An Improve of “EMERGE: Enhancing Multimodal Electronic Health Records Predictive Modeling with Retrieval-Augmented Generation”

Guoqian Zeng[1], Paul Grigas[1]
j_zeng@berkeley.edu, pgrigas@berkeley.edu

1: University of California, Berkeley, Berkeley, California.

I. Summary of Contributions

The public healthcare infrastructures and system has a high demand on intelligent renewing of the delivering system, like staff planning or readmission risk management, to improve treatment success rate and operational efficiency. Operational work in the health system produces vast, continuous, and heterogeneous records that could be used in multimodal predictive learning. [1] The emerging Electronic Health Records (EHRs) consolidate digitized health information from multiple sources, organized around the patient, often containing a patient’s medical history, diagnoses, immunization dates, and clinical notes. [2] Such data type was used in previous studies with neural network approaches singular-model predictions. Existing multimodal models process to learn a mapping from heterogeneous inputs to output labels on other data types. In concern to the lack of medical expertise to weigh health information, some recent studies have begun to integrate knowledge graphs (KG) in their analysis.

However, the paper notices a gap between the EHRs’ abundant source of information—especially previously ignored time-series data and text-free clinical notes—and the effectiveness of knowledge graph insertion for prediction modeling.

Therefore, a predictive modeling framework with integrated RAG, EMERGE, was proposed by the paper as a modern approach to improve effectiveness of healthcare predictive tasks with EHRs data. For this report, we divided it into two-stages: (1) an extra summary generation driven by a RAG subframework; followed by (2) a predictive multimodal modeling.

EMERGE is the state-of-the-art framework that tackles these challenges to elevate predictive modeling in EHRs data, with its combined structure of the Retrieval Augmented Generation and the multimodal fusion. This structure helps it utilize a broad range of professional information, while connecting them in a biomedical perspective, to elaborate prompts for health specific predictive tasks.

To ensure the goal-oriented feature, the paper further concludes two major challenges to focus while developing solutions: 1. How to perform entity extraction and matching from EHR analysis to incorporate with KG? 2. What helps integrating long-text retrieval for task-specific prompting?

Initially, they want to get the representation from the original data as a part of input to the LLM before using LLM as an enhancement, this would mine the meaningful information from the original EHRs data to further improve the effectiveness of their innovative models. They use Gated Recurrent Unit network as the encoder to capture the time-dependencies in sequence data and encode this temporally linked information. They then utilize a medical domain language model to obtain text embeddings.

Let’s now focus on the Retrieval Augmented Generation part, this part extracts medical entities from time-series data and texts information from the EHRs to generate a concise health summary. Purposes of generating the summary is to use as supplemental information like an extension with the mixed of the original information for the BERT model in the later part to perform specific predictive tasks. Why do we need these extension data? Because it would provide the accurate information, specific sourced information that is cited directly from the patient records, and overcome terminology confusion to provide domain-specific results for the task, accurating predicting.

In terms of time-series data extraction, they use a z-score method, which is also widely being used in other medical fields, to find the outliers of the data. The outliers represent the abnormal high or low in the tabulated data of measurement, such as the temperature is too low. These outliers are significant to identify the patient disease or parts to be noticed.

Text extraction, they use an iterative method to first use LLM to conclude the important disease words in the clinical notes, and then do a refinement to improve quality. In the first stage, they designed a prompt with clear instructions and an example to tell the LLM to extract disease terminologies that the patient might suffer from. They then use the prompt to generate information, extended with the original inputs. In refinement, they deal with hallucinations by three steps: filter out the terminologies that are not in the original input, filter out the words that are not in their defined disease type, and filter out duplicate words to avoid semantic redundancy. They replicate this process for multiple times until achieving convergence.

Now, after getting the entities as extraction of time-series data and clinical notes, the paper goes on to use a semantic-based victor retrieval approach to tackle the problems in matching of medical entities and knowledge graphs. Initially, they used the same sentence embedding model to encode nodes from Knowledge Graph and the entities from EHRs, ensuring that these embeddings are aligned with the same vector space. [3] Then, for the matching process, they calculate the similarities between the entity and each node in the Knowledge graph using the cosine similarities. They published a threshold along with their methods to select the matched nodes while maintaining qualities. This means that they will search for the most nodes with the highest similarity with the KG node and filter out the maximums below a certain level. After the framework selects a set of matched nodes, it gets definitions and descriptions of entities from the KG. Later it obtains the relationships between diseases which act as edges of KG.

The framework walks into the procedure to generate a prompt for the concise patient health summary with the information entities drawing from EHR time-series data and clinical notes with their extended information gained. The framework uses RAG to dense information because a combination of complete information of extracted entities from the original data set and supplementary information might be overlong for the input of LLM. A prompt template consisting of (1) role & instructure (2) extracted entities in time series (3) extracted entities in notes (4) retrieved knowledge graph nodes (5) retrieved knowledge graph triplets is provided.

Now, with extracted entities from RAG and separate extracted entities from clinical text and time-series data, the Multimodal Fusion Network is structured to integrate these information in an unified version and fuse for prediction. The embeddings from clinical notes and RAG summaries, both texts, are concatenated and passed through a text fusion module to generate a combined text representation. Then, a cross-attention mechanism is employed in both directions: the time-series representation attends to the fused text representation, and vice versa. This bidirectional attention allows each modality to incorporate contextual information from the other. The outputs from both attention pathways are then concatenated and passed through a Multi-Layer Perceptron (MLP) to get the final fused representation for the prediction tasks. The BCE loss is selected for downstream prediction tasks such as in-hospital mortality and 30-day readmission.

This structure is strongly correlated with our course because it elevates the LLM for an engineering design to assist the healthcare operations. This is related to the time-series modeling in 242A and the recurrent neural network modeling in 242B. We are able to use the understanding of the neural network architectures to improve this study.

II. Evaluation of the Contributions

The EMERGE framework demonstrates its state-of-the-art performance as an integration of RAG with multimodal EHR data to enhance predictive modeling in the healthcare delivery system. This evaluation critically assesses the strengths and limitations of EMERGE, providing insights into its contributions and areas for improvement.

1. Strengths

It innovates RAG integration on LLMs with both time-series data and clinical notes to extract and summarize relevant medical entities. By aligning extracted entities with a professional medical knowledge graph (PrimeKG), EMERGE ensures consistency and semantic richness in the data representation. This approach enables model learning to absorb multimodal data with contextual information and capture professional relationships between entities, which is a potential cause for the outperformed model accuracy in the experiments.

When bridge the extractions from multimodal resources, the framework employs an adaptive cross-attention mechanism to fuse information on an uniform dimension effectively. This design allows for a more comprehensive understanding of patient pattern, as it considers the interplay between various data types.

In experiments, it demonstrates superior prediction and AUROC performance on benchmark datasets (MIMIC-III and MIMIC-IV) for in-hospital mortality and 30-day readmission prediction. Then, it also demonstrates the workflow design process with performance comparison of critical points like GRU and adaptive features with their counterparts. The inclusion of extensive ablation studies further validates the effectiveness of each component within the framework.

2. Weakness & Improvements

However, EMERGE uses complex integration of multiple components, including LLMs, knowledge graphs, and cross-attention mechanisms, which increases the computational demands of the framework. This complexity may hinder real-time application in clinical settings where resources are limited. Therefore, streamlining the framework through model pruning or more efficient algorithms could possibly reduce computational overhead.

Notably, its reliance on external knowledge graphs like PrimeKG introduces potential challenges related to data availability and maintenance. Any inconsistencies or updates in the knowledge base could affect the model's performance.

In the experiments, its validation is primarily based on MIMIC-III and MIMIC-IV datasets for in-hospital mortality and 30-day readmission prediction. Its performance and adaptability to other healthcare systems or diverse patient populations remain to be explored.

We must also recognize that the difficulties to apply dynamic knowledge graphs of this model would cause loss of model accuracy in clinical study, but developing advanced techniques to filter and prioritize retrieval could mitigate the risk of information loss and improve model interpretability.

III. Further Extensions

Building on EMERGE’s RAG-driven fusion, we propose integrating a sparse Mixture-of-Experts (MoE) gating layer to specialize processing for each modality (textual summaries vs. time-series data). Specifically, a top-1 sparse gate routes each patient instance to the most relevant expert—either a ClinicalBERT-based text expert or a GRU-based time-series expert—mirroring the sparsely-gated MoE paradigm shown to separate patient heterogeneity effectively in EHR data [6] and applied in healthcare [7]. This sparse gating reduces computation by activating only one expert per instance, while enabling each expert to specialize, improving both efficiency and performance over dense fusion.

We recommend an ordered modality pipeline: first process clinical notes through a fine-tuned ClinicalBERT to produce text embeddings, then inject these embeddings, along with the RAG-generated summaries, as inputs to the sparse MoE gate. The selected text expert transforms these embeddings, and its output is concatenated with structured features before passing to the time-series expert (a GRU or Transformer encoder). This ordering ensures that rich textual context informs subsequent temporal modeling, reflecting the design of the Multi-Modal Forecaster, which aligns textual and time-series features in sequence (TimeText Corpus) [8].

To evaluate the impact of sparse gating vs. alternative gating mechanisms, we propose small-scale numerical experiments on a held-out subset of MIMIC-IV (or UCI Diabetes). Three configurations would be compared:

No gating (baseline fusion): concatenate text and time-series embeddings (EMERGE style) [5].

Dense gating: fuse both experts' outputs via learned dense weights (as in early FuseMoE work) [10].

Sparse top-1 gating: activate only the selected expert per patient [6].

For each configuration, we measure ROC-AUC, Precision-Recall AUC, and inference latency. Prior studies show that sparse MoEs achieve comparable or superior accuracy with reduced compute compared to dense MoE baselines. We hypothesize that sparse gating will improve readmission prediction AUC by 1–2% over EMERGE's original fusion, while halving inference time.

Methodologically, this extension addresses two limitations: reducing pipeline complexity and computational cost, and capturing patient-specific modality relevance via expert specialization. Such an MoE-enhanced EMERGE could be a step toward robust, real-time clinical decision support, where efficient inference and interpretability (by inspecting gate activations) are critical.

IV. References

[1] Chandak P, Huang K, Zitnik M. Building a knowledge graph to enable precision medicine. Sci Data. 2023 Feb 2;10(1):67. doi: 10.1038/s41597-023-01960-3. PMID: 36732524; PMCID: PMC9893183.

[2] Jha, A. K., DesRoches, C. M., Campbell, E. G., Donelan, K., Rao, S. R., Ferris, T. G., Shields, A., Rosenbaum, S., & Blumenthal, D. (2009). Use of electronic health records in U.S. hospitals. New England Journal of Medicine, 360(16), 1628–1638.

[3] Lewis, P., Oguz, B., Rinott, R., Riedel, S., & Stenetorp, P. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Advances in Neural Information Processing Systems 33 (NeurIPS 2020).

[4] Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M Dai, Nissan Hajaj, Michaela Hardt, Peter J Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, et al. 2018. Scalable and

accurate deep learning with electronic health records. NPJ digital medicine 1, 1 (2018), 18.

[5] Yinghao Zhu, Changyu Ren, Zixiang Wang, Xiaochen Zheng, Shiyun Xie, Junlan Feng, Xi Zhu, Zhoujun Li, Liantao Ma, and Chengwei Pan. 2024. EMERGE: Enhancing Multimodal Electronic Health Records Predictive Modeling with Retrieval-Augmented Generation. In Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM '24), October 21–25, 2024, Boise, ID, USA. ACM, New York, NY, USA, 11 pages. DOI: 10.1145/3627673.3679582

[6] Shazeer, N. et al. "Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer," NeurIPS 2021.

[7] Z. Huo et al., "Sparse Gated Mixture-of-Experts to Separate and Interpret Patient Heterogeneity in EHR data," 2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI), Athens, Greece, 2021, pp. 1-4, doi: 10.1109/BHI50953.2021.9508549.

[8] Kim, K., Tsai, H, Sen, R., Das A., Zhou Z., Tanpure A., Luo M., & Yu R. (2024). Multi-Modal Forecaster: Jointly Predicting Time Series and Textual Data. Clinical Orthopaedics and Related Research.

[10] Han, X., Nguyen, H., Harris, W. C., Ho, H. et al. (2024). FuseMoE: Mixture-of-Experts Transformers for Fleximodal Fusion. (NeurIPS 2024).