January 2025

Medical Retrieval-Augmentation Generation Framework for Healthcare Prediction

Yanchao TAN [a], Jie ZHANG [a], Jiamin ZHUANG [b], Guofang MA [c,1], and Carl YANG [d]
[a] *College of Computer and Data Science, Fuzhou University*
[b] *College of Maynooth International Engineering, Fuzhou University*
[c] *School of Computer Science and Technology, Zhejiang Gongshang University*
[d] *Department of Computer Science, Emory University*

Abstract. Electronic Health Records (EHRs) are pivotal for healthcare prediction tasks, offering rich patient data such as symptoms, diagnoses, and treatments. Recent advances in Retrieval-Augmented Generation (RAG) have gained attention due to the ability to retrieve relevant information from medical sources to improve EHR-based predictions. However, existing RAG approaches for medical applications often struggle with flat data representations, which fail to capture the complex inter-dependencies among medical entities, leading to fragmented and verbose responses. In this work, we propose **MedGR**, a novel framework for healthcare prediction that incorporates graph-based clinical text indexing with a dual-level medical retrieval architecture. By leveraging graph-structured knowledge, we synthesize information from multiple sources into coherent and contextually enriched responses in an efficient manner. The experiment results showed that our medical RAG framework achieved high precision performance on both diagnosis and medical code prediction.

Keywords. Electronic Health Record, Medical Retrieval, RAG, LLMs

1. Introduction

Electronic Health Records (EHRs) encompass a wealth of patient data, including symptoms, diagnoses, and medications, making them crucial for healthcare prediction tasks. Recent research has explored integrating external knowledge into EHR data using large language models (LLMs) to enhance predictive performance. For instance, Jiang et al. [1] introduced patient-specific knowledge graphs created by LLMs for bi-attention augmented graph neural networks, Chi et al. [2] developed a multi-label hierarchical classification model that integrates medical knowledge and EHR data to improve disease prediction, and Xu et al. [3] proposed a Retrieval-Augmented Generation (RAG) framework that incorporates multiple knowledge sources for improved EHR predictions. However, current RAG-based approaches are limited by flat data representations that fail to capture intricate relationships among medical entities. As depicted in Fig. 1, when queried about Alzheimer's disease, the existing RAG-based methods often return fragmented responses, ignoring the connections between conditions such as hypertension and vascular dementia. In addition, broader medical topics such as the correlation between cardiovascular issues and Alzheimer's disease are ignored. This limitation underscores the need for more effective RAG systems that can contextualize and synthesize insights from diverse medical domains and provide responses for healthcare data.

In this paper, we propose **MedGR**, a Medical Retrieval-Augmentation generation framework for healthcare prediction. By leveraging graph-structured knowledge, our approach synthesizes information from

[1]Corresponding Author: Guofang Ma, maguofang@zjgsu.edu.cn

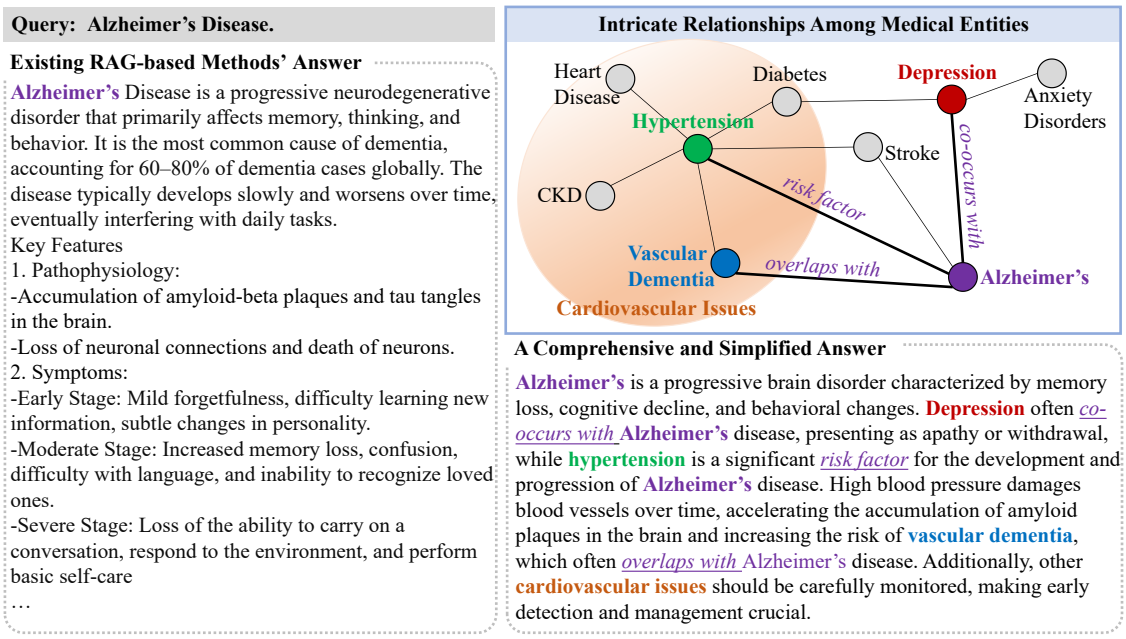

Figure 1. An illustrative example of Alzheimer's disease prediction.

multiple sources into coherent and contextually enriched responses. Specifically, we novelly leverage one simple and efficient generation method (i.e., LightRAG [4]) to incorporate the graph-based clinical text indexing paradigm and dual-level retrieval framework. The low-level retrieves focus on specific medical entities and their relationships, while the high-level captures broader medical themes, ensuring relevant and accurate responses tailored to medical queries. Furthermore, **MedGR** reduces computational overhead and accelerates adaptation, enabling rapid adaptation to new medical data without requiring full index reconstruction. Our proposed model improved predictive performance by 1.13%, demonstrating that our Medical RAG framework effectively captures complex medical relationships and integrates diverse knowledge sources to enhance healthcare predictions from EHR data.

2. Methods

2.1. Problem Definition

The EHR data consists of a group of patients P, each patient i is associated with a sequence of hospital visits $V_i = \{v_{i,1}, v_{i,2}, \ldots, v_{i,j}, \ldots, v_{i,|V_i|}\}$. Each visit $v_{i,j}$ includes a series of medical events, which can be represented as a set of unique medical codes $C_{i,j} \subset C$. In our study, C represents the complete set of medical codes for all patients in P, encompassing diagnoses, procedures, and medications. For the EHR data, we use MIMIC-III dataset [5], which includes 846 diagnoses, 4525 medications, 2032 procedures, and 12353 health records. Following the approach outlined by Ma et al. [6], we filtered the dataset to include only patients with at least two visits. Given history visits V_i and its associated medical codes C_i, the task is to predict the diagnostic labels for a patient's next visit, focusing on 264 Clinical Classifications Software (CCS) codes.

2.2. Graph-based Medical Retrieval-Augmentation Generation Framework

The integration of graph-structured knowledge facilitates the synthesis of information from multiple sources into coherent and contextually rich responses. In this section, we introduce the graph-based medical retrieval augmentation generation framework, **MedGR**, for healthcare prediction, which is extended from LightRAG [4]. We leverage graph-structured knowledge to synthesize information from multiple sources into

contextually enriched responses, enhancing the EHR data for more accurate clinical predictions. **MedGR** incorporates the following four key processes: knowledge retrieval, graph-based knowledge generation, patient visits augmentation and prediction model construction.

Knowledge Retrieval. The retrieval process leverages external knowledge from diverse sources to enhance the coverage of clinical information. We compile this knowledge from resources such as PubMed, DrugBank, MeSH (Medical Subject Headings), Wikipedia, and PrimeKG, representing each piece as the raw text for efficient retrieval. Specifically, each medical code c_k in the patient i's code set C_i corresponds to a specific clinical concept and is associated with a name s_k in the form of short text snippets. For each diagnosis, procedure, and drug code, we use the mapped name as the keyword to query external medical knowledge databases. These queries retrieve the most relevant passages or descriptions related to each keyword, denoted as $T_k = \{T_{k,1}, \ldots, T_{k,m}\}$, where $T_{k,m}$ refers to the m-th retrieved article corresponding to the relevant code c_k.

Graph-based Knowledge Generation. To streamline the context and extract relevant information from T_k and motivated by the efficiency of graph-structured knowledge in synthesizing information from multiple sources into context-rich responses, we propose a Graph-based Knowledge Generation method. In detail, this knowledge generation method combines the graph-based clinical text indexing paradigm with the dual-level retrieval framework as follows:

(a) Graph-based Clinical Text indexing Paradigm. For passages T_k, we first segment them into small and manageable pieces $t_{k,p}$ for quick identification. Next, we leverage LLM to identify and extract various medical entities V_k (e.g., diseases, medications, and procedures) along with the relationships E_k between them in each piece. The information extracted through this process, denoted as Recog$(\cdot)$, will be used to create a comprehensive medical knowledge graph G_k that highlights the connections and insights across the entire external knowledge T_k, as expressed in Eq.1:

$$G_k = (V_k, E_k) = \bigcup_{t_{k,p} \in T_k} \text{Recog}(t_{k,p}) \tag{1}$$

(b) Dual-level Retrieval: We construct code-specific Query for each medical code in the visit, targeting specific entities and their relationships within the medical knowledge graph G_k. The dual-level retrieval approach is applied: low-level retrieval focuses on specific entities, while high-level retrieval aggregates broader information. This process retrieves and summarizes the most relevant knowledge e_k related to the medical code c_k from the graph G_k, as $e_k = \text{DLR}(\text{Query}, G)$, where DLR refers to Dual-level Retrieval.

Patient Visits Augmentation. For a patient i having a visit $v_{i,j}$ with involved medical codes $C_{i,j}$, we now have three subsets: $C^d_{i,j}$ for diagnoses, $C^p_{i,j}$ for procedures, and $C^m_{i,j}$ for medication codes. For each type of medical code, we augment by pairing the code with its corresponding knowledge, e.g., for diagnosis codes $C^d_{i,j}$, we generate $D_{i,j} = \{(c,e) \mid c \in C^d_{i,j}, e = \text{augment knowledge for } c\}$. Then we flatten the visit in sequential order and obtain the code-augmented visit series, which is $X^d_i = \{D_{i,1}, D_{i,2}, \ldots, D_{i,j}\}$. We then initialized a learnable embedding to obtain the initial vector representation of the code-augmented visit series. Subsequently, we utilize Transformer [7], leveraging attention and positional encoding mechanisms, to obtain a more comprehensive visit embedding h^d_i that captures sequence dependencies, denoted as Eq.2:

$$h^d_i = \text{Transformer}(\text{Initial}(X^d_i)) \tag{2}$$

So far, we have acquired the visit embedding corresponding to the diagnosis code feature, and the generation process for the procedure and the medication code is the same.

Prediction Model Construction. We choose a multi-layer perceptron (MLP) classification head for prediction, with the flattened document representations as inputs. The classification process involves aggregating the representations from the different types of medical codes (diagnoses, procedures, and medications) into a single vector, which is then passed through the MLP. The formula for this process is denoted as Eq.3:

$$\widehat{Y}_i = \sigma(\text{MLP}\left(h^d_i \| h^p_i \| h^m_i\right)) \tag{3}$$

The representations h^d_i, h^p_i, and h^m_i correspond to the visit embeddings with the features of diagnosis, procedure, and medication codes respectively. The function σ maps the prediction results to the interval $(0,1)$. The prediction $\widehat{Y}_i$ represents the predicted labels for patient i for prediction of the next visit diagnosis.

3. Results

Metrics. For evaluation, we use visit-level precision@K and code-level accuracy@K to assess diagnosis prediction performance. The true labels and the predicted labels for patient i are denoted as Y_i and $\widehat{Y}_i$, respectively, where $Y_i = \{y_1, y_2, \ldots, y_n\}$ and $y_i \in \{0, 1\}$. Visit-level precision@K measures predictive performance for a patient's next visit and is calculated by summing the correct top-K predictions over the minimum of K and the number of diagnostic codes in Y_i. Code-level accuracy@K evaluates the overall accuracy of code predictions across all patients by dividing the sum of correct top-K predictions by the total diagnostic codes in the dataset, with a higher ratio indicating greater code prediction accuracy.

Baselines. We evaluate **MedGR** against sequence- and graph-based baselines. Sequence models include StageNet [8], combining stage-aware LSTM with stage-adaptive convolutions, and HiTANet [9], a hierarchical time-aware Transformer. Graph models include KAME [6], integrating ICD code-based ontology, DDHGNN [10], using dynamic graph convolutions, and TRANS [11], a temporal graph transformer.

Table 1.Performance on MIMIC-III with different baselines(%). We use **bold** and underlined to indicate the best and second-runner results.

Model	Visit-Level Precision@K			Code-Level Accuracy@K		
	10	20	30	10	20	30
StageNet	51.49	59.57	69.33	38.08	55.72	66.61
HiTANet	56.33	63.66	73.38	40.95	59.87	70.68
KAME	52.94	60.76	70.74	39.56	57.48	68.66
DDHGNN	56.80	63.73	72.99	40.59	59.90	70.59
TRANS	57.65	64.18	73.69	41.30	60.19	71.17
MedGR	**57.78**	**64.77**	**74.54**	**41.72**	**60.92**	**72.29**

Performance Analysis. Table 1 shows that on the MIMIC-III dataset, **MedGR** outperforms all baseline models, including sequence-based models like StageNet and graph-based models like KAME. Compared to the TRANS model, **MedGR** achieves a 1.13% improvement in Visit-Level Precision@30 and a 1.12% increase in Code-Level Accuracy@30. This enhancement is attributed to the integration of advanced retrieval mechanisms with graph-based reasoning in **MedGR**, which provides the most relevant and logically consistent augmented information for medical codes, thereby significantly supporting clinical decision-making.

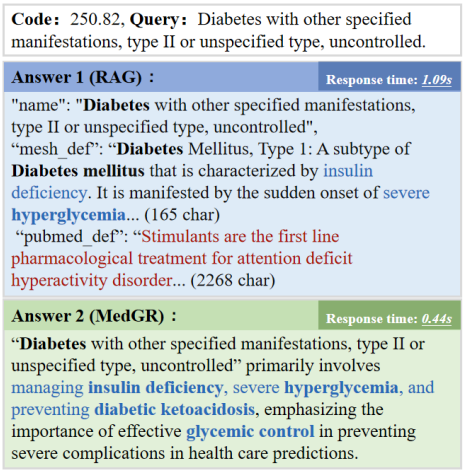

Figure 2.A case study with *Diabetes*. **Bold** denotes disease concepts, Blue and Red indicate useful and irrelevant knowledge.

4. Discussion

We present a case study on MIMIC-III *Diabetes* prediction in Figure 2 to compare medical retrieval knowledge generated by the RAG model and **MedGR**. We observe that our framework provides more relevant information, particularly regarding the Diabetes outcome in this case. Answer 2 (**MedGR**) is more comprehensive in terms of professionalism and specificity because it not only covers the key features of the disease but also emphasizes the importance of preventing diabetic ketoacidosis and highlights the key role of effective glycemic control in preventing serious complications. This is a better description for healthcare professionals because it provides more specific management goals and preventive measures, which are core elements in the treatment and management of diabetes. Therefore, answer 2 is a better summary because it more comprehensively covers the clinically relevant aspects of the disease and emphasizes key points in treatment and prognosis.

5. Conclusions

The **MedGR** framework enables the integration of graph-based retrieval-augmentation generation into EHR systems. This approach improves the ability to capture complex medical relationships and enhances predictive performance, making it a valuable tool to support clinical decision-making.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (62302098), Zhejiang Provincial Natural Science Foundation of China (LQ23F020007), Zhejiang Provincial Department of Agriculture and Rural Affairs Project (2024SNJF044), and Zhejiang Gongshang University "Digital+" Disciplinary Construction Management Project (SZJ2022B001). Carl Yang was not supported by any funds from China.

References

[1] Pengcheng Jiang, Cao Xiao, Adam Richard Cross, and Jimeng Sun. Graphcare: Enhancing healthcare predictions with personalized knowledge graphs. In *The Twelfth International Conference on Learning Representations*.

[2] Shengqiang Chi, Yuqing Wang, Ying Zhang, Weiwei Zhu, and Jingsong Li. Graph neural network based multi-label hierarchical classification for disease predictions in general practice. In *MEDINFO 2023—The Future Is Accessible*, pages 725–729. IOS Press, 2024.

[3] Ran Xu, Wenqi Shi, Yue Yu, Yuchen Zhuang, Bowen Jin, May Dongmei Wang, Joyce Ho, and Carl Yang. Ram-ehr: Retrieval augmentation meets clinical predictions on electronic health records. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 754–765, 2024.

[4] Zirui Guo, Lianghao Xia, Yanhua Yu, Tu Ao, and Chao Huang. Lightrag: Simple and fast retrieval-augmented generation. *CoRR*, abs/2410.05779, 2024.

[5] Alistair Johnson, Tom Pollard, Lu Shen, Li-wei Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Celi, and Roger Mark. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3:160035, 05 2016.

[6] Fenglong Ma, Quanzeng You, Houping Xiao, Radha Chitta, Jing Zhou, and Jing Gao. Kame: Knowledge-based attention model for diagnosis prediction in healthcare. In *Proceedings of the 27th ACM international conference on information and knowledge management*, pages 743–752, 2018.

[7] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

[8] Junyi Gao, Cao Xiao, Yasha Wang, Wen Tang, Lucas M Glass, and Jimeng Sun. Stagenet: Stage-aware neural networks for health risk prediction. In *Proceedings of The Web Conference*, pages 530–540, 2020.

[9] Junyu Luo, Muchao Ye, Cao Xiao, and Fenglong Ma. Hitanet: Hierarchical time-aware attention networks for risk prediction on electronic health records. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 647–656, 2020.

[10] Qianlong Wen, Zhongyu Ouyang, Jianfei Zhang, Yiyue Qian, Yanfang Ye, and Chuxu Zhang. Disentangled dynamic heterogeneous graph learning for opioid overdose prediction. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 2009–2019, 2022.

[11] Jiayuan Chen, Changchang Yin, Yuanlong Wang, and Ping Zhang. Predictive modeling with temporal graphical representation on electronic health records. In *International Joint Conference on Artificial Intelligence*, 2024.