December 2024

Multimodal Fusion of EHR in *Structures* and *Semantics*: Integrating Clinical Records and Notes with Hypergraph and LLM

Hejie CUI [a*], Xinyu FANG [b*], Ran XU [b], Xuan KAN [b], Joyce C. HO [b] and Carl YANG [b,1]
[a] *Center for Biomedical Informatics Research, Stanford University*
[b] *Department of Computer Science, Emory University*

Abstract. In recent decades, Electronic Health Records (EHRs) have become increasingly useful to support clinical decision-making and healthcare. EHRs usually contain heterogeneous information, such as structural data in tabular form and unstructured data in textual notes. Different types of information in EHRs can complement each other and provide a comprehensive picture of a patient's health status. While there has been a lot of research on representation learning of structured EHR data, the fusion of different types of EHR data (multimodal fusion) is not well studied. This is mostly because of the complex medical coding systems and the noise and redundancy in the written notes. In this work, we propose a new framework called **MINGLE**, which effectively integrates both *structures* and *semantics* in EHR. Our framework uses a two-level infusion strategy to combine medical concept semantics and clinical note semantics into hypergraph neural networks, which learn the complex interactions between different types of data to generate visit representations for downstream prediction. Experiment results on two EHR datasets, the public MIMIC-III and private CRADLE, show that **MINGLE** can effectively improve predictive performance by 11.83% relatively, enhancing semantic integration as well as multimodal fusion for structural and textual EHR data.

Keywords. Electronic Health Record, Clinical Note, Multimodal Fusion, LLMs

1. Introduction

Electronic Health Records (EHRs) are widely used in healthcare and comprise heterogeneous data, including tabular records and clinical notes. Tabular records contain individual visits and are composed of a set of medical concepts like diagnoses and medications. Clinical notes are long documents written by healthcare providers containing detailed information such as patient history, clinical findings, and laboratory test results.

Previous research has focused on modeling structured EHR data for predictive purposes [1], typically using traditional machine learning (ML) models. However, this approach overlooks complex interactions and does not capture hidden structures within the data. To address this, graph neural networks (GNNs) [2, 3] and hypergraph models [4] have been introduced to better capture interactions among visits and medical codes. In this study, we aim to integrate structured EHR data with textual data, combining structures and semantics using medical knowledge from LLMs. We focus on two types of textual information: medical code concept names and clinical notes. Integrating these presents challenges due to the diversity of coding systems (e.g., ICD-10, CPT, SNOMED) and the errors or irrelevant information often found in clinical notes.

[1]Corresponding Author: Carl Yang, j.carlyang@emory.edu.

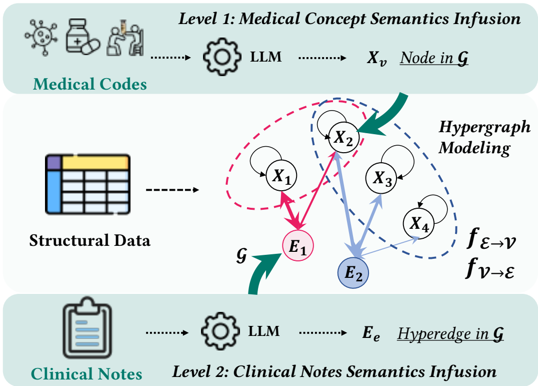

Figure 1.: The EHR fusion framework **MINGLE**.

Integrating the semantics from medical code concept names and clinical notes is crucial for accurate patient record modeling. Recent advances in large language models (LLMs) offer new opportunities for this integration. We explore LLMs to generate semantic embeddings of medical concepts and fuse them with structural information from medical codes and clinical notes to enhance visit-level reasoning.

We propose **MINGLE**, a multimodal EHR fusion framework that integrates structures and semantics from clinical records and notes, as shown in Figure 1. Our approach uses a hypergraph neural network as the backbone and infuses *medical concept semantics* and *clinical notes semantics* into the structural modeling process with a two-level semantics infusion strategy and LLMs. Experiment results on two EHR datasets demonstrate that **MINGLE** effectively enriches the representation of patient information. It jointly leverages the power of hypergraph GNNs to model complex relationships and harnesses the domain knowledge in LLMs and their strengths in natural language understanding. As a result, this integrated approach offers a comprehensive and nuanced analysis of EHR data, leading to more accurate and domain knowledge-enriched decision-making in healthcare.

2. Preliminaries

Hypergraph modeling for multimodal EHR data. EHR data combines structured clinical records and unstructured clinical notes. The structured records are tabular, with each row representing a patient visit and columns for medical codes, while clinical notes offer additional textual context. Building on prior work [4], we transform EHR data into a hypergraph, where each visit is modeled as a hyperedge $\mathcal{E}$ connecting nodes $\mathcal{V}$ corresponding to medical codes. This hypergraph, denoted as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, captures interactions among visits and codes, enabling more effective downstream predictive modeling.

Patient visit representation learning. We use a hypergraph neural network (HGNN) to learn node and hyperedge embeddings via iterative message passing. Node embeddings are updated by aggregating information from connected hyperedges, and hyperedge embeddings are derived from the nodes using multi-head self-attention. The final patient visit embeddings are obtained by aggregating outputs from L layers of message passing, which are then processed through a classification MLP for downstream tasks.

Risk prediction. Given a multimodal EHR dataset $\mathcal{D} = \{\mathcal{T}, \mathcal{N}\}$, $\mathcal{T} = \{\mathcal{T}_p\}_{p=1}^{P}$ represents the structured patient record that include P rows of individual patient visits, and $\mathcal{N}_p$ represents the corresponding clinical notes to each visit. The goal of our method is to train a predictive model that makes a clinical prediction for each given p-th visit $\mathcal{D}_p = \{\mathcal{T}_p, \mathcal{N}_p\}$.

3. Method

Two textual semantics resources exist in the multimodal EHR dataset - the concept names of medical codes in tabular data and clinical notes. To infuse semantics into structural learning of hypergraph modeling, we propose a two-level strategy, as illustrated below.

3.1. Infusing Medical Concept Semantics into the Structural Modeling of EHR data

To reflect the structural contexts of nodes in the graph, we utilize the Deep Walk algorithm [5] to learn a structural latent representation $\boldsymbol{s}_v \in \mathbb{R}^{d_1}$ for each node v in the hypergraph. This is particularly useful in the EHR modeling task as edges are sparse. To model the medical codes from different coding systems in a unified way, we first map the original code v to the corresponding concept name c_v, then utilize GPT `text-embedding-ada-002` model to generate a semantic embedding $\boldsymbol{c}_v \in \mathbb{R}^{d_2}$, which contains clinical knowledge and context background from LLMs. Different ways to combine network-based and knowledge-based encoding are investigated, and the simple concatenation achieves the best performance. Specifically, the node embedding $\boldsymbol{X}_v^{(0)}$ is initialized as the concatenation of both the structural feature $\boldsymbol{S}_v$ and the semantic feature $\boldsymbol{C}_v$ of the nodes in the hypergraph,

$$\boldsymbol{X}_v^{(0)} = [\boldsymbol{S}_v; \boldsymbol{C}_v]. \tag{1}$$

These fused node embeddings are utilized as the node feature initialization of the message-passing process, which induces the initial hyperedge embedding.

3.2. Infusing Clinical Note Semantics into the Structural Modeling of EHR data

EHR datasets contain various types of clinical notes, each serving a specific role in documenting patient care. These include progress notes (tracking condition and treatment during hospitalization), nursing notes (detailing daily care and treatment response), radiology reports (interpreting imaging results), and discharge summaries (providing an overview of the hospital stay, including diagnoses, treatments, and follow-up care instructions). Out of all these types of notes, *discharge summaries* are particularly valuable for integration with structured EHR data as they offer a granular summary that is instructive for continuous patient care.

In **MINGLE**, for each individual patient visit record $\mathcal{T}_p$ (correspond to a hyperedge $e \in \mathcal{E}$), we match the corresponding discharge summary $\mathcal{N}_p$ and filter irrelevant sections such as admission dates, services, etc. A document representation $\boldsymbol{n}_p$ is generated for each *discharge summary* $\mathcal{N}_p$ with the GPT embedding model, resulting in a corpus semantic matrix $\boldsymbol{N}_e$ across all visits. In order to further incorporate fine-grained semantics, we treat single nodes as additional hyperedges in the hypergraph by adding a self-loop on each node. The overall hyperedge semantics embeddings $\boldsymbol{H}_e$ is then the combination of the corpus semantic matrix $\boldsymbol{N}_e$ and the medical concept semantic matrix $\boldsymbol{C}_v$:

$$\boldsymbol{H}_e = \mathrm{MLP}_1\left(\begin{bmatrix}\boldsymbol{N}_e \\ \boldsymbol{C}_v\end{bmatrix}\right). \tag{2}$$

This leads to an enhancement of the central node semantics during its update from connected hyperedges, which also helps to establish a soft collaboration between fine-grained concept semantics and coarse-grained document semantics. Finally, we improve the hyperedge representation updating rule in Eq. (**??**) as below:

$$\boldsymbol{E}_e^{(l)} = \mathrm{MLP}_2([f_{\mathcal{V}\to\mathcal{E}}\left(\mathcal{V}_{e,\boldsymbol{X}^{(l-1)}}\right); \boldsymbol{H}_e]). \tag{3}$$

This means that the hyperedge semantic embeddings $\boldsymbol{H}_e$ are incorporated into each message passing layer, along with the aggregated information from its connected nodes, to update the hyperedge representation.

4. Experiments

Datasets. We have performed experiments on two clinical prediction datasets, MIMIC-III and CRADLE. The CRADLE dataset was collected from a large healthcare system in the United States. The MIMIC-III [6] dataset contains 36,875 visits in all, represented by 7423 medical codes, with 12,353 visits being labeled. The CRADLE dataset contains 36,611 visits with 12,725 codes. We divided them into a train, a validation, and a test set in the ratio of 7:1:2. As natural notes are not included in the CRADLE dataset, we convert individual visits into natural language through textualization.

Tasks. On the MIMIC-III [6] dataset, we perform phenotyping prediction, which involves predicting the presence of 25 care conditions in patients' next visits [7], given their current ICU records. This can be useful for detecting morbidity, repurposing drugs, and diagnosis. On the CRADLE dataset, the task aims to determine if patients diagnosed with type 2 diabetes will experience cardiovascular disease (CVD) endpoints within a year of their diagnosis. CVD endpoint is defined by the presence of coronary heart disease (CHD), congestive heart failure (CHF), myocardial infarction (MI), or stroke. As CVD affects around 32% of patients with diabetes [8], it is essential to have a systematic CVD risk prediction.

Table 1. Performance (100%) on MIMIC-III and CRADLE compared with different baselines. The result is averaged over 5 runs. We use * to indicate statistically significant results ($p < 0.05$). Bold and underlined indicate the best and second-runner results.

Model	MIMIC-III				CRADLE			
	ACC	AUROC	AUPR	F1	ACC	AUROC	AUPR	F1
LR	68.66 ± 0.24	64.62 ± 0.25	45.63 ± 0.32	13.74 ± 0.40	76.22 ± 0.30	57.22 ± 0.28	25.99 ± 0.26	42.18 ± 0.35
SVM	72.02 ± 0.12	55.10 ± 0.14	34.19 ± 0.17	32.35 ± 0.21	68.57 ± 0.13	53.57 ± 0.11	23.50 ± 0.15	52.34 ± 0.22
MLP	70.73 ± 0.24	71.20 ± 0.22	52.14 ± 0.23	16.39 ± 0.30	77.02 ± 0.17	63.89 ± 0.18	33.28 ± 0.23	45.16 ± 0.26
GCT	76.58 ± 0.23	78.62 ± 0.21	63.99 ± 0.27	35.48 ± 0.34	77.26 ± 0.22	67.08 ± 0.19	35.90 ± 0.20	56.66 ± 0.25
GAT	76.75 ± 0.26	78.89 ± 0.12	66.22 ± 0.29	34.88 ± 0.33	77.82 ± 0.20	66.55 ± 0.27	36.06 ± 0.18	56.43 ± 0.26
HGNN	77.93 ± 0.41	80.12 ± 0.30	68.38 ± 0.24	40.04 ± 0.35	76.77 ± 0.24	67.21 ± 0.25	37.93 ± 0.18	58.05 ± 0.23
HyperGCN	78.01 ± 0.23	80.34 ± 0.15	67.68 ± 0.16	39.29 ± 0.20	78.18 ± 0.11	67.83 ± 0.18	38.28 ± 0.19	60.24 ± 0.21
HCHA	78.07 ± 0.28	80.42 ± 0.17	68.56 ± 0.15	37.78 ± 0.22	78.60 ± 0.15	68.05 ± 0.17	39.23 ± 0.13	59.26 ± 0.21
HypEHR	79.07 ± 0.31	82.19 ± 0.13	71.08 ± 0.17	41.51 ± 0.25	79.76 ± 0.18	70.07 ± 0.13	40.92 ± 0.12	61.23 ± 0.18
MINGLE	**80.17 ± 0.08***	**83.54 ± 0.06***	**72.50 ± 0.07***	**46.26 ± 0.61***	78.87 ± 0.48	**73.01 ± 0.06***	**45.76 ± 0.13***	**63.49 ± 0.49***
MINGLE w/o Medical Concept Semantics	79.08 ± 0.18	82.37 ± 0.14*	70.98 ± 0.26	41.83 ± 1.89	**80.07 ± 0.38***	72.49 ± 0.26*	44.63 ± 0.24*	60.62 ± 1.53
MINGLE w/o Clinical Note Semantics	79.77 ± 0.33*	83.14 ± 0.18*	72.02 ± 0.32*	45.69 ± 2.68*	75.39 ± 1.34	70.83 ± 0.62*	43.90 ± 0.90*	63.19 ± 0.60*

Implementation details. We use Adam as the optimizer with a $1e^{-3}$ learning rate. The weight decay is set to $1e^{-3}$. We tune hyperparameters, including the hidden dimension d (24, 48, 72, 96) , the number of layers L (1, 2, 3, 4) in the hypergraph neural network, and the ratio (0.5. 0.67, 1, 1.5, 2) between structural and semantical embeddings. Results are omitted here due to the space limit.

Baselines. We compare **MINGLE** with several baselines: (1) *Conventional ML Baselines.* Logistic Regression (LR), SVM, and MLP, which are non-graph models. (2) *GNN Baselines.* In graph-based methods, the graph is constructed based on pair-wise relations among medical codes: an edge is created between two codes if they co-occur in the same visit. We choose GCT [3] and GAT [9]. (3) *Hypergraph Modeling Baselines.* These baselines are tested using the same hypergraph structured as **MINGLE** but with various neural network architectures. We include HGNN [10], HyperGCN [11], HCHA [12], and HypEHR [4].

5. Results

The results of **MINGLE** compared to baseline models on two EHR datasets are shown in Table 1. **MINGLE** outperforms baselines across four metrics on the MIMIC-III dataset, excelling in F1 score. On the CRADLE dataset, it improves AUROC and AUPR, demonstrating effectiveness in handling unbalanced datasets. A slight accuracy drop may result from better classification of minority classes. We conducted ablation and hyperparameter studies to analyze the model's components and configurations. The ablation study (Table 1, last two rows) highlights the importance of medical concept semantics, whose removal causes a significant performance drop. Clinical note semantics have less impact, likely due to challenges with noisy document representation. In CRADLE, the cleaner, concept-based notes make hyperedge fusion more critical, especially for AUROC and AUPR.

Case study. We present two case studies on MIMIC-III *Cardiac Dysrhythmias* phenotype prediction (Figure 2) to highlight differences in important medical node selection between **MINGLE** and the baseline, based on attention weights in the self-attention mechanism.

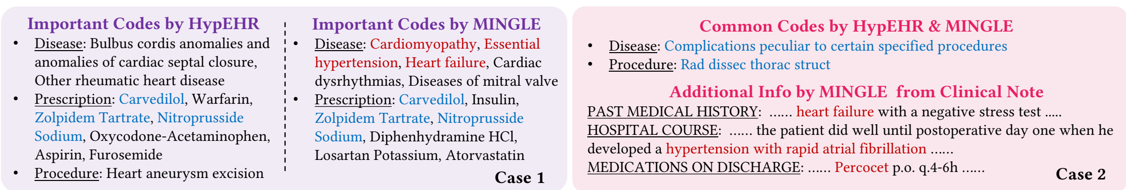

Figure 2. Case studies on important codes identified by HypEHR and **MINGLE**. Blue shows common ones, while Red highlights additional information identified by **MINGLE**.

⋆ **Case 1.** Both HypEHR and **MINGLE** identified key codes like *Carvedilol*, *Nitroprusside Sodium*, and *Zolpidem Tartrate*, which are relevant to *Dysrhythmias*. However, **MINGLE** uniquely identified diseases re-

lated to cardiac function, such as *Heart Failure*, *Cardiomyopathy*, and *Cardiac Dysrhythmias*, demonstrating its ability to incorporate medical concept semantics for deeper clinical insights.
$\star$ **Case 2.** Both models identified similar important codes, but **MINGLE** also utilized clinical notes for additional context. For example, the patient's *past medical history* showed *heart failure near admission*, and the *hospital course* noted *rapid atrial fibrillation* and *hypertension*. The medication *Percocet*, known for cardiovascular effects, added further insight. By combining clinical notes with EHR data, **MINGLE** provided a more comprehensive patient profile.

6. Conclusion and Discussion

We propose a framework called **MINGLE**, which is designed to combine structured EHR and clinical notes using a two-level semantic infusion strategy. The framework uses a hypergraph model and additional semantic information from LLMs to enable the joint learning of complex interactions among medical codes and patient visits. Results demonstrate the benefits of integration, particularly the concept name semantics.

Acknowledgements

This research was partially supported by US National Institute of Diabetes and Digestive and Kidney Diseases of the US National Institutes of Health under Award Number K25DK135913.

References

[1] Benjamin Shickel, Patrick James Tighe, Azra Bihorac, and Parisa Rashidi. Deep ehr: a survey of recent advances in deep learning techniques for electronic health record analysis. *BHI*, 22:1589–1604, 2017.
[2] Juan G Diaz Ochoa and Faizan E Mustafa. Graph neural network modelling as a potentially effective method for predicting and analyzing procedures based on patients' diagnoses. *Artificial Intelligence in Medicine*, page 102359, 2022.
[3] Edward Choi, Zhen Xu, Yujia Li, Michael Dusenberry, Gerardo Flores, Emily Xue, and Andrew Dai. Learning the graphical structure of electronic health records with graph convolutional transformer. In *AAAI*, 2020.
[4] Ran Xu, Mohammed K Ali, Joyce C Ho, and Carl Yang. Hypergraph transformers for ehr-based clinical predictions. *AMIA*, 2023.
[5] Bryan Perozzi, Rami Al-Rfou, and Steven Skiena. Deepwalk: Online learning of social representations. In *KDD*, 2014.
[6] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, et al. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.
[7] Hrayr Harutyunyan, Hrant Khachatrian, David C Kale, Greg Ver Steeg, and Aram Galstyan. Multitask learning and benchmarking with clinical time series data. *Scientific data*, 6(1):1–18, 2019.
[8] Thomas R Einarson, Annabel Acs, Craig Ludwig, and Ulrik H Panton. Prevalence of cardiovascular disease in type 2 diabetes: a systematic literature review of scientific evidence from across the world in 2007–2017. *Cardiovascular diabetology*, 17:1–19, 2018.
[9] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *ICLR*, 2018.
[10] Yifan Feng, Haoxuan You, Zizhao Zhang, Rongrong Ji, and Yue Gao. Hypergraph neural networks. In *AAAI*, volume 33, pages 3558–3565, 2019.
[11] Naganand Yadati, Madhav Nimishakavi, Prateek Yadav, Vikram Nitin, Anand Louis, and Partha Talukdar. Hypergcn: A new method for training graph convolutional networks on hypergraphs. *NeurIPS*, 2019.
[12] Song Bai, Feihu Zhang, and Philip HS Torr. Hypergraph convolution and hypergraph attention. *Pattern Recognition*, 110:107637, 2021.