DisGraph-RP: Graph-Augmented Temporal Modeling with Aspect-Based Contrastive Encoding of Discharge Summary for Readmission Prediction

Sudeshna Jana[1,2], Manjira Sinha[1], Tirthankar Dasgupta[1], Pabitra Mitra[2]

[1]Tata Consultancy Services Research, Kolkata, India
[2]Indian Institute of Technology Kharagpur, India
Corresponding Author: sudeshna.jana@tcs.com

Abstract

Predicting hospital readmissions is a critical clinical task with substantial implications for patient outcomes and healthcare cost management. We propose *DisGraph-RP*, a graph-augmented temporal modeling framework that integrates structured discourse-aware text representation with cross-admission relational reasoning. Our approach introduces a Section-Aware Contrastive Encoder that leverages section segmentation and aspect-based supervision to produce fine-grained representations of discharge summaries. These representations are then composed over time using a Graph-Based temporal module that encodes inter-visit dependencies through learned edge relations, enabling the model to capture disease progression, treatment history, and recurrent risk signals. Experiments on multiple real-world datasets demonstrate that DisGraph-RP achieves significant improvements over strong baselines, including transformer-based clinical models and prompting-based LLM approaches. Our findings highlight the importance of combining discourse-informed text encoding with temporal graph reasoning for robust clinical outcome prediction.

1 Introduction

Hospital readmission particularly within a short period after discharge, remains a major challenge for healthcare systems worldwide, negatively affecting patient outcomes and increasing healthcare expenditures (Burke and Coleman, 2013; Lu et al., 2016). Recent data indicate that nearly 15% of hospitalized patients in the U.S. are readmitted soon after discharge[1]. This rate is even higher for chronic and acute conditions - 23% for heart failure, 20% for stroke, 21% for chronic obstructive pulmonary disease (COPD), and 18% for pneumonia [2]. Importantly, almost 60% of these readmissions are considered potentially avoidable with adequate follow-up care, revealing persistent gaps in care continuity. In the U.S., CMS administers the Hospital Readmissions Reduction Program (HRRP), imposing substantial financial penalties on hospitals with excessive readmission rates and costing the healthcare system billions annually (Psotka et al., 2020; Gupta and Fonarow, 2018). These clinical and economic pressures have intensified the demand for scalable, automated models that accurately estimate patient-specific readmission risk at discharge.

Predictive modeling using EHRs - combining structured data with unstructured clinical narratives, has gained prominence for readmission prediction (Ashfaq et al., 2019; Rojas et al., 2018; Cai et al., 2016). Prior approaches, ranging from rule-based systems to statistical and deep learning models, including disease-specific solutions for heart failure, sepsis, and pneumonia (Shin et al., 2021; Liu et al., 2019; Amrollahi et al., 2022; Huang et al., 2022), primarily rely on structured features and underutilize long-form clinical notes. Yet, these narratives contain crucial insights into clinical status, reasoning, and care decisions. Moreover, a patient's current health status is shaped by prior hospitalizations, comorbidities, and chronic relapsing conditions, emphasizing the need for effective longitudinal modeling.

To address these limitations, we propose a framework, *DisGraph-RP*, that predicts 30-day readmissions using discharge summaries. These summaries encapsulate critical details of a patient's hospitalization, including clinical course, procedures, treatments, discharge status, and follow-up plans. To improve contextual representation, our approach integrates discharge summaries from prior admissions, enabling the model to capture longitudinal patterns of chronicity, recurrence, and treatment response. Our key contributions are as follows:

[1] https://www.definitivehc.com/resources/healthcare-insights/average-hospital-readmission-state

[2] https://wifitalents.com/hospital-readmission-rates-statistics/

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 5: Industry Track), pages 801–812
March 25-27, 2026 ©2026 Association for Computational Linguistics

- We introduce a Section-Aware Contrastive Encoder that embeds long-form discharge summaries into task-specific, semantically rich representations.

- We propose a Graph-Augmented Temporal Model that encodes a patient's longitudinal hospitalization history and refines the current admission embedding for improved prediction.

- We construct two large-scale datasets - 42,573 admissions from 33,107 patients in MIMIC-III (Johnson et al., 2016) and 9,947 admissions from 4,539 patients in MIMIC-IV (Johnson et al., 2023) and demonstrate that our framework consistently outperforms state-of-the-art baselines.

2 Problem Formulation

Let the hospitalization history of a patient p be represented as a sequence of admissions:

$$\mathcal{X}_p = \{\mathcal{H}_p^{[a_1,d_1]}, \mathcal{H}_p^{[a_2,d_2]}, \cdots, \mathcal{H}_p^{[a_n,d_n]}\}$$

where n denotes the total number of hospital visits of p, and each $\mathcal{H}_p^{[a_i,d_i]}$ corresponds to the i-th admission with admission date a_i and discharge date d_i. Each admission comprises multimodal electronic health records (EHRs), including structured clinical measurements and unstructured clinical notes. At discharge, a summary is generated detailing the hospitalization course, diagnoses, procedures, treatments, discharge condition, and follow-up recommendations, providing a comprehensive view of the inpatient episode.

A *readmission* for patient p is defined as an admission j occurring within 30 days of the previous discharge, i.e., $a_j - d_{j-1} \leq 30$. Given a patient p, the discharge summary of the current hospitalization $\mathcal{DS}_p^{(d_t)}$ and summaries from all prior hospitalizations $[\mathcal{DS}_p^{(d_1)}, \ldots, \mathcal{DS}p^{(dt-1)}]$, our task is to predict whether the patient will be readmitted within 30 days of the current discharge.

3 Methodology

The schematic overview of the proposed DisGraph-RP framework is shown in Figure 1. It comprises two main components: (i) a Section-Aware Contrastive Encoder that extracts section-level semantics to embed long-form discharge summaries, and (ii) a Graph-Augmented Temporal Model that encodes a patient's longitudinal hospitalization history to dynamically refine the current admission representation.

3.1 Section-Aware Contrastive Encoder for Discharge Summary Representation

Discharge summaries provide a comprehensive narrative of hospitalization, including medical history, diagnoses, treatments, complications, discharge condition, and follow-up plans (see Appendix A). Their richness makes them highly informative for readmission prediction. However, their length (often exceeding 2,500 tokens) poses a major modeling challenge. Transformer-based encoders such as ClinicalBERT (Huang et al., 2019) are limited to 512 tokens, forcing summaries to be split into segments and disrupting global structure and inter-section dependencies. To address this, we segment each summary into clinically meaningful sections using prompt-based extraction with LLaMA-3.1-8B-Instruct (Grattafiori et al., 2024) (details in Appendix B).

3.1.1 Section-Level Semantic Embedding

Furthermore, we design an additional text-processing pipeline to embed each discharge summary section. To handle variability and non-standard terminology in clinical narratives, we construct unified section representations by combining ontology-based latent embeddings with contextual embeddings from pretrained language models.

To obtain ontology-based latent embeddings (detailed in Appendix C), we first extract diverse clinical entity types such as *diseases*, *symptoms*, *abnormalities*, *lifestyle factors*, *mental health conditions*, *procedures*, and *medications*, using MetaMap (Aronson, 2001). In addition, employ negEx (Mehrabi et al., 2015) to identify negated expressions commonly found in clinical narratives, such as *"no history of sob"* or *"absence of pain"*. To resolve synonymy and terminological variation (e.g., "pulmonary edema" vs. "fluid in lungs"), all entities are standardized to their corresponding UMLS Concept Unique Identifiers (CUIs) (Schuyler et al., 1993). For a discharge summary $\mathcal{DS}$ segmented into k sections, each section is represented as a vector $v_i \in \{-1, 0, 1\}^{|E|}$ over the entity vocabulary E, where $v_i[e] = 1$ denotes the presence, -1 denotes a negated mention, and 0 indicates absence of entity e in that section. Because these vectors are high-dimensional and sparse, we

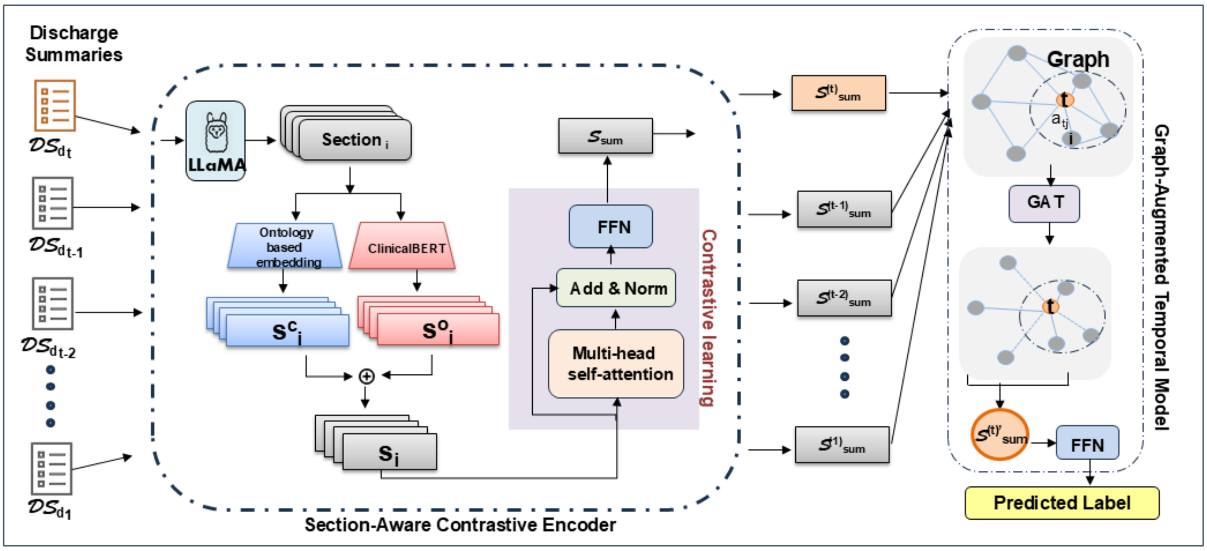


Figure 1: Schematic overview of the DisGraph-RP framework for hospital readmission prediction.

employ an unsupervised autoencoder (Wang et al., 2016) to derive dense latent embeddings $s_i^o \in \mathbb{R}^m$, $m \leq |E|$. The encoder $\phi_{\text{enc}} : \mathbb{R}^{|E|} \rightarrow \mathbb{R}^m$ captures the underlying semantics of the section, while the decoder $\phi_{\text{dec}} : \mathbb{R}^m \rightarrow \mathbb{R}^{|E|}$ attempts to reconstruct the original sparse vector by preserving key clinical content: $s_i^o = \phi_{\text{enc}}(v_i)$, $\hat{v}i = \phi_{\text{dec}}(s_i^o)$. The model is trained to minimize the following reconstruction loss: $\mathcal{L}_{\text{AE}} = \sum_i |v_i - \hat{v}_i|_2^2$.

In addition, we encode each section of the discharge summary using ClinicalBERT to obtain contextualized linguistic features. Specifically, we extract the embedding corresponding to the special token [CLS] as the section-level representation: $s_i^c = ClinicalBERT\ ([CLS]\,(x_i)\,[SEP])_{[CLS]}$, where x_i is the token sequence corresponding to the ith section.

We obtain the final representation of each section by concatenating the ontology-based latent embedding and the contextual embedding, $s_i = s_i^o, ||, s_i^c \in \mathbb{R}^{m'}$, where $m' = m + 768$. The discharge summary is thus represented as a sequence of k section-level vectors: $[s_1, s_2, \ldots, s_k] \in \mathbb{R}^{k \times m'}$. This fusion preserves clinically grounded semantic structure while capturing fine-grained contextual cues, yielding richer and more informative section representations.

3.1.2 Section-Aware Contrastive Encoder

Although discharge summaries are segmented and encoded using ontology-based or contextual representations, not all sections contribute equally to readmission prediction. Sections like *Chief Complaint*, *Brief Hospital Course*, and *Discharge Condition* carry stronger predictive signals than less informative ones such as *Administrative Information* or *Allergies*. Conventional aggregation methods (e.g., uniform averaging or fixed-order concatenation) ignore this variability. To address this, we introduce a Section-Aware Contrastive Encoder that learns adaptive attention over sections, emphasizing clinically informative content conditioned on patient context.

Let the discharge summary be represented as a sequence of k section embeddings $\mathcal{S} = [s_1, s_2, \ldots, s_k]$, where each $s_i \in \mathbb{R}^{m'}$. To model inter-sectional dependencies, we apply multi-head self-attention (Vaswani et al., 2017), projecting the inputs into query, key, and value spaces: $Q = \mathcal{S}W^Q$, $K = \mathcal{S}W^K$, $V = \mathcal{S}W^V$, where $W^Q, W^K, W^V \in \mathbb{R}^{m' \times m'_h}$ are learnable projection matrices, and $Q, K, V \in \mathbb{R}^{k \times m'_h}$ are the corresponding head-specific representations for h attention heads with $m'_h = m'/h$. Then, the self-attention mechanism computes a weighted combination of all sections based on pairwise similarity: $\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{m'_h}}\right)V$ and outputs from all heads are concatenated and passed through a residual connection, layer normalization, and a position-wise feedforward network to produce the updated section representations $\mathcal{S}' \in \mathbb{R}^{k \times m'}$.

To obtain a fixed-size discharge summary representation, we apply attention pooling over the section embeddings.

$$\mathcal{S}_{\text{sum}} = \sum_{i=1}^{k} \alpha_i {s'}_i \tag{1}$$

$$\alpha_i = \frac{\exp(w^\top \tanh(W{s'}_i))}{\sum_{j=1}^{k} \exp(w^\top \tanh(W{s'}_j))} \tag{2}$$

where W and w are learnable parameters, and $\mathcal{S}_{\text{sum}} \in \mathbb{R}^{m'}$ denotes the final discharge summary embedding.

To train the encoder, we adopt a contrastive learning objective that encourages clinically similar discharge summaries to be close in the embedding space. Positive pairs are constructed by selecting summaries that share the same readmission label and exhibit high semantic similarity in their Chief Complaint sections, ensuring clinically meaningful alignment. Given a batch of N discharge summaries, the $\mathcal{L}_{\text{NT-Xent}}$ contrastive loss is defined as:

$$\mathcal{L}_{\text{NT-Xent}} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\sum\limits_{j \in \mathcal{P}(i)} \beta_{i,j}}{\sum\limits_{k=1,\, k \neq i}^{2N} \beta_{i,k}} \tag{3}$$

$$\beta_{i,j} = \exp(\text{sim}(\mathcal{S}_{sum_i}, \mathcal{S}_{sum_j})/\tau) \tag{4}$$

where $\text{sim}(\cdot,\cdot)$ denotes cosine similarity, τ is a temperature scaling parameter, $\mathcal{P}(i)$ denotes the set of positive indices for the i-th sample, defined as:

$$\mathcal{P}(i) = \{ j \mid j \neq i; y_j = y_i; \text{sim}(c_i, c_j) \geq \delta \} \tag{5}$$

with y_i and y_j as the readmission labels and c_i, c_j as the embeddings of the *Chief Complaint* sections. The similarity threshold δ ensures that only semantically similar diagnoses are considered clinically aligned. This contrastive framework enhances the encoder's ability to learn outcome-aware and context-sensitive discharge representations, improving their effectiveness for readmission prediction.

3.2 Graph-Augmented Temporal Model for Readmission Prediction

After obtaining the discharge summary representations, we incorporate the patient's historical hospitalization records to model temporal and clinical dependencies relevant to readmission risk. Let $\mathcal{S}_{\text{sum}}^{(t)}$ denote the current discharge embedding and $[\mathcal{S}_{\text{sum}}^{(1)}, \ldots, \mathcal{S}_{\text{sum}}^{(t-1)}]$ the embeddings of prior admissions. Notably, past hospitalizations contribute unevenly to future risk, depending on both temporal proximity and clinical similarity. For example, a cardiac-related admission from a year earlier may be more informative for a current cardiac visit than a recent admission for an unrelated condition say, leg injury.

To model these asymmetric temporal dependencies, we construct a patient-specific graph $G = (V, E)$, where each node $v_i \in V$ corresponds to a hospitalization episode represented by its discharge summary embedding $\mathcal{S}_{\text{sum}}^{(i)}$. For a patient with t admissions, the graph thus contains t nodes - the current admission and $t-1$ historical ones. Edges E capture both clinical relatedness and temporal proximity between episodes. Specifically, we construct a fully connected graph whose edge weights reflect a joint function of semantic similarity and time interval. The resulting adjacency matrix $\mathcal{A} \in \mathbb{R}^{t \times t}$ is defined as:

$$A_{ij} = \begin{cases} \text{sim}(\mathcal{S}_{\text{sum}}^{(i)}, \mathcal{S}_{\text{sum}}^{(j)}) \cdot e^{-\lambda \Delta t_{ij}}, & \text{if } i \neq j \\ 0, & \text{if } i = j \end{cases} \tag{6}$$

where $\text{sim}(\cdot,\cdot)$ denotes the cosine similarity between admissions i and j, Δt_{ij} is the time gap (in days), and λ is a time–decay coefficient. This formulation prioritizes temporally recent and clinically similar past admissions while downweighting distant or less relevant episodes.

Then, we apply a Graph Attention Network (GAT) (Veličković et al., 2017) to the patient-specific graph, allowing the current admission embedding $\mathcal{S}\text{sum}^{(t)}$ to be refined through attention-weighted aggregation of information from past admissions. Let $\mathcal{H} = \{\mathcal{S}\text{sum}^{(i)}\}_{i=1}^{t}$ denote the set of hospitalization embeddings, the GAT layer computes contextualized updates (eq. 7) by attending to clinically relevant neighbors.

$$\mathcal{S}_{\text{sum}}^{(t)'} = \sigma\left(\sum_{j=1}^{t} \alpha_{tj} W \mathcal{S}_{\text{sum}}^{(j)} \right) \tag{7}$$

where W is a learnable weight matrix, σ is a nonlinear activation function, and α_{tj} is the attention weight from node t to neighbor j, defined as:

$$\alpha_{tj} = \frac{\exp(e_{tj})}{\sum_{k=1}^{t} \exp(e_{tk})} \tag{8}$$

$$e_{tj} = \text{LeakyReLU}\Big(a^\top [W \mathcal{S}_{\text{sum}}^{(t)} \mid W \mathcal{S}_{\text{sum}}^{(j)}] \Big) + \gamma A_{tj} \tag{9}$$

where a is a learnable attention vector and γ is a hyperparameter. The final summary representation $\mathcal{S}_{\text{sum}}^{(t)\prime}$ is passed through a feedforward network followed by a softmax activation to obtain the readmission probability. To address the severe class imbalance inherent in hospital readmission data, we adopt the Focal Loss (Lin et al., 2017), defined as $\mathcal{L}_{\text{focal}}(p_t) = -\alpha(1 - p_t)^{\gamma} \log(p_t)$, where p_t denotes the predicted probability corresponding to the ground-truth class, α balances positive and negative instances, and γ modulates the emphasis on hard-to-classify samples. The model is trained using this objective and optimized with Adam optimizer to ensure stable and effective convergence.

4 Experiment

Dataset: As our primary data source, we have used two publicly available critical-care datasets, MIMIC-III v1.4 and MIMIC-IV v2.2. An admission is labeled as a *Readmission* if the patient is hospitalized again within 30 days of the index discharge.

- From MIMIC-III, we build a readmission dataset comprising 42,573 hospital admissions from 33,107 patients, of which 2,794 admissions are labeled as 30-day readmissions.
- From MIMIC IV, we construct a cohort of 9,947 admissions from 4,539 patients, including 2,709 admissions labeled as 30-day readmissions.

Both datasets are partitioned into training (60%), validation (20%), and test (20%) splits at the patient level to ensure that all admissions for a given patient appear exclusively in a single split.

Experiment Environment and Evaluation Metrics: All experiments were conducted on a server with an NVIDIA Tesla V100 GPU (32 GiB), 9 vCPUs, and 60 GiB RAM. The model was implemented in PyTorch. The final hyperparameters for our prediction model include a learning rate of 0.001, dropout rate of 0.1, 100 training epochs, and 2 GAT layers; training uses the Adam optimizer with Focal Loss (α : 0.25 & γ : 2.0). Model performance is evaluated using Accuracy (Acc), Precision (P), Recall (R), F1-score (F1), and Area Under the ROC Curve(ROC-AUC).

Baselines: We compare our proposed framework, *DisGraph-RP*, against several strong baselines. First, we benchmark against several widely used LLMs such as, Bio-Medical-LLaMA-3-8B (Con, 2024), BioMistral-7B (Labrak et al., 2024), GPT-4 (Waisberg et al., 2023), and GPT-5 (Hou et al., 2025), using zero-shot prompting due to token-length constraints that make few-shot prompting infeasible for long discharge summaries. We additionally fine-tune ClinicalBERT (Huang et al., 2019) for the readmission prediction task. To assess the impact of temporal modeling, we also compare against two time-aware architectures, T-LSTM (Mou et al., 2019) and HiTANet (Luo et al., 2020), both of which incorporate prior admissions.

Moreover, we perform an ablation study to isolate the contribution of each module. **DisGraph-RP w/o CE** removes the contextualized encoder, using only ontology-based section representations. **DisGraph-RP w/o OE** drops the ontology-based embedding, retaining only contextualized features. **DisGraph-RP w/o GT** excludes the Graph-Augmented Temporal module, evaluating prediction using only the current discharge summary.

4.1 Results and Discussion

The comparative results of all models are presented in Table 1, with the best scores highlighted in bold. *DisGraph-RP* consistently surpasses all state-of-the-art readmission prediction baselines across every metric, and the ablation results further validate the contribution of each component in the architecture. Since the dataset exhibits a high degree of class imbalance, accuracy is not a reliable performance metric, as it remains artificially high in most cases due to the dominance of the majority class.

Notably, incorporating the Graph-Augmented Temporal Module substantially improves prediction performance, boosting the overall F1 score by 42% on the MIMIC-III cohort and 11% on MIMIC-IV. This underscores the importance of modeling temporal dependencies in prior hospitalizations for identifying high-risk patients. We further observed that the evaluated LLMs, despite achieving high recall, exhibit markedly low precision, indicating a strong tendency to over-predict readmissions. Such bias toward the positive class limits their reliability in real clinical decision-making.

Furthermore, Figure 4 in Appendix D shows how our model assigns differentiated attention weights to discharge summary sections, prioritizing task relevant content for generating more informative embeddings. Figure 5 in Appendix E presents a t-SNE visualization of Section-Aware embeddings for patients with Pneumonia and Cardiovascular diseases

Table 1: Performance comparison of baseline models and the proposed *DisGraph-RP* model for readmission prediction task.

Category	Model	MIMIC III					MIMIC IV				
		Acc*	P	R	F1	ROC-AUC	Acc*	P	R	F1	ROC-AUC
w/o Temporal context	ClinicalBERT	0.76	0.08	0.29	0.12	0.56	0.52	0.20	0.22	0.21	0.65
	BioMistral-7B	0.81	0.06	0.15	0.09	0.51	0.42	0.26	0.73	0.42	0.61
	Bio-LLaMA	0.65	0.09	0.63	0.16	0.67	0.51	0.35	0.81	0.49	0.72
	GPT-4	0.78	0.11	0.30	0.14	0.59	0.70	0.44	0.12	0.20	0.57
	GPT-5	0.79	0.09	0.31	0.13	0.59	0.51	0.34	0.71	0.45	0.56
	DisGraph-RP w/o GT	0.89	0.29	0.44	0.34	0.68	0.79	0.69	0.52	0.59	0.79
with Temporal context	T-LSTM	0.95	0.63	0.47	0.54	0.72	0.68	0.42	0.30	0.36	0.78
	HiTANet	0.93	0.42	0.51	0.46	0.72	0.72	0.52	0.35	0.42	0.81
	DisGraph-RP w/o OE	0.93	0.42	0.57	0.48	0.76	0.78	0.61	0.66	0.63	0.81
	DisGraph-RP w/o CE	0.95	0.57	0.59	0.54	0.78	0.79	0.63	0.69	0.66	0.82
	DisGraph-RP	**0.97**	**0.76**	**0.75**	**0.76**	**0.87**	**0.82**	**0.71**	**0.69**	**0.70**	**0.84**

(CVD), revealing distinct clusters that highlight the encoder's ability to capture discriminative and meaningful features.

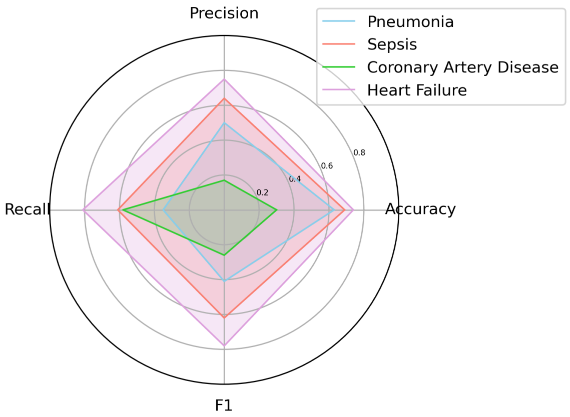


Figure 2: Case study: Comparative performance of DisGraph-RP across different disease types in terms of accuracy, precision, recall, and F1-score.

Although our framework is disease-agnostic and generalizable across chronic and acute conditions, we further evaluate its performance on specific disease groups. The four most frequent diagnoses in our dataset are Pneumonia, Sepsis, Coronary Artery Disease (CAD), and Heart Failure. Analysis reveals notable variation in model behavior across these conditions. As shown in Figure 2, for Pneumonia patients the model exhibits low recall (35%), indicating under-prediction of readmissions. In contrast, for CAD cases, recall is higher (58%) but precision drops to 17%, suggesting an inclination to over-predict. These findings underscore the need for condition-specific calibration when deploying readmission prediction models in practice.

Moreover, to better understand the limitations of our framework, we conducted an error analysis by examining false positive and false negative cases on the test set. A substantial portion of false positives cases, where the model incorrectly predicted readmission involved discharge summaries characterized by high clinical complexity, including multiple procedures such as 'bypass grafting' and 'mitral valve replacement', or ambiguous discharge instructions. Although these cases were clinically severe, the patients did not return within 30 days. These findings suggest that the model tends to associate clinical severity with readmission risk, potentially overestimating risk in well-managed cases. Conversely, false negatives often stemmed from incomplete summaries or lack of prior admissions, limiting the model's ability to capture latent clinical risks. These findings highlight the need for comprehensive documentation and longitudinal context for accurate prediction.

5 Conclusion

In this paper, we presented *DisGraph-RP*, a framework that integrates section-aware contrastive encoding with ontology-guided and contextualized discharge-summary representations, complemented by a graph-augmented temporal module to model longitudinal patient history. Experiments show that DisGraph-RP consistently outperforms strong baselines, including domain-specific LLMs, which are shown to be less reliable for clinical decision-making tasks such as readmission prediction. Future work will extend the framework with multimodal clinical data to further improve patient modeling and predictive accuracy.

6 Limitations

This work relies heavily on discharge summaries whose structure and sectioning vary widely across healthcare organizations, making the note-processing pipeline sensitive to formatting inconsistencies and limiting generalizability. Most of the real world data including our datasets is highly imbalanced, with a strong bias toward the majority (non-readmitted) class, which affects model stability and reduce predictive performance for minority cases. Additionally, the model depends on accurate and complete past admission information; however, prior hospital records from external institutions are often unavailable, leading to incomplete temporal histories and potentially underestimated readmission risk. Furthermore, the approach may capture institution-specific linguistic patterns that do not fully transfer to other clinical environments, and errors in clinical text can propagate through the embedding and attention modules, introducing noise into the final predictions.

References

2024. Contactdoctor-bio-medical: A high-performance biomedical language model. https://huggingface.co/ContactDoctor/Bio-Medical-Llama-3-8B.

Fatemeh Amrollahi, Supreeth P Shashikumar, Angela Meier, Lucila Ohno-Machado, Shamim Nemati, and Gabriel Wardi. 2022. Inclusion of social determinants of health improves sepsis readmission prediction models. *Journal of the American Medical Informatics Association*, 29(7):1263–1270.

Alan R Aronson. 2001. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In *Proceedings of the AMIA Symposium*, page 17. American Medical Informatics Association.

Awais Ashfaq, Anita Sant'Anna, Markus Lingman, and Sławomir Nowaczyk. 2019. Readmission prediction using deep learning on electronic health records. *Journal of biomedical informatics*, 97:103256.

Robert E Burke and Eric A Coleman. 2013. Interventions to decrease hospital readmissions: keys for cost-effectiveness. *JAMA internal medicine*, 173(8):695–698.

Xiongcai Cai, Oscar Perez-Concha, Enrico Coiera, Fernando Martin-Sanchez, Richard Day, David Roffe, and Blanca Gallego. 2016. Real-time prediction of mortality, readmission, and length of stay using electronic health record data. *Journal of the American Medical Informatics Association*, 23(3):553–561.

Wendy W Chapman, Will Bridewell, Paul Hanbury, Gregory F Cooper, and Bruce G Buchanan. 2001. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of biomedical informatics*, 34(5):301–310.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Ankur Gupta and Gregg C Fonarow. 2018. The hospital readmissions reduction program—learning from failure of a healthcare policy. *European journal of heart failure*, 20(8):1169–1174.

Yu Hou, Zaifu Zhan, Min Zeng, Yifan Wu, Shuang Zhou, and Rui Zhang. 2025. Benchmarking gpt-5 for biomedical natural language processing. *arXiv preprint arXiv:2509.04462*.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*.

Yinan Huang, Ashna Talwar, Ying Lin, and Rajender R Aparasu. 2022. Machine learning methods to predict 30-day hospital readmission outcome among us adults with pneumonia: analysis of the national readmission database. *BMC Medical Informatics and Decision Making*, 22(1):288.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, and 1 others. 2023. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. Biomistral: A collection of open-source pretrained large language models for medical domains. *arXiv preprint arXiv:2402.10373*.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

Xiong Liu, Yu Chen, Jay Bae, Hu Li, Joseph Johnston, and Todd Sanger. 2019. Predicting heart failure readmission from clinical notes using deep learning. In *2019 IEEE international conference on bioinformatics and biomedicine (BIBM)*, pages 2642–2648. IEEE.

Ning Lu, Kuo-Cherh Huang, and James A Johnson. 2016. Reducing excess readmissions: promising effect of hospital readmissions reduction program in us hospitals. *International Journal for Quality in Health Care*, 28(1):53–58.

Junyu Luo, Muchao Ye, Cao Xiao, and Fenglong Ma. 2020. 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.

Saeed Mehrabi, Anand Krishnan, Sunghwan Sohn, Alexandra M Roch, Heidi Schmidt, Joe Kesterson, Chris Beesley, Paul Dexter, C Max Schmidt, Hongfang Liu, and 1 others. 2015. Deepen: A negation detection system for clinical text incorporating dependency relation into negex. *Journal of biomedical informatics*, 54:213–219.

Luntian Mou, Pengfei Zhao, Haitao Xie, and Yanyan Chen. 2019. T-lstm: A long short-term memory neural network enhanced by temporal information for traffic flow prediction. *Ieee Access*, 7:98053–98060.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. Scispacy: fast and robust models for biomedical natural language processing. *arXiv preprint arXiv:1902.07669*.

Mitchell A Psotka, Gregg C Fonarow, Larry A Allen, Karen E Joynt Maddox, Mona Fiuzat, Paul Heidenreich, Adrian F Hernandez, Marvin A Konstam, Clyde W Yancy, and Christopher M O'Connor. 2020. The hospital readmissions reduction program: nationwide perspectives and recommendations: a jacc: heart failure position paper. *JACC: Heart Failure*, 8(1):1–11.

N Reimers. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Juan C Rojas, Kyle A Carey, Dana P Edelson, Laura R Venable, Michael D Howell, and Matthew M Churpek. 2018. Predicting intensive care unit readmission with machine learning using electronic health record data. *Annals of the American Thoracic Society*, 15(7):846–853.

Peri L Schuyler, William T Hole, Mark S Tuttle, and David D Sherertz. 1993. The umls metathesaurus: representing different views of biomedical concepts. *Bulletin of the Medical Library Association*, 81(2):217.

Sheojung Shin, Peter C Austin, Heather J Ross, Husam Abdel-Qadir, Cassandra Freitas, George Tomlinson, Davide Chicco, Meera Mahendiran, Patrick R Lawler, Filio Billia, and 1 others. 2021. Machine learning vs. conventional statistical models for predicting heart failure readmission and mortality. *ESC heart failure*, 8(1):106–115.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. *arXiv preprint arXiv:1710.10903*.

Ethan Waisberg, Joshua Ong, Mouayad Masalkhi, Sharif Amit Kamran, Nasif Zaman, Prithul Sarker, Andrew G Lee, and Alireza Tavakkoli. 2023. Gpt-4: a new era of artificial intelligence in medicine. *Irish Journal of Medical Science (1971-)*, 192(6):3197–3200.

Yasi Wang, Hongxun Yao, and Sicheng Zhao. 2016. Auto-encoder based dimensionality reduction. *Neurocomputing*, 184:232–242.

A Example of a Discharge Summary

Example Discharge Summary

Admission Date: —- **Discharge Date:** —-
Date of Birth: —- **Sex:** F
Service: MEDICINE
Allergies: Haldol
Chief Complaint: pneumonia, lethargy, sepsis
Major Surgical or Invasive Procedure: none
History of Present Illness: 35F with pneumonia who presented today from daycare after her healthcare providers noted that she was lethargic. They were initially unable to obtain a blood pressure...... A CT-A was negative for a PE. The patient was transferred to the MICU for further mgmt.
Past Medical History: Anemia
Social History: Patient lives at home with sister and brother. Father passed away.
Physical Exam: T 97.7, HR 65–68, BP 91–97/61–63, R 14–21... ABD: flat, soft, NT, ND, +BS...
Pertinent Results: CT-A IMPRESSION: Poorly defined opacities within the lungs bilaterally... may have been infection. Will continue infectious workup and treatment.
Discharge Medications: Amiodarone 200 mg Tablet Sig, Amoxicillin-Pot Clavulanate 250–62.5 mg/5 mL...
Discharge Disposition: Home With Service
Discharge Diagnosis: supraventricular

tachycardia, pneumonia
Discharge Condition: stable and improving
Discharge Instructions: You will be discharged home today ...you will be sent home with two new medications. If you develop any chest pain, shortness of breath, fever, or any other symptoms, please call Dr. ** or return to the emergency department.
Followup Instructions: Please follow up with Dr. ** within the next week. Dr. ** should arrange a follow-up with Cardiology within the next month.

B System Instruction for Discharge Summary Segmentation

Prompt template for Discharge Summary Segmentation

Task Description: You are a specialized medical expert. Given a patient discharge summary, extract all clinically relevant information into predefined subsections. Each subsection should be a concise paragraph, including all key details. If a subsection is missing, output *"Not mentioned"*.
Required Subsections:

1. **Administrative Information**: Details identifying the patient and the hospital stay, including admission/discharge dates, date of birth, sex, the service they were on, and the attending physician.

2. **Chief Complaint**: The primary reason or main symptom(s) that led to the patient's admission.

3. **Allergies**: A clear statement of the patient's known allergies or if none are recorded.

4. **Major Procedures**: A summary of any significant surgical or invasive procedures performed during the hospital stay.

5. **History of Present Illness**: A chronological narrative of the patient's symptoms and conditions leading up to and during the initial phase of admission. Do not include past medical history.

6. **Past Medical History**: A concise overview of the patient's relevant chronic or significant past health conditions.

7. **Social History**: Information about the patient's lifestyle, family history, living situation, habits (e.g., smoking, alcohol), and occupation.

8. **Physical Exam**: A summary of the patient's physical findings upon admission and/or discharge, focusing on pertinent observations.

9. **Pertinent Results**: A brief overview of all significant laboratory, imaging, and other diagnostic test results.

10. **Medications During Treatment**: A summary of key medications administered to the patient during their hospitalization, explicitly excluding discharge medications.

11. **Brief Hospital Course**: A concise, problem-oriented summary of the patient's overall progress, management, and significant events throughout the hospital stay.

12. **Discharge Medications**: A list of medications prescribed to the patient upon their release from the hospital.

13. **Discharge Condition**: A brief description of the patient's overall status and health at the time of discharge.

14. **Discharge & Follow-up Instructions**: Key instructions given to the patient for post-discharge care, including activity restrictions, wound care, medication adherence, and scheduled follow-up appointments.

Output Format: Return strictly in following JSON format:

```
{
 "Administrative Information": "...",
  "Chief Complaint": "...",
  "Allergies": "...",
  "Major Procedures": "...",
 "History of Present Illness": "...",
  "Past Medical History": "...",
  "Social History": "...",
  "Physical Exam": "...",
  "Pertinent Results": "...",
 "Medications During Treatment": "...",
  "Brief Hospital Course": "...",
  "Discharge Medications": "...",
  "Discharge Condition": "...",
 "Discharge & Follow-up Instructions":
}
```

No additional explanations or reasoning should be included.

C Pre-processing Clinical Note: Extraction of clinical details

Clinical notes vary greatly in style and content. Some document only symptoms, while others detail absences of symptoms, adverse reactions, psychological states, and appetite changes, often using non-standard terminology and abbreviations. To manage this variability, we added a processing layer that uses biomedical dictionaries to create a structured representation of clinical details, as shown in Figure 3. Details of this processing pipeline are presented below.

C.1 Entity Extraction

We employed two BioNER tools, ScispaCy (Neumann et al., 2019) and Metamap (Aronson, 2001), for the extraction of patients' health conditions from clinical notes. The pre-trained scispaCy model, was utilized for recognizing "disease" names. We use Metamap to identify eight medical entities, including "Sign or Symptom", "Disease or Syndrome", "Acquired Abnormality", "Anatomical Abnormality", "Congenital Abnormality", "Injury or Poisoning", "Mental Process", and "Mental or Behavioral Dysfunction" within these notes.

C.2 Detecting Negations

Subsequently, the Negex algorithm (Chapman et al., 2001), designed to identify negative modifiers such as "no", "not", etc., is employed to detect negative mentions of entities within the text. The initial list was expanded to encompass commonly occurring negation concepts like 'deny", "refuse", "absent", "decline", etc., frequently encountered in clinical notes. For instance, in a sentence like "The patient has shortness of breath but denies any chest pain", the two symptoms identified would be "shortness of breath" and "neg chest pain." These negative symptoms play a crucial role in providing a comprehensive understanding of individual patients.

C.3 Clinical Entity Normalization

Clinical notes often encompass diverse non-standard terminology, abbreviations, formats, and coding systems to represent clinical concepts. For instance, a single medical condition like "Hemorrhage" may be referred to as "Bleeding", "Blood loss", or "oozing of blood" by different healthcare professionals. To address this variability, we have standardized all extracted entities using the UMLS Metathesaurus (Schuyler et al., 1993), which includes a comprehensive list of terms and assigns a "Concept Unique Identifier (CUI)" to each. However, we observed that certain entities did not yield an exact match with any UMLS concept. To resolve this issue, we employed the 'all-mpnet-base-v2' SBERT model (Reimers, 2019) to compute the semantic textual similarity between the unmatched entities and the retrieved UMLS concepts. The SBERT model generates embeddings for each entity and calculate the similarities between them. For entity pairs with a similarity score exceeding a empirically defined threshold of 0.9, we considered the terms to be semantically similar. For entities that could not be mapped to any UMLS concept, we created unique identifiers to ensure that no health condition was overlooked.

C.4 Encoding the clinical notes

Let $V = \{d_1, d_2, ..., d_{|V|}\}$ denote the comprehensive vocabulary of CUIs of all extracted clinical entities, including descriptions of diseases, symptoms, injuries, abnormalities and so on, relevant to the study. For a patient p, the health condition at a timestamp t is represented by a vector $H_t^p =< h_i >, i = 1, 2, ..., |V|$, and

$$h_i = f(d_i) = \begin{cases} 1 & \text{if } d_i \text{ present,} \\ -1 & \text{if } d_i \text{ negative,} \\ 0 & \text{if } d_i \text{ absent.} \end{cases}$$

However, the high number of unique clinical entities and the variability in individual manifestations result in vectors that are often high-dimensional and sparse. To address this challenge,

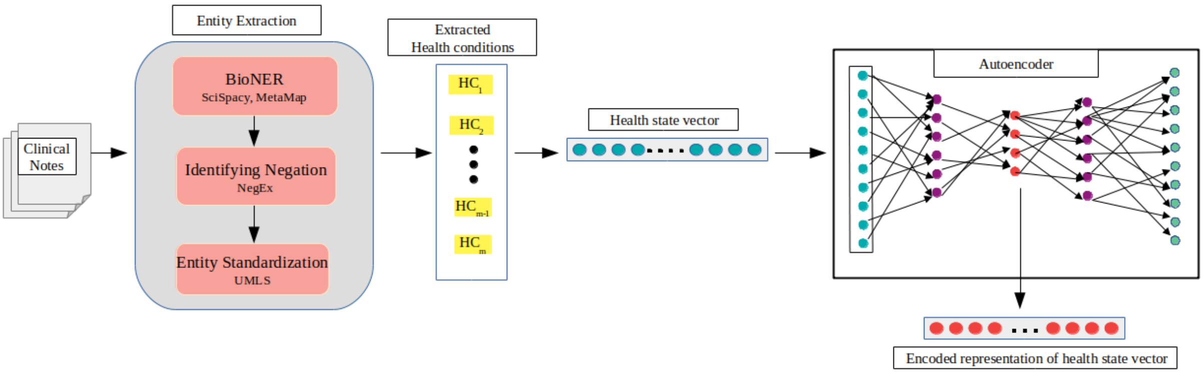


Figure 3: Overview of the process for extraction and representation of patient health conditions from clinical notes.

we employed a standard autoencoder (AE) framework (Wang et al., 2016) to obtain a dense, lower-dimensional representation. The AE is an unsupervised model where the "encoder" network compresses the input data by capturing its key features, and the "decoder" network reconstructs the original data from this compressed form, aiming to preserve the essential information.

Let $H_t^p \in \mathbb{R}^{1 \times |V|}$ represent health condition at stage t of a patient p. The AE optimizes the following loss function to minimize the reconstruction error:

$$\mathcal{L}(\mathbf{H_t^P}, \hat{\mathbf{H_t^P}}) = \frac{1}{|V|} \sum_{i=1}^{|V|} [h_i - g_\phi(f_\theta(h_i))]^2$$

, where f_θ is the encoder function parameterized by θ, g_ϕ is the decoder function parameterized by ϕ, and $\hat{\mathbf{H_t^P}}$ is the reconstructed input.

In our experiment, the encoder was implemented as a multi-layer neural network that mapped the input data into a low-dimensional latent space, while the decoder adopted a mirrored architecture to reconstruct the original input. The model was trained using the reconstruction error as the loss function, and the Adam optimizer with a learning rate of 0.01 was employed to ensure convergence. The resulting autoencoded representations, denoted as $EH_t^p = f_\theta(H_t^p)$, offer a more compact and informative health vector representation for patient p at timestamp t.

D Attention distribution across discharge summary sections

Figure 4 illustrates attention distribution across discharge summary sections as assigned by the Section-Aware Encoder model.

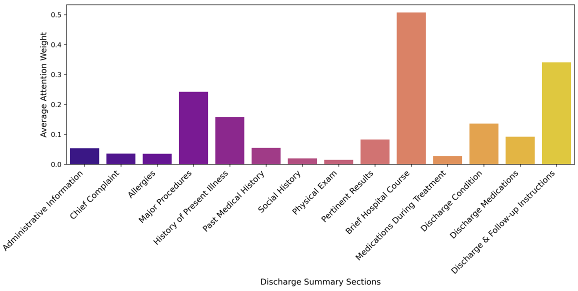
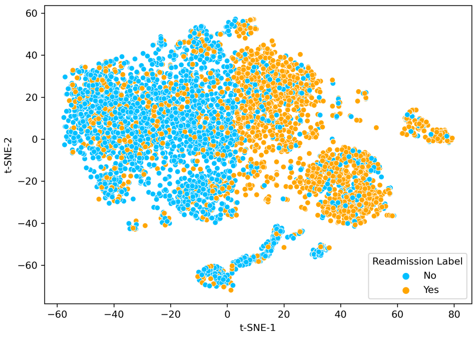


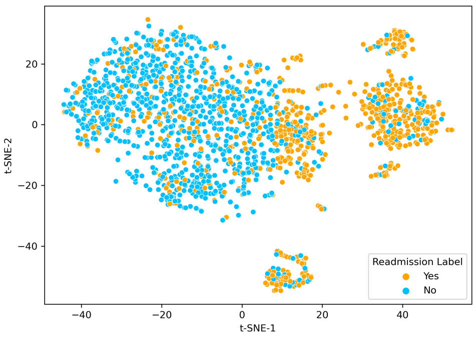
Figure 4: Attention distribution across discharge summary sections as assigned by the Section-Aware Encoder model.

E t-SNE visualizations of discharge summary embeddings

Figure 5 illustrates the t-SNE visualizations of discharge summary embeddings for patients with cardiovascular disease (CVD) (Fig. a) and pneumonia (Fig. b). The plots reveal clear and well-defined separations between readmitted and non-readmitted patient groups, indicating distinct embedding patterns associated with readmission status.



(a)



(b)

Figure 5: t-SNE visualizations of discharge summary embeddings of CVD (fig. a) and Pneumonia (fig. b) patients.