

³Dept. of Computer Science and Engineering, Kangwon Nat'l Univ., Chuncheon, Republic of Korea

Keywords: Heterogeneous Graph, Node Embedding, Disease Prediction, Electronic Health Records, Toxicogenomics Data.


Abstract: Disease prediction is an important technology in the field of medicine. Several studies have been conducted on disease prediction using electronic health records (EHR). However, existing methods have several limitations, such as predicting only a single disease and utilizing limited data sources of textual or drug-related data; thus, they cannot capture the relationship between a patient and a disease, or among diseases. Furthermore, they suffer from the problem that additional information other than EHR exists only for a limited set of diseases and cannot be used for a wide range of diseases. To mitigate these problems, we utilize Toxicogenomics Data (TD) that contains extensive information about most diseases, and analyze this complicated data using a heterogeneous graph embedding technique. We utilize metapath and graph neural network for graph embedding of heterogeneous relationships in EHR-TD, and then develop a novel disease prediction framework. To achieve this goal, we first present a process for the collection and processing of EHR and TD data to improve their reliability. Secondly, we propose a method for efficiently constructing heterogeneous EHR-TD graphs, and present an embedding model that can be effectively used. Finally, we propose a metapath interaction encoder that can address the problems of RNN-based encoders in previous models. Thereafter, we validate the effectiveness of the proposed framework and modules with extensive evaluations of various designs for disease prediction using EHR and TD data.


1 INTRODUCTION


Recent advancements in medical technology have led to the generation of large amounts of high-quality data in hospitals. These data are commonly referred to as electronic health records (EHRs) and typically include a wide range of information, such as visit and hospitalization records, symptoms, and patient information. Due to the wealth of valuable information contained in EHRs, there has been a rapid increase in the number of data analysis studies utilizing EHRs for medical applications, such as hospital stay prediction (Gentimis et al., 2017), diagnostic prediction (Ma et al., 2018), and mortality prediction (DeSalvo et al.,


2006). Among these studies, disease prediction is a valuable technology that can reduce the cost of medical care. In existing studies on disease prediction, gene-based approaches have been used, which are limited in accessibility due to the high cost and resource requirements. Therefore, we aim to develop a disease prediction method that can be used with reasonable costs and resources by utilizing EHR data generated by patients visiting a hospital.

Existing disease prediction studies address specific disease prediction using only EHR (Jin et al., 2018), which contains textual information, including user examination records and disease predictions using drug information databases, such as PubMed and DrugBank. However, these disease predictions do not employ sufficient information for disease prediction, although they utilize additional data, such as images, text, or drug-related data. Furthermore, these additional data cannot be consistently used for all diseases because their presence depends on the characteristics

^a <https://orcid.org/0009-0003-7817-1774>

^b <https://orcid.org/0000-0002-3594-8948>

^c <https://orcid.org/0000-0002-6696-6276>

^d <https://orcid.org/0000-0003-1125-3938>

*Corresponding authors

of each disease. In addition, existing methods do not capture the relationship between a patient and disease or among different diseases.

Our primary goal is to achieve a more accurate disease prediction that is practical, as well as solve the problem in which the existing methods are available to limited diseases and show limited accuracy owing to a lack of information for prediction. To supplement information about relatively rare diseases, we introduced Toxicogenomics Data (TD), which includes various environmental factors affecting health-care. TD contains a variety of information about disease-related genes and chemicals, which is useful for inferring information about the disease of a patient as well as other diseases similar to a particular disease. As TD appears in most diseases, it does not depend on the characteristics of a disease; thus, TD can be actively utilized.

To understand the complex structure that combines EHR and TD, we adopted a graph-based embedding approach. A graph-based approach can effectively model the data structure in which each object has similar information, and multiple relationships between any two objects, as validated in social graph analysis (Leskovec and McAuley, 2012) and protein-protein graph analysis (Nasiri et al., 2021). In our work, to cope with the heterogeneous objects of patients and diseases, we proposed a heterogeneous graph and heterogeneous node embedding model via metapath design. In addition, we proposed a metapath interaction encoder that can be utilized in the proposed node-embedding model. Through experimental evaluations of a couple of graph configurations and node embedding methods, we suggested a promising framework for disease prediction for a patient and validated the effectiveness of the proposed framework and encoder.

The contributions of this study are summarized as follows.

- We present a novel framework to achieve further accurate disease prediction that can be applied to different diseases.
- We construct a reliable heterogeneous graph data structure that represents the complex data structure and relationship among objects in EHR and TD.
- We suggest a heterogeneous graph embedding model with a metapath interaction encoder to learn the EHR-TD graph data effectively.
- We achieve outperforming performance to existing approach without using complex data, such as images and text.

2 PROPOSED METHOD

In this study, we proposed an EHR-TD combined graph-utilized disease prediction framework to supplement patient disease information. The proposed framework consists of four steps: Data Construction, Graph Configuration, Graph Embedding, and Prediction Process. Here, we introduce each step and describe how it is performed. Figure 1 illustrates the overall process of the proposed framework.

2.1 Data Construction

In this study, we extracted information from two databases: MIMIC-III (Johnson et al., 2016) and CTDBase (Davis et al., 2020). This section describes how the data used in the experiments were extracted and pre-processed from the two databases.

The ADMISSIONS table in MIMIC-III, which contains information on patients, such as insurance status and religion, has been used in addition to the disease that the patient suffers from. Before extracting patient information, we first used ADMITTIME and DISCHTIME in the table to calculate the patient's visit time and discharge time to obtain hospitalization periods and remove records with hospitalization periods of less than 1 d. Thereafter, disease prediction was performed only for patients who appeared more than twice, based on SUBJECT_ID representing the patient.

CTDBase includes genes and chemicals associated with the disease that appear in MIMIC-III. To collect disease-related genes or chemicals from CTDBase, the disease MeshID is required. Therefore, we first performed text-to-MeshID matching to convert a disease name(text) in the MIMIC-III table into MeshID in CTDBase. However, the disease presented in the ADMISSIONS table of MIMIC-III has not been pre-processed, such as abbreviated words or containing words that are not commonly used. For more accurate text-to-MeshID matching, we pre-processed each disease name, and then matched it to a MeshID. When a search term(text) is entered, the Mesh Browser provides approximate information about the disease and MeshID for the search term, and slightly corrects the synonyms. The MeshID corresponding to the disease name was collected using Web crawling. Thereafter, text-to-MeshID matching was performed using a text search. Because the disease expression between MIMIC-III and CTDBase is inconsistent, they may not be perfectly matched. For further accurate matching, two types of text pre-processing were performed, without using text information. First, all special characteristics were re-

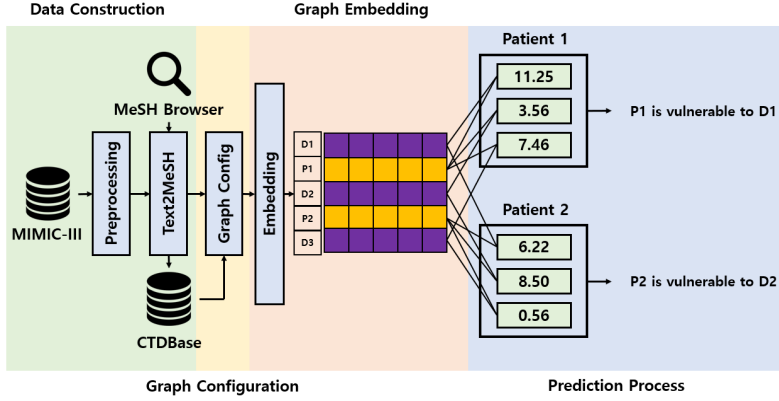


Figure 1: Overall process of the proposed framework.

moved, and second, if there were several diseases in one column, they were separated.

When processing was completed, we collected genes and chemicals associated with MeshID. We hypothesized that many genes and chemicals, compared to the number of diseases, would not be helpful in disease prediction. Thus, when collecting Gene and Chemical, only up to 200 genes and chemicals were collected for each. When there were fewer than 20 associated Genes and Chemicals, we did not collect them. Additionally, the disease was collected only if the disease name exactly coincided with one another.

Finally, we added features to the collected objects. Among the above-described objects, Gene did not have essential information to be used as a feature; thus, the feature was assigned by a One-Hot Vector, and the feature was added as follows for patients, diseases, and chemicals.

- Patient: Patient information in the ADMISSIONS table was encoded into a multi-hot vector as a feature vector, depending on the specific patient information (admission type, admission location, discharge location, insurance status, religion, marital status, and race).
- Disease and Chemical: The words in Definition of CTDBase were encoded into a multi-hot vector as a feature. When collecting Definition, the feature of an object without Definition was assigned by a zero vector.

2.2 Heterogeneous Graph Configuration

In the previous section, we described the data collection and processing for EHR and TD, respectively. This section describes the configuration of the EHR-TD combined graph required for the proposed framework.

For disease prediction, a graph provides similarity between a patient’s outbreak disease and other diseases. We define the heterogeneous graph G as follows: G has two or more node types, including vertex V , edge E , and type T : $G = (V, E, T)$. Each node v in the graph can have heterogeneous neighbor nodes connected along the edges in E , and the set of neighbor nodes \mathcal{N}_v for v is defined as $\mathcal{N}_v = \{v_n | (v, v_n) \in E; v, v_n \in V\}$.

We first add the node types of the patient (P) and disease (D) to reflect the patient’s outbreak information in the graph. The similarity between any two disease nodes can be obtained indirectly from the disease node features or outbreak information of similar patients. This indirectly obtained similarity might degrade the performance of the disease prediction. Therefore, we used TD to provide the similarity in the graph. We added node types, such as genes (G) and chemicals (C) in TD to the graph, and the diseases sharing the types were regarded as similar diseases to each other. Finally, the type T_v of each node v in the heterogeneous graph is defined as $T_v \in \{P, D, G, C\}, v \in V$. Edge type $T_e, e \in E$ for a heterogeneous graph is defined as $T_e \in \{O, R_G, R_C\}, e \in E$, where $O = \text{Occur}(P-D)$ denotes the relationship between a patient and a disease, $R_G = \text{RelatedGene}(D-G)$ denotes the relationship between a disease and a gene, and $R_C = \text{RelatedChemical}(D-C)$ denotes the relationship between a disease and chemical.

2.3 Heterogeneous Graph Embedding

2.3.1 Metapath Feature Transform

A data sample, such as an image or text, can be easily vectorized into a point in the Euclidean space. However, because computers cannot analyze (or calculate) complex non-Euclidean data, such as graph-structured data, a process of vectorizing graphs is re-

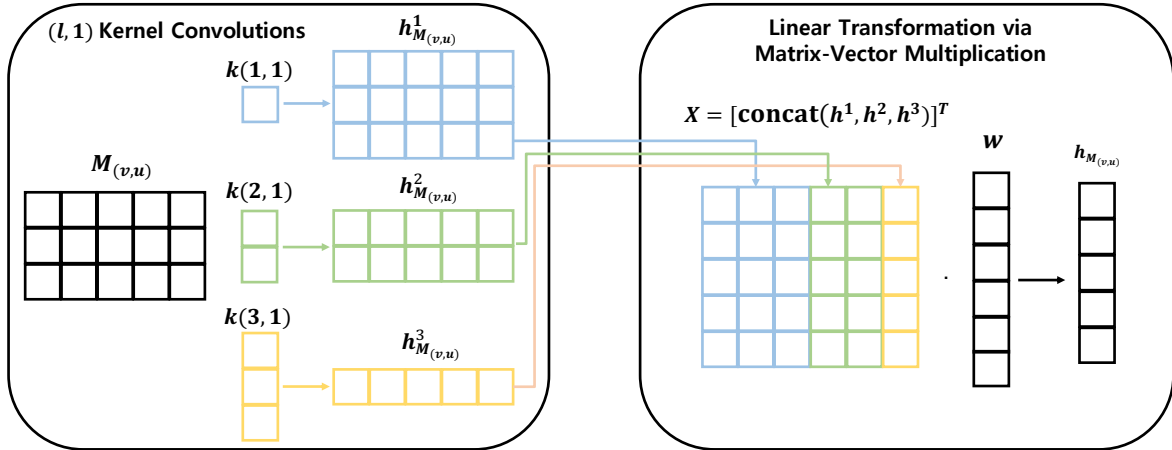


Figure 2: Entire operation of metapath interaction encoder.

quired. In addition, the heterogeneous graph we target can express more information compared to a homogeneous graph; however, there are many factors to be considered when constructing a heterogeneous graph, such as node types and relations among heterogeneous nodes. Therefore, when vectorizing a graph, diverse methods can be used, depending on the graph structure. In this study, we designed a heterogeneous graph structure and node-embedding scheme for disease prediction based on MAGNN(Fu et al., 2020), which uses metapath and graph neural network (GNN).

In a heterogeneous graph, the shape of the features of each node is different from that of the other nodes. For a node $v \in V$ of type $a \in A$, encoding is conducted through a single-layer neural transformation as

$$h'_v = W_a \cdot x_v^a, \quad (1)$$

where x_v is the input feature vector and W_a indicates a trainable matrix for type $a \in A$.

Thereafter, we generated a metapath instance to apply heterogeneous graph data to an embedding model. The metapath M is defined by a serial path of user-specified node types, as follows:

$$M = \{T_{v_1} - T_{v_2} - \dots - T_{v_n}\}, \quad T_{v_i} \in T_V, \quad (2)$$

where the start node v_1 and end node v_n are of the same type. metapath is used in most heterogeneous graph-embedding models. In our study, we defined four types of metapath. M_1, M_2 are defined to reflect the patient-disease relationship to embedding, Patient-Disease-Patient ($P-D-P$), and Disease-Patient-Disease ($D-P-D$). Furthermore, M_3, M_4 are defined to reflect the relationship between disease and gene, and disease and chemical to embedding, disease-gene-disease ($D-G-D$), and disease-chemical-disease ($D-C-D$), respectively.

In the graph, every metapath instance $M_{(v,u)}$ with the start node v , and end node u has one of four metapath types. $M_{(v,u)}$ was created by stacking the features in Eq. (1) as follows:

$$M_{(v,u)} = \text{stack}(h'_v, \{h'_t, \forall t \in \{m^{M(v,u)}\}\}, h'_u), \quad (3)$$

where $m^{M(v,u)}$ denotes the intermediate node of (v, u) . The metapath instances in (3) are expressed by a matrix that becomes an input to the GNN in MAGNN(Fu et al., 2020) for embedding. GNN consists of two stages: intra metapath aggregation and inter metapath aggregation. In our work, instead of intra metapath aggregation in MAGNN(Fu et al., 2020), we suggested a new encoder called metapath interaction encoder, and is described in the following subsection.

2.3.2 Metapath Interaction Encoder

The intra-metapath aggregation is a process of collecting information on metapaths of a target node into one vector. This process uses RNN structure in MAGNN(Fu et al., 2020). However, the RNN structure has a problem in that it can not fully utilize all the information appearing in metapath because of their sequential message passing. To mitigate this problem, we propose a CNN-based encoder called metapath interaction encoder. We applied a trainable kernel convolution operation to the input matrix $M_{(v,u)}$ in (3), where the $(i, j)^{th}$ element of the filtered result for $M_{(v,u)}$ is expressed as

$$h^l_{M_{(v,u)}}(i, j) = \sum_{q=1}^l M_{(v,u)}(i+q-1, j) \mathbf{k}(q, 1), \quad (4)$$

where $\mathbf{k} \in R^{l \times 1}$ denotes a trainable kernel, whereas $0 < l \leq L$ for L denotes metapath instance's length. Subsequently, the filtered matrices $h^l_{M_{(v,u)}}, 0 < l \leq L$,

are concatenated and transformed by matrix-vector multiplication of the concatenated matrix and a trainable vector $\mathbf{w} \in R^{\sum_{l=1}^L l \times 1}$ as

$$\mathbf{X} := [\text{concat}(\{h_{M(v,u)}^l, 0 < \forall l \leq L\})]^T, \quad (5)$$

$$h_{M(v,u)} = \mathbf{X}\mathbf{w}. \quad (6)$$

The entire operation of metapath interaction encoder is depicted in Figure 2.

2.3.3 Inter Metapath Aggregation

The vector $h_{M(v,u)}$ in (5) is input to the Inter Metapath Aggregation (IMA) process in MAGNN (Fu et al., 2020) as

$$h_{(v,u)} = \text{IMA}(h_{M(v,u)}), \forall v, u \in V, \quad (7)$$

where v is the target node and u is the node connected to v in the same cluster type. For embedding, we adopted the GNN in MAGNN (Fu et al., 2020), and considered the original paper for the detailed structure and hyperparameters. We employed the cross-entropy loss, which is denoted by

$$\begin{aligned} \text{Loss} = & - \sum_{(v,u) \in E} \log \sigma(\langle h_{(v,u)}, h_{(v,u)} \rangle) \\ & - \sum_{(v',u') \in E^c} \log \sigma(1 - \langle h_{(v',u')}, h_{(v',u')} \rangle), \quad (8) \end{aligned}$$

where $\sigma(\cdot)$ is the sigmoid function, E indicates a set of connected edges through a metapath instance, and E^c denotes the complement of E , that is, a set of unconnected edges through any metapath instance.

2.4 Prediction Process

The node vectors generated through node embedding become closer to each other as they become similar. Therefore, it can be observed that the dot product value between vectors indicates how similar the vectors are to each other. Through embedding, vectors are adjusted such that the distance between a patient and the disease the patient is suffering from or the distance between a disease and another disease with similar genes and chemicals is reduced. These embedding vectors that reflect the graph structure and graph feature can be used in various applications, such as link prediction and node classification. In this study, we used link prediction to predict a patient's likelihood of an outbreak of a specific disease.

Link prediction predicts the presence or absence of an edge that already exists or that will be created in the future in graph G . By predicting the link between a patient and disease, we can determine the probability of how risky a patient is for a particular disease.

For a patient node p , its similarity with a disease node d is obtained by $D = \langle h_p, h_d \rangle$, $p \in V_P, d \in V_D$, where V_P, V_D are the patient and disease node sets, respectively, and $\langle \cdot, \cdot \rangle$ denotes the inner product. The probability that p has disease d is obtained using the sigmoid function as follows: $\text{Prob}(p, d) = 1/(1 + e^{-D})$.

The larger the $\text{Prob}(p, d)$, the higher is the probability that disease d will occur in patient p . To evaluate the link prediction performance, we measured the prediction errors of the edge existence. In this evaluation, if $\text{Prob}(p, d)$ is higher than the threshold η , we determine that the edge between p and d is connected; otherwise, it is unconnected.

3 EXPERIMENTS

In this section, we show the validity of the proposed framework through a performance comparison according to three types of graph configurations and six types of embedding models.

3.1 Experiment Settings

The experimental settings were configured by applying three graph-configuration methods and six graph-embedding methods to demonstrate the validity of the proposed framework. The true-positive and false-positive rates can differ depending on the threshold η . Therefore, we used the area under the curve of the receiver operating characteristic curve (AUROC), and average precision score (AP).

3.1.1 Types of Graph Configuration

The configurations of the graph types are denoted according to the collected data and constructed features as follows:

- HoEHR-TD Graph: This graph has four node types (patient, disease, gene, chemical) and three edge types (patient-disease, disease-gene, disease-chemical). Although it has several node types, it is treated as homogeneous graph and graph embedding method suitable to be employed.
- BiEHR Graph: This has two node types (patient, disease) and one edge types (patient-disease). Because nodes can be divided into two groups, graph can be classified as bipartite graphs, and an embedding method specialized for bipartite graphs is used.
- HeEHR-TD Graph (Proposed): This is a heterogeneous graph proposed in this study. This graph

Table 1: Results of link prediction experiment.

Framework	10%		30%		50%		70%	
	AUC	PR	AUC	PR	AUC	PR	AUC	PR
HoEHR-TD + Deepwalk	0.3903	0.6325	0.4932	0.7059	0.6383	0.7994	0.8632	0.9288
HoEHR-TD + Node2Vec	0.3903	0.6349	0.4893	0.7033	0.6413	0.8013	0.8013	0.9290
BiEHR + BiNE	0.8599	0.8677	0.8806	0.8822	0.8917	0.8839	0.8876	0.8772
BiEHR + BiGI	0.8455	0.8101	0.8618	0.8198	0.8627	0.8220	0.8700	0.8357
HeEHR-TD + Metapath2Vec	0.4142	0.6520	0.5181	0.7236	0.6504	0.8071	0.8629	0.9287
HeEHR-TD + MAGNN	0.9763	0.9740	0.9900	0.9903	0.9907	0.9902	0.9907	0.9907
HeEHR-TD + MAGNN + MIE (Proposed)	0.9870	0.9862	0.9903	0.9900	0.9892	0.9894	0.9922	0.9923

has four node types (patient, disease, gene, chemical) and three edge types (patient-disease, disease-gene, disease-chemical). Because it has several node types and edge types, we designed a specialized graph embedding model suitable to this heterogeneous graph.

3.1.2 Type of Graph Embedding Methodologies

Six embedding models were selected based on the graph configuration. The model selected according to each graph configuration is described as follows. In this study, we designed a framework based on the MAGNN(Fu et al., 2020) model.

- Models for Homogeneous Graph
 - Deepwalk(Perozzi et al., 2014): This is a random walk-based homogeneous graph embedding method. This is to learn embeddings of nodes via random walks using skip-gram or c-bow methods.
 - Node2Vec(Grover and Leskovec, 2016): This is a random walk-based homogeneous graph embedding method. Similar to DeepWalk, but when generating a random walk, the probability of moving to a neighbor is adjusted by parameters p and q to achieve higher performance than DeepWalk.
- Models for Bipartite Graph
 - BiNE(Gao et al., 2018): This is a random walk-based bipartite graph embedding method. This attempts to learn the explicit relationship expressed by an edge, as well as the implicit relationship expressed by a transitive link that is not observed.
 - BiGI(Cao et al., 2021): This is a GCN-based bipartite graph embedding method. This model introduces local-global infomax to capture the global property.
- Models for Heterogeneous Graph
 - Metapath2Vec(Dong et al., 2017): This is a GCN-based bipartite graph embedding method.

This addresses a heterogeneous graph by generating random walks through the metapath, which is a list of predefined nodes.

- MAGNN(Fu et al., 2020): This is a model using both metapath and GCN used in the design of the proposed framework. The metapaths generated from the heterogeneous nodes are compressed into a single vector, and these vectors for a metapath are compressed into one vector, and used as the embedding of the starting node.
- MAGNN + MIE: This is a model that applied our proposed metapath interaction encoder. As a metapath interaction encoder, we attempted to utilize the interaction within metapath that could not be used in the existing encoders.

3.2 Experiment Result

Based on the experimental settings above, the six frameworks that were compared in our experiment are presented as follows:

- HoEHR-TD + (Deepwalk, Node2Vec): we design (Deepwalk, Node2Vec) for homogeneous EHR-TD graph.
- BiEHR + (BiNE, BiGI): we design (BiNE, BiGI) for Bipartite EHR graph.
- HeEHR-TD + (Metapath2Vec, MAGNN): we design (Metapath2Vec, MAGNN) for heterogeneous EHR-TD graph.
- HeEHR-TD + MAGNN + MIE: we design MAGNN using metapath interaction encoder for heterogeneous EHR-TD graph.

Table 1 lists the results of the link prediction experiment. The edges were divided using four training data ratios (10%, 30%, 50%, and 70%) and learned using them. Random walk-based models (Deepwalk, Node2Vec, BiNE, Metapath2Vec) did not predict the nodes that did not appear. Thus, when the training data ratio was low, the performance degradation was

Table 2: Metapath Comparison in metapath used model.

Metapath	Framework	10%		30%		50%		70%	
		AUC	PR	AUC	PR	AUC	PR	AUC	PR
M_1, M_2	Metapath2Vec	0.3959	0.6349	0.5115	0.7413	0.6417	0.7946	0.7847	0.8738
	MAGNN	0.4928	0.4949	0.5059	0.5068	0.4949	0.4967	0.4958	0.5054
M_1, M_2, M_3	Metapath2Vec	0.4281	0.6584	0.5298	0.7281	0.6673	0.8155	0.8778	0.9360
	MAGNN	0.9795	0.9806	0.9767	0.9757	0.9748	0.9741	0.9571	0.9551
M_1, M_2, M_4	Metapath2Vec	0.4229	0.6545	0.5308	0.7288	0.6640	0.8132	0.8689	0.9313
	MAGNN	0.9627	0.9642	0.9472	0.9472	0.9551	0.9573	0.9503	0.9533
M_1, M_2, M_3, M_4	Metapath2Vec	0.4213	0.6539	0.5299	0.7276	0.6673	0.8153	0.8839	0.9392
	MAGNN	0.9850	0.9864	0.9868	0.9869	0.9634	0.9623	0.9865	0.9883

large compared to that of the graph neural network-based models (BiGI and MAGNN). In addition, the model using our proposed metapath interaction encoder(MAGNN + MIE) showed better performance than before.

Comparing the results according to the graph configuration, it can be observed that the performance is roughly in the order of HeEHR-TD, HoEHR-TD, and BiEHR. BiEHR embedding models (BiNE and BiGE) do not reflect the proposed TD. HoEHR-TD embedding models (Deepwalk, Node2Vec) considered EHR and TD to be of the same type; therefore, the benefit from the added TD was relatively small. As shown in the table, we can observe that HeEHR-TD embedding models exhibit high performance by fully utilizing TD. In particular, the proposed framework, HeEHR-TD + MAGNN, outperformed the other frameworks.

Table 2 lists the performance variation depending on the design of metapaths. As shown in the table, the design of metapath significantly affected the prediction performance. The results in Table 2 show that the result using plenty of metapaths through the disease, gene, and chemical nodes is significantly better than that using only patient-disease metapaths. In metapath2vec, similar to the results in Table 1, the performance is similar at a low training data ratio. However, it can be observed that the higher the training data ratio, the better the results of using the metapath reflecting TD. In MAGNN, the results ($\{M_1, M_2\}$) show worse performance than the others, implying that simple metapaths cannot reflect sufficient relationships among heterogeneous node types. However, other results ($\{M_1, M_2, M_3\}$, $\{M_1, M_2, M_4\}$, and $\{M_1, M_2, M_3, M_4\}$) show that adding many relationships between a disease and gene, and a disease and chemical can contribute to heterogeneous node embedding.

3.3 Discussion

This study aimed to present a new framework for improving disease prediction performance by composing EHR-TD heterogeneous graphs on the relationships among patients, diseases, and genes/chemicals related to diseases that cannot be captured by existing disease prediction studies. This implies TD data can be helpful for disease prediction.

One of our contributions is that, by using toxicogenomics data, specific diseases as well as all diseases that appear in MIMIC-III (Johnson et al., 2016) are covered. As introduced in the section on data construction and graph embedding, the probability can be calculated for all diseases in MIMIC-III because the embedding vectors are created with edges for all diseases that have emerged through TD between diseases. Maximum prediction of diseases, rather than considering a single disease, will provide greater benefits to users.

4 CONCLUSION

In this study, we proposed and introduced an EHR-TD combined with a heterogeneous graph-based disease prediction framework to further improve disease prediction. Through the proposed framework, we aimed to maximally predict diseases, rather than the existing single disease prediction, to show that the combined heterogeneous data can improve disease prediction performance and present a heterogeneous graph structure that is effective for improving disease prediction performance. The proposed framework consists of data construction, which collects and pre-processes data, graph configuration, graph embedding, which creates representations for each node with constructed graphs, and a prediction process that uses representations of generated nodes. As contributions, we suggested a new heterogeneous graph representing EHR-

TD data, and designed a heterogeneous graph embedding model along with metapath design. The proposed frameworks were validated through a comparison with possible frameworks using combinations of graph configurations and embedding models. In addition, through ablation experiments, we demonstrated the usefulness of TD for disease prediction, and effects of the metapath design were investigated. Although the proposed framework shows outstanding performance compared to existing embedding models, further study for an enhanced embedding model specific to our EHR-TD data can be conducted in the future. We expect that the proposed framework will contribute to more accurate disease prediction and disease management in patients.

ACKNOWLEDGEMENTS

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. 2021R1F1A1059255).

REFERENCES

- Cao, J., Lin, X., Guo, S., Liu, L., Liu, T., and Wang, B. (2021). Bipartite graph embedding via mutual information maximization. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining, WSDM '21*, page 635–643, New York, NY, USA. Association for Computing Machinery.
- Davis, A. P., Grondin, C. J., Johnson, R. J., Sciaky, D., Wiegiers, J., Wiegiers, T. C., and Mattingly, C. J. (2020). Comparative Toxicogenomics Database (CTD): update 2021. *Nucleic Acids Research*, 49(D1):D1138–D1143.
- DeSalvo, K. B., Bloser, N., Reynolds, K., He, J., and Muntner, P. (2006). Mortality prediction with a single general self-rated health question. *Journal of General Internal Medicine*, 21(3):267–275.
- Dong, Y., Chawla, N. V., and Swami, A. (2017). Metapath2vec: Scalable representation learning for heterogeneous networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '17*, page 135–144, New York, NY, USA. Association for Computing Machinery.
- Fu, X., Zhang, J., Meng, Z., and King, I. (2020). Magnn: Metapath aggregated graph neural network for heterogeneous graph embedding. In *Proceedings of The Web Conference 2020, WWW '20*, page 2331–2341, New York, NY, USA. Association for Computing Machinery.
- Gao, M., Chen, L., He, X., and Zhou, A. (2018). Bine: Bipartite network embedding. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR '18*, page 715–724, New York, NY, USA. Association for Computing Machinery.
- Gentimis, T., Alnaser, A. J., Durante, A., Cook, K., and Steele, R. (2017). Predicting hospital length of stay using neural networks on mimic iii data. In *2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress(DASC/PiCom/DataCom/CyberSciTech)*, pages 1194–1201.
- Grover, A. and Leskovec, J. (2016). node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 855–864.
- Jin, B., Che, C., Liu, Z., Zhang, S., Yin, X., and Wei, X. (2018). Predicting the risk of heart failure with ehr sequential data modeling. *IEEE Access*, 6:9256–9261.
- Johnson, A. E., Pollard, T. J., Shen, L., Lehman, L.-w. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., and Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1):160035.
- Leskovec, J. and McAuley, J. (2012). Learning to discover social circles in ego networks. In Pereira, F., Burges, C., Bottou, L., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc.
- Ma, F., You, Q., Xiao, H., Chitta, R., Zhou, J., and Gao, J. (2018). Kame: Knowledge-based attention model for diagnosis prediction in healthcare. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM '18*, page 743–752, New York, NY, USA. Association for Computing Machinery.
- Nasiri, E., Berahmand, K., Rostami, M., and Dabiri, M. (2021). A novel link prediction algorithm for protein-protein interaction networks by attributed graph embedding. *Computers in Biology and Medicine*, 137:104772.
- Perozzi, B., Al-Rfou, R., and Skiena, S. (2014). Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14*, page 701–710, New York, NY, USA. Association for Computing Machinery.