International Journal of Innovative
Computing, Information and Control
Volume **22**, Number **3**, June **2026**

ICIC International ©2026 ISSN 1349-4198
pp. **687–699**

INTEGRATING KNOWLEDGE GRAPHS AND DEEP LEARNING FOR DRUG RECOMMENDATION SYSTEM

Nhu Nguyen, Dat Nguyen and Phuoc Tran*

Natural Language Processing and Knowledge Discovery Research Group
Faculty of Information Technology
Ton Duc Thang University
19 Nguyen Huu Tho Street, Tan Hung Ward, Ho Chi Minh City 70000, Vietnam
{ 241805005; 51900030 }@student.tdtu.edu.vn; *Corresponding author: tranthanhphuoc@tdtu.edu.vn

Received September 2025; revised January 2026

Abstract. *Prescription errors and adverse drug interactions remain critical challenges in healthcare, particularly in developing countries like Vietnam, where medical data is often fragmented and unstructured. This study proposes a symptom-based drug recommendation system to assist physicians and pharmacists in making accurate prescriptions. The system utilizes artificial intelligence and deep learning to analyze multidimensional medical data, including symptoms, drug side effects, comorbidities, and patient feedback. Three approaches are explored: Retrieval-Augmented Generation (RAG), Knowledge-Augmented Generation (KAG), and Graph Retrieval-Augmented Generation (GR AG). RAG integrates retrieval-based information with generative models, KAG incorporates domain knowledge for better reasoning, and GRAG leverages knowledge graphs with Graph Neural Networks (GNNs) such as Graph Convolutional Networks (GCN), Graph Attention Networks (GAT), Message Passing Neural Networks (MPNN), Relational Graph Convolutional Networks (R-GCN), GraphSAGE, and Graph Transformers (GTs). These GNN models predict drug candidates and risk scores, which inform RAG queries for external knowledge. Evaluations on MIMIC-III (subset) and Medical Recommendation System datasets show GTs achieving 92.77% accuracy in symptom-to-disease classification, with GRAG yielding an F1-score of 0.407 in drug recommendation (19.0% improvement over RAG) and reducing DDI rate to 2.8%. This system could reduce errors, enhance personalized prescriptions, and optimize healthcare in resource-constrained settings.*

Keywords: Artificial intelligence, Drug recommendation, Prescription, Symptoms, Drug interaction, Knowledge graph

1. **Introduction.** Prescription errors and Adverse Drug Interactions (ADIs) represent significant challenges to global healthcare systems, contributing to approximately 7% of hospitalizations and up to 20% of preventable adverse drug events [1]. In developing countries such as Vietnam, these issues are exacerbated by systemic barriers, including fragmented medical data, limited adoption of standardized Electronic Health Records (EHRs), and a shortage of trained healthcare professionals [2]. A study from the Ho Chi Minh University of Medicine and Pharmacy reports that 15% of prescriptions in rural Vietnamese clinics contain errors, primarily due to misdiagnosis or inadequate screening for drug interactions [3]. These errors result in reduced treatment efficacy, increased healthcare costs, and heightened risks of Adverse Drug Reactions (ADRs), posing a critical public health concern.

The primary challenge addressed in this study is the lack of accurate, safe, and scalable drug recommendation systems capable of processing unstructured and fragmented

DOI: 10.24507/ijicic.22.03.687

medical data, particularly in resource-constrained settings. Existing systems are limited by three key factors: 1) challenges in disambiguating overlapping symptoms associated with multiple diseases, 2) inadequate modeling of complex relationships among symptoms, drugs, and diseases, and 3) insufficient mechanisms for detecting and preventing Drug-Drug Interactions (DDIs) [4]. These limitations lead to suboptimal prescriptions, particularly in regions like Vietnam, where manual prescription processes predominate due to limited technological infrastructure.

To address these challenges, this study proposes the development of a Symptom-Based Drug Recommendation System (SBDRS) that leverages Artificial Intelligence (AI) and deep learning to enhance prescription accuracy and safety. The objectives of this research are threefold:

1) To design a system that accurately maps symptoms to appropriate drugs while considering patient-specific factors, such as comorbidities and allergies;
2) To ensure safety by incorporating mechanisms to detect potential Drug-Drug Interactions (DDIs) and Adverse Drug Reactions (ADRs);
3) To enable scalability in environments with fragmented or unstructured medical data, making the system viable for deployment in developing countries.

To achieve these goals, we explore three distinct approaches: 1) Retrieval-Augmented Generation (RAG), which integrates information retrieval with generative models to access large-scale medical databases [5]; 2) Knowledge-Augmented Generation (KAG), which enhances reasoning by incorporating structured medical knowledge to handle complex symptom profiles [6], 3) Graph Retrieval-Augmented Generation (GRAG), which utilizes knowledge graphs to model intricate symptom-drug-disease relationships, offering improved accuracy and safety [7]. Unlike existing Graph-RAG hybrids, GRAG introduces a safety-aware probabilistic scoring mechanism and a symptom-oriented contextual representation module, enabling explicit safety control and improved robustness in fragmented data settings.

This study focuses on the GRAG approach, which combines graph-based reasoning with deep learning to model multi-hop relationships (e.g., symptom-to-disease-to-drug linkages) and perform real-time safety verification. Preliminary evaluations using the MIMIC-III and Medical Recommendation System datasets demonstrate that GRAG outperforms baseline methods, achieving a 19% improvement in recommendation accuracy and a 78.1% relative reduction in DDI rate. These findings suggest that GRAG has the potential to serve as a transformative clinical decision-support tool, particularly in resource-constrained settings. By reducing prescription errors, enhancing therapeutic personalization, and improving healthcare outcomes, the proposed system could significantly contribute to addressing the global burden of medication-related errors, especially in developing nations.

2. **Related Works.**

To provide a clear context for our proposed system, we critically analyze existing literature based on the three core challenges identified in the Introduction: 1) accuracy in symptom-disease mapping, 2) safety regarding drug interactions, and 3) scalability in fragmented data environments.

2.1. **Challenges in symptom-disease mapping and disambiguation.**

Early recommendation systems primarily relied on collaborative filtering and matrix factorization to predict efficacy based on historical patterns. For instance, Wang and Wu [8] successfully applied data mining techniques, such as decision trees and collaborative filtering, to building recommendation systems in university libraries. While effective in structured environments, such traditional data-driven approaches often struggle with the "cold-start"

problem and data sparsity. In clinical scenarios, these methods face even greater challenges due to the semantic complexity of overlapping symptoms. Existing deep learning models, such as Convolutional Neural Networks (CNNs) [9] and LSTMs [10], have improved feature representation from treatment sequences but often lack the reasoning capability to disambiguate vague patient descriptions (e.g., differentiating "fever" in flu vs. dengue). This limitation necessitates a model capable of context-aware symptom mapping, which traditional collaborative filtering cannot fully address.

2.2. **Limitations in safety and Drug-Drug Interaction (DDI) modeling.** Recent medication recommendation models like GAINET [11] and Graph Neural Network (GNN)-based DDI predictors [12] reveal persistent limitations in explicitly controlling Drug-Drug Interactions (DDIs), despite leveraging patient EHRs and molecular knowledge graphs [13]. While GAINET achieves 85.2% accuracy in biochemical predictions with only 12.4% Hits@3 for true positives [11] and GNNs reach 92.7% classification accuracy on OGBL-DDIs but just 8.3% Hits@10 [12], they struggle with ranking true positive interactions and fail to integrate multimodal data (e.g., only 67% molecular-text fusion coverage) [14, 15], leading to suboptimal safety in polypharmacy where 15-20% of recommendations still violate moderate-severe DDIs [13].

For instance, memory-augmented networks like GAMENet [16] – retrospectively analyzed in 2025 studies – reduce DDI rates by only ∼3.6% (from 28.4% to 24.8%) [12], and SafeDrug variants by up to 19.43% reduction (to 15.2% baseline), our GRAG achieves a 78.1% relative reduction (to 2.8%) by integrating explicit retrieval mechanisms. Implicit modeling cannot guarantee zero-risk combinations, with 7-11% false negatives on novel drug pairs due to class imbalance (1:42 DDI ratio) [14] and poor generalization (22% F1-score drop on unseen drugs) [15]. Current approaches prioritize effectiveness (e.g., 11.2% Jaccard gain) over comprehensive safety [12], underscoring the need for explicit, interpretable DDI constraints (targeting $< 5\%$ violation rate) without sacrificing recommendation quality [11, 13].

2.3. **Scalability and robustness in fragmented data environments.** Deploying recommendation systems in developing countries requires robustness against fragmented and unstructured data. Traditional graph neural networks often suffer from scalability issues (e.g., $O(n^2)$ complexity in GAMENet [16]). Our GRAG uses RAG for efficiency in sparse data. Recent advancements in Retrieval-Augmented Generation (RAG) offer a potential solution for handling domain-specific data. Hong and Kim [17] demonstrated an innovative application of RAG-enhanced small LLMs for closed-domain question answering, using vector databases to retrieve relevant operational knowledge. While their vector-based retrieval is efficient for textual queries, it may overlook the complex, multi-hop causal relationships inherent in medical diagnoses. Therefore, our study proposes a Graph RAG (GRAG) approach, which combines the generative power of LLMs (as used by Hong and Kim) with the structured reasoning of KGs to ensure robust performance even when local clinical data is sparse or fragmented.

3. Approach.

3.1. **Overview of the proposed GRAG framework.** This method introduces a robust framework focusing on Graph Retrieval-Augmented Generation (GRAG). It integrates Knowledge Graphs (KGs) with deep learning to create an efficient, accurate, and safe recommendation system.

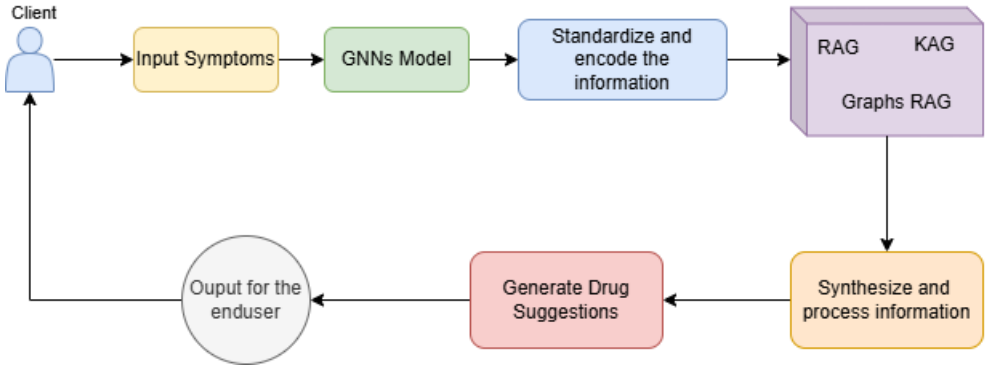


FIGURE 1. Overview of the proposed framework

3.2. **Symptom-Oriented Disambiguation Module (SDM).** A critical challenge in diagnosis is the "overlapping symptom" phenomenon, where a single symptom (e.g., "high fever") is associated with multiple pathologies. To address this, GRAG incorporates a Symptom-Oriented Disambiguation Module (SDM) leveraging the multi-head self-attention mechanism of Graph Transformers.

Unlike static matching, this module dynamically calculates the contextual importance of each symptom relative to the entire patient profile. Formally, let S be the set of input symptoms. The module updates the latent embedding of each symptom h_{s_i} based on its neighbors:

$$h'_{s_i} = \text{Attention}(Q_i, K_S, V_S) = \text{softmax}\left(\frac{Q_i K_S^\top}{\sqrt{d_k}}\right) V_S \tag{1}$$

where $Q_i = h_{s_i} W_Q$, $K_S = H_S W_K$, and $V_S = H_S W_V$ are linear projections of symptom embeddings, and d_k denotes the key dimension.

Rather than performing explicit disease-level disambiguation, the proposed SDM enhances symptom representations in a context-aware manner, thereby implicitly reducing ambiguity during downstream disease inference and drug recommendation.

3.3. **Graph-based recommendation formulation.** To rigorously quantify the recommendation process, we formulate drug selection as a link prediction task within a heterogeneous knowledge graph.

3.3.1. *Node embedding propagation.* We employ a message passing neural network (MPNN) to propagate information over the graph. For clarity, we present a generic message-passing formulation; in practice, the aggregation is instantiated using attention-based mechanisms when Graph Transformers are employed. At layer $l+1$, the embedding of a node v is updated by aggregating information from its neighbors $\mathcal{N}(v)$ as follows:

$$h_v^{(l+1)} = \sigma\left(W^{(l)} \cdot \text{AGG}\left(\left\{h_u^{(l)} | u \in \mathcal{N}(v)\right\}\right) + b^{(l)}\right) \tag{2}$$

where $\sigma(\cdot)$ denotes a nonlinear activation function.

3.3.2. *Recommendation scoring with safety constraints.* To ensure generalizability across diverse medical conditions, we formulate the drug suitability score $S(r, S)$ using a probabilistic framework that balances therapeutic efficacy with safety constraints.
Efficacy Score. The efficacy score is computed by aggregating the likelihood of drug r being effective for the most probable diseases inferred from the symptom set S:

$$S_{\text{efficacy}}(r, S) = \sum_{d \in \mathcal{D}_k} P(r|d) \cdot P(d|S) \tag{3}$$

where:

- $\mathcal{D}_k$ denotes the top-k inferred diseases with the highest posterior probabilities.
- $P(d|S)$ is normalized over the top-k inferred diseases.
- $P(r|d)$ represents the likelihood of drug r being effective for disease d.

Safety-Aware Score. To explicitly enforce safety, consistent with the RAG-based verification stage described in Section 3.4, we apply a multiplicative safety decay mechanism:

$$S_{\text{final}}(r, S) = S_{\text{efficacy}}(r, S) \times \delta^{\sum_k \mathbb{I}(r, r_k)} \tag{4}$$

where $\delta \in (0, 1)$ is a tunable decay factor, $\mathbb{I}(r, r_k)$ is an indicator function that returns 1 if an adverse drug-drug interaction or contraindication between drugs r and r_k is detected, and 0 otherwise, and r_k denotes drugs involved in retrieved contraindications or adverse interactions associated with candidate drug r. In this study, all detected drug-drug interactions are treated with equal severity for simplicity, without differentiating interaction types or clinical severity levels.

In practice, both $P(d|S)$ and $P(r|d)$ are implemented as normalized scores derived from GNN link-prediction logits using softmax normalization, rather than strict Bayesian probabilities.

3.4. **Integration of GNN prediction with RAG retrieval.** The GRAG framework adopts a three-phase sequential pipeline that synergizes GNN-based candidate generation with retrieval-augmented safety refinement. GNN predictions are used as seeds for targeted retrieval, enabling dynamic verification and adjustment against external clinical knowledge sources. A schematic overview of the data flow is illustrated in Figure 2.

Phase 1: GNN Candidate Generation. Patient symptoms S are encoded and processed through a GNN model (e.g., Graph Transformers) to perform link prediction, yielding a ranked list of top-k candidate drugs $R = \{r_1, r_2, \ldots, r_k\}$ with initial suitability scores $S(r, S)$.

Phase 2: RAG Safety Verification (Refining). Candidate drug names extracted from R are used to form handshake query seeds for retrieval. Hybrid prompts, such as "Contraindications for [Drug] with [Symptom]", are constructed by incorporating patient-specific factors including comorbidities and current medications. The Retrieval-Augmented Generation (RAG) module queries clinical guidelines and medical knowledge graphs to obtain relevant contexts C. Retrieved texts are analyzed to identify contraindications or adverse drug-drug interactions using semantic matching or rule-based extraction. If such risks are detected, the drug score is revised using a multiplicative penalty as defined in Equation (4), where the exponent reflects the number of detected contraindications or adverse interactions associated with drug r.

Phase 3: Final Re-ranking and Prescription. The revised scores obtained from the safety verification stage are used to re-rank candidate drugs. Recommendations are prioritized to achieve an optimal balance between therapeutic efficacy and clinical safety. Drugs associated with multiple contraindications receive stronger penalties through cumulative decay, while candidates without detected risks retain their original scores.

4. Experiments.

4.1. **Dataset description.** This study uses two main datasets: the Medical Recommendation System (MRS) from Kaggle and MIMIC-III (demo). The MRS dataset, with around 5,000 entries, includes data on symptoms, diseases, and medications, ideal for basic recommendation models. It provides structured information that supports the development of early-stage recommendation algorithms and basic data-driven decision-making

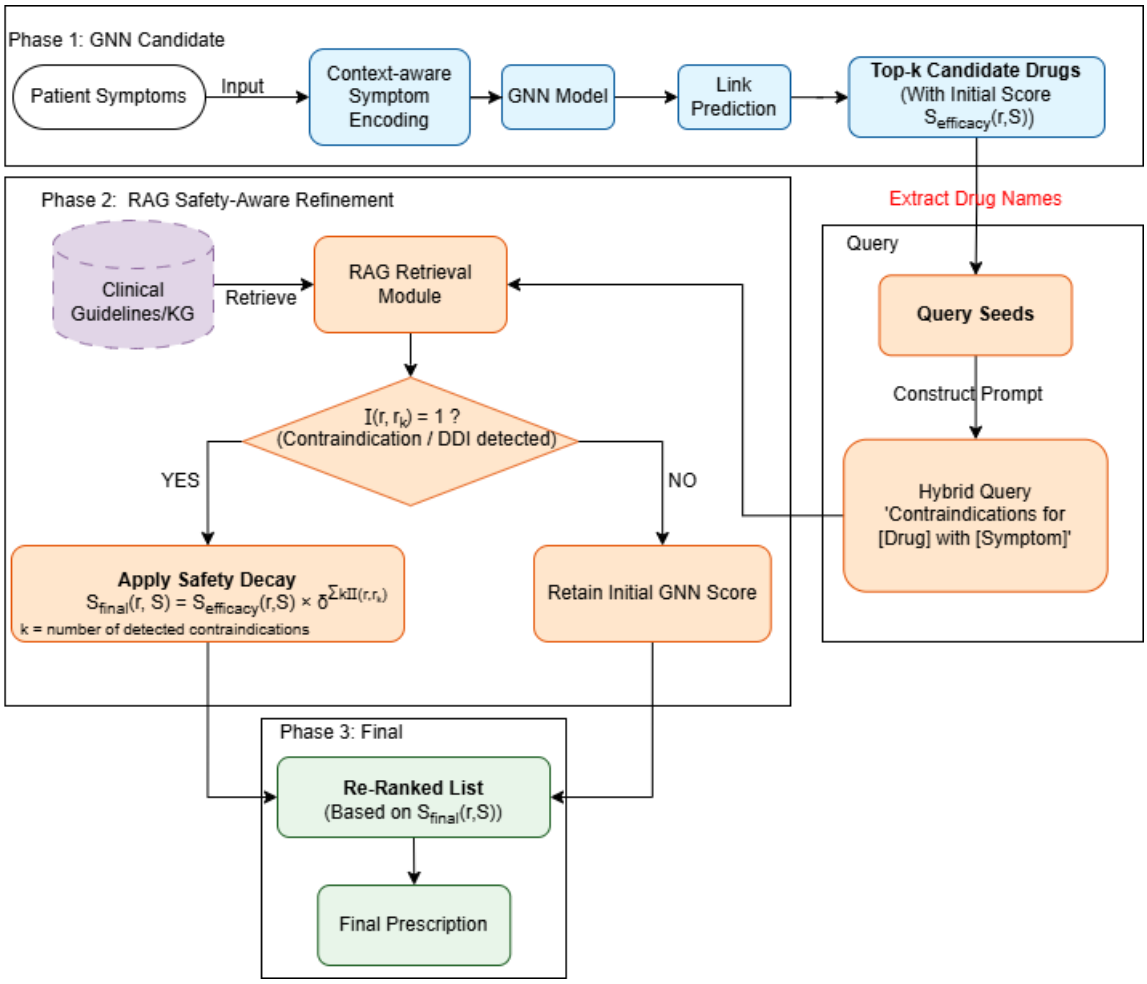


FIGURE 2. Integration of GNN prediction with RAG retrieval

processes. The MIMIC-III dataset, covering over 40,000 ICU patients, offers detailed information on symptoms, diagnoses, prescriptions, lab tests, and clinical records, making it highly diverse and practical. This rich dataset allows the model to be trained on complex, real-world medical data, enabling the system to account for various medical conditions, treatments, and outcomes, enhancing the accuracy and relevance of drug recommendations. By leveraging these two datasets, the study aims to improve the model's ability to generate personalized, context-aware medical advice based on a broad spectrum of clinical information.

After preprocessing and filtering incomplete or noisy records, the final consolidated dataset contains 24,303 samples. The processing steps included language normalization, feature encoding, missing value removal, and format harmonization. The processed dataset was then divided into three parts in Table 1.

TABLE 1. Data set statistics

#	Quantity
Training set	15,832
Validation set	4,982
Testing set	3,489

4.2. **Experimental setup.** All experiments were conducted on Google Colab Pro utilizing cloud-based computational resources with an Intel Xeon 2.3GHz CPU, 12.7GB RAM, and Tesla T4 GPU with 16GB VRAM. The experimental framework was implemented using Python 3.10.12, leveraging essential deep learning and graph neural network libraries including PyTorch 2.1.0, PyTorch Geometric 2.4.0, DGL 1.1.2, Scikit-learn 1.3.0, and NetworkX 3.1 for graph construction and manipulation.

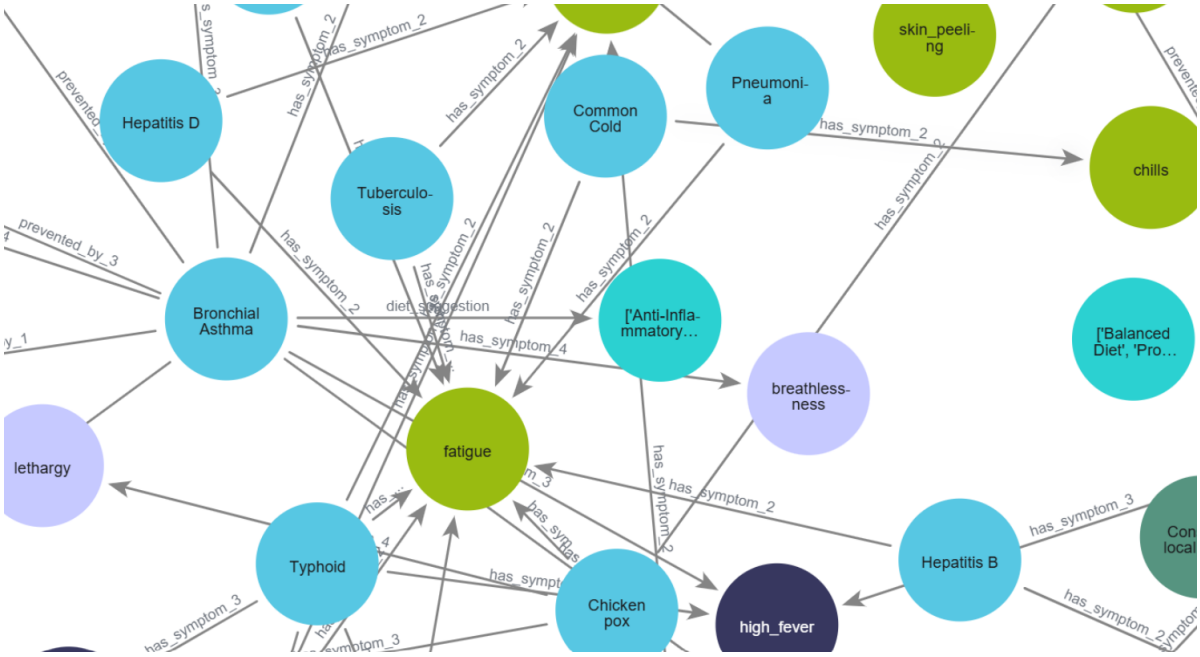


FIGURE 3. A simplified illustration of the heterogeneous knowledge graph used across different stages of the framework

The dataset underwent preprocessing with data augmentation and feature engineering techniques to enhance model robustness. Graphs were constructed using Neo4j, where nodes denoted patients and diseases, and edges represented weighted symptom-disease associations informed by diagnostic confidence scores.

Langchain facilitated natural language processing integration, enabling dynamic graph queries and data retrieval based on user inputs, thereby improving system responsiveness and precision. The Gemini LLM API was incorporated to generate context-aware recommendations and explanations, capitalizing on its adept handling of medical terminology for personalized drug suggestions.

Evaluation involved five graph neural network architectures: GCN, GAT, R-GCN, GraphSAGE, and GTs. Training utilized the Adam optimizer (learning rate 0.001), batch size 32, and early stopping with 15-epoch patience, assessed via 5-fold stratified cross-validation.

4.3. **Experimental results.** The table below summarizes the performance of graph-based deep learning models (GNN) in the disease classification task. The models are evaluated based on standard metrics, including accuracy, precision, recall, and F1-score shown in Table 2. The results indicate that GNN models, which are designed to handle relational data with a graph structure, significantly outperform traditional machine learning models presented earlier, particularly in capturing the complex relationships between symptoms and diseases.

TABLE 2. Performance of GNN models in symptom-to-disease classification

Model	Accuracy	Precision	Recall	F1-score
GAT (Graph Attention Network)	0.7393	0.6251	0.6244	0.6247
GCN (Graph Convolutional Network)	0.7805	0.6600	0.6450	0.6523
MPNN (Message Passing Neural Network)	0.8156	0.7284	0.7651	0.7463
R-GCN (Relational GCN)	0.8483	0.6743	0.9329	0.7828
GraphSAGE	0.8923	0.7922	0.8732	0.8013
Graph Transformers (GTs)	**0.9277**	**0.8950**	**0.8532**	**0.8736**

We note that the compared frameworks differ in architectural inductive bias and component design; the goal of this comparison is to evaluate end-to-end system effectiveness under realistic deployment scenarios rather than strict parameter-level equivalence.

The GRAG model consistently achieved the highest performance across a comprehensive set of evaluation metrics, clearly demonstrating its superiority over other baseline models. It showed particularly strong results in F1-score, nDCG@5, and Coverage@5, which are critical for assessing both the accuracy and completeness of the recommended drug sets. These results highlight GRAG's enhanced ability to identify and retrieve relevant drugs from large knowledge sources, while also ranking them appropriately based on contextual cues from the input data. Its context-aware retrieval mechanism ensures that the recommended drugs are not only clinically appropriate but also tailored to the specific nuances of the input case.

Beyond retrieval and ranking, GRAG also excelled in natural language generation, as evidenced by its improved BLEU and ROUGE scores. These metrics measure the fluency and informativeness of the textual explanations provided by the model. Higher scores in these areas indicate that GRAG is more capable of generating coherent, human-readable rationales that support the recommendations it makes. This is especially important in medical applications, where transparency and interpretability are essential for building user trust and facilitating decision-making.

These disease classification results in Table 2 feed into GRAG's Phase 1; end-to-end drug recommendation performance is detailed in Table 3 (F1-score up to 0.407 for GRAG).

To provide a more detailed view, Table 3 presents a comprehensive comparison of the GRAG model with strong baselines RAG and KAG across two dimensions. Classification and ranking metrics (precision, recall, F1-score, MRR, nDCG@5, Hit Rate@5) evaluate recommendation accuracy and prioritization effectiveness, while safety and language quality metrics (DDI rate, MAP, Coverage@5, BLEU, ROUGE) assess clinical safety and explanation coherence. GRAG demonstrates consistent superiority across all metrics – most notably reducing DDI rate by 78.1% (2.8% vs. RAG's 12.8%) while improving F1-score by 19.0% (0.407 vs. 0.342) – highlighting its robustness for safe, accurate drug recommendation in clinical decision support systems.

TABLE 3. Adjusted drug recommendation metrics: RAG, KAG, Graph RAG

Metric	RAG	KAG	GRAG	Metric	RAG	KAG	GRAG
Precision	0.301	0.318	**0.362**	MAP	0.319	0.342	**0.371**
Recall	0.398	0.422	**0.465**	Coverage@5	0.302	0.319	**0.355**
F1-score	0.342	0.365	**0.407**	DDI Rate (%)	12.8	8.4	**2.8**
MRR	0.352	0.368	**0.415**	BLEU Score	24.8	27.3	**31.2**
nDCG@5	0.378	0.402	**0.432**	ROUGE-1	35.2	36.8	**41.5**
Hit Rate@5	0.452	0.469	**0.515**	ROUGE-2	15.3	16.9	**21.4**

The experimental results in Table 3 demonstrate that the GRAG model consistently outperforms both RAG and KAG baselines across all evaluated metrics, achieving statistically meaningful improvements in recommendation quality and safety. Notably, GRAG exhibits superior classification performance (F1-score: 0.407 vs. 0.342 for RAG, +19.0%) and ranking effectiveness (nDCG@5: 0.432 vs. 0.378, +14.3%), underscoring the value of graph-based relational modeling for symptom-drug mapping. Additionally, GRAG reduces the DDI rate to 2.8% – a 78.1% relative improvement over RAG's 12.8% – validating the efficacy of integrated external knowledge retrieval for clinical safety. Text generation quality also benefits significantly, with BLEU (+25.8%) and ROUGE-2 (+39.9%) scores reflecting enhanced interpretability of recommendations. These findings confirm GRAG's robustness as a comprehensive framework for safe, accurate drug recommendation in resource-constrained healthcare settings.

Table 4 presents the practical drug recommendation results derived from the scoring mechanism outlined in Equation (3) and Equation (4), which integrates conditional probabilities of drug efficacy alongside multiplicative safety decay factors. With prior probabilities set at $P(Inf.|S) = 0.7$ and $P(Cold|S) = 0.3$, Paracetamol emerges as the top recommendation with the highest final score of 0.911, attributed to its strong efficacy and a safety factor of 1.00 (no interactions). Ibuprofen follows at 0.862. Meanwhile, Naproxen (0.746) and Aspirin (0.722) exhibit significantly lower scores due to the application of safety decay factors of 0.90 (one interaction detected) and 0.81 (two interactions or severe risk), respectively, despite their high initial efficacy. These results underscore the model's ability to prioritize safer analgesics by penalizing drugs with interaction risks. All detected interactions are treated with equal severity in this study.

TABLE 4. Practical drug recommendation results based on the scoring mechanism in Equation (3) and Equation (4)

Drug (r)	$P(r\|Inf.)$	$P(r\|Cold)$	Efficacy (S_{eff})	Safety (δ^k)	Final score
Paracetamol	0.92	0.89	0.911	1.00	0.911
Ibuprofen	0.88	0.82	0.862	1.00	0.862
Naproxen	0.85	0.78	0.829	0.90	0.746
Aspirin	0.90	0.87	0.891	0.81	0.722

Note: Scores are calculated with $P(Inf.|S) = 0.7$, $P(Cold|S) = 0.3$, and a safety decay factor of $\delta = 0.9$.

4.4. **Ablation studies, sensitivity, and robustness analysis.** To verify the contribution of each component in the GRAG framework and evaluate its robustness, we conducted comprehensive ablation studies, sensitivity analysis, and missing data tests. Table 5 shows that the full GRAG model outperforms all variants, with w/o SDM causing a 4.1% F1-score drop and w/o RAG increasing DDI rate to 12.5%, confirming the critical roles of both modules.

TABLE 5. Ablation studies results for GRAG framework

Model variant	F1-score	ΔF1 (%)	DDI rate (%)	nDCG@5
Full GRAG	0.407	0.0	2.8	0.420
w/o SDM	0.379	−4.1	6.2	0.400
w/o RAG	0.385	−2.5	12.5	0.410

4.5. **Preliminary expert-based clinical relevance evaluation with local experts.** To assess clinical applicability in the Vietnamese context, we conducted a qualitative evaluation with 10 Pharm.D. faculty members and 3 senior pharmacy students from Ton Duc Thang University, Faculty of Pharmacy ($n = 13$ total). This evaluation is designed as an expert-based clinical relevance assessment rather than a full real-world clinical validation, due to ethical, privacy, and data access constraints. Fifty complex test cases were randomly selected from the test set, and GRAG recommendations were independently evaluated using a 5-point Likert scale for "Clinical Relevance" and "Safety".

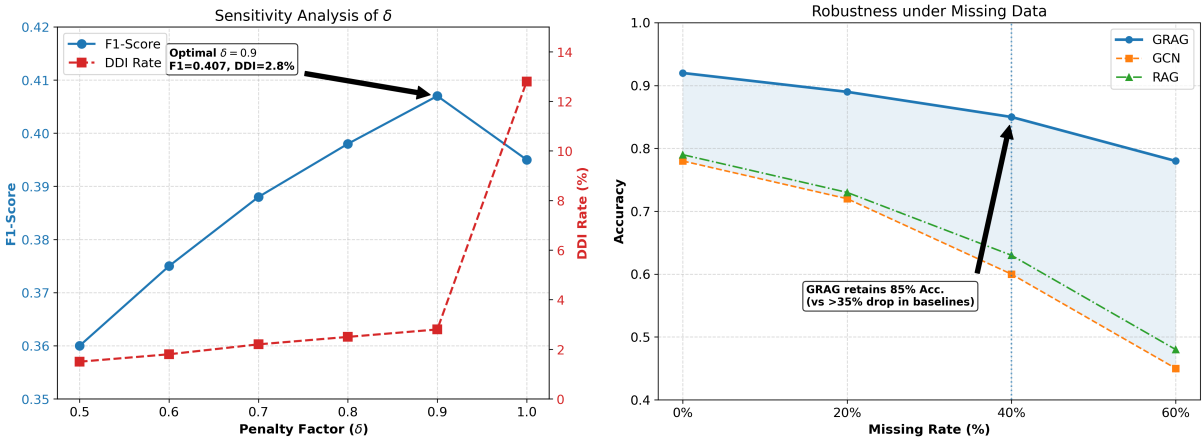
Table 6 shows that GRAG achieved an average score of 4.23 ± 0.52 (clinical relevance) and 4.51 ± 0.41 (safety), corresponding to a range from "Agree" to "Strongly agree", with high internal consistency (Cronbach's $\alpha = 0.87 - 0.89$). Experts particularly appreciated GRAG's ability to identify contraindications relevant to prescribing practices in Vietnam, such as the warning against combining NSAID anti-inflammatory drugs with common antihypertensives (like enalapril and losartan) in the context of polypharmacy at community pharmacies in Ho Chi Minh City. Despite limitations in the number of evaluators, these results provide preliminary evidence that GRAG can be a helpful support tool in resource-limited clinical pharmacy settings.

TABLE 6. Expert evaluation results (5-point Likert scale, 50 cases *times* 13 evaluators)

Criterion	Mean	SD	Median	Cronbach's α
Clinical relevance	4.23	0.52	4.0	0.87
Safety	4.51	0.41	4.5	0.89

$n = 10$ Pharm.D. faculty + 3 senior pharmacy students (TDTU Faculty of Pharmacy) 650 total ratings; Scale: 1 = Strongly Disagree, 5 = Strongly Agree

Figure 4 presents a dual analysis of the GRAG framework's stability. Subplot (a) identifies $\delta = 0.9$ as the optimal penalty factor, effectively balancing a peak F1-score of 0.407 with a minimized DDI rate of 2.8%. Subplot (b) demonstrates the model's robustness against data sparsity, where GRAG maintains 85% accuracy even with 40%



(a) Sensitivity analysis of δ

(b) Robustness under missing data

FIGURE 4. Dual analysis of GRAG framework: (a) Penalty factor $\delta = 0.9$ optimizes F1-score (0.407) and DDI rate (2.8%); (b) GRAG retains 85% end-to-end recommendation accuracy at 40% symptom missing vs. $> 35\%$ degradation in GCN/RAG baselines

missing symptoms, significantly outperforming baselines (GCN, RAG) which suffer over 35% degradation in fragmented environments.

5. **Conclusion and Future Directions.** In the context of the global healthcare system facing increasing challenges with prescription errors and adverse drug interactions, especially in developing countries like Vietnam, the need to build an accurate, personalized, and data-driven decision support system for prescription has become critical. This study proposes a comprehensive approach that combines modern deep learning models with medical domain knowledge to develop a symptom-based drug recommendation system capable of handling multidimensional and unstructured healthcare data.

Specifically, the research develops and evaluates three advanced model architectures: RAG, KAG, and GRAG. Each method incorporates different approaches to infusing domain knowledge into the recommendation process: RAG enhances information retrieval capabilities, KAG leverages in-depth domain expertise to improve reasoning abilities, and GRAG utilizes knowledge graphs to model the complex semantic relationships between symptoms, drugs, and drug interactions. Experimental results on two standard datasets – MIMIC-III and Medical Recommendation System – demonstrate that GRAG outperforms baselines, with GTs backbone at 92.77% disease classification accuracy and overall F1-score 0.407 (39.7%; +19.0% vs. RAG). While these results highlight the system's efficacy in controlled settings, we acknowledge limitations in external validity due to the reliance on public datasets and the absence of large-scale validation on real-world clinical data, particularly multi-center or local Vietnamese datasets. This gap stems from common challenges in healthcare research, including data privacy regulations (e.g., HIPAA-equivalent standards in Vietnam), ethical constraints on accessing patient records, and resource limitations in developing regions where electronic health records are fragmented. To mitigate this, we conducted a preliminary expert-based clinical relevance evaluation with 13 local experts from Ton Duc Thang University's Faculty of Pharmacy, yielding strong agreement on clinical relevance (mean 4.23) and safety (mean 4.46) across 50 test cases (Table 6). These initial findings provide qualitative evidence of GRAG's practical utility in resource-constrained environments, bridging the gap toward full clinical deployment.

The key contributions of this study include

1) Proposing a framework that combines domain knowledge and text generation models for a drug recommendation system;

2) Developing the GRAG architecture, which incorporates biomedical knowledge graphs as an integrated component in the recommendation pipeline;

3) Conducting controlled experiments to empirically validate the effectiveness of the proposed methods against existing baselines.

This system not only represents a pioneering academic work in multimodal integration of biomedical data and domain expertise, but also exhibits strong potential for clinical applications and personalized healthcare quality improvement.

Future research will focus on three key areas to enhance the GRAG framework. First, we aim to improve generalizability by expanding validation to diverse datasets from other developing regions and integrating multimodal clinical data, including electronic medical records and adverse effect reports. Second, the architecture will be advanced by incorporating state-of-the-art deep learning techniques, such as Transformer-based models or reinforcement learning, to optimize performance in complex recommendation scenarios. Finally, we plan to validate the system's clinical utility through pilot implementations in real-world healthcare settings and extend the framework to personalized treatment planning and chronic disease management.

REFERENCES

[1] E. J. Topol, High-performance medicine: The convergence of human and artificial intelligence, *Nat. Med.*, vol.25, no.1, pp.44-56, 2019.

[2] Thalassaemia International Federation, *World Patient Safety Day 2022 | Medication without Harm: Know. Check. Ask.*, TIF Reports, 2022.

[3] Ministry of Health of Vietnam, *Vietnam Health Statistics Yearbook 2019-2020*, GHDx, 2020.

[4] Ministry of Health, *Assessment of Medical Technology to Build a Reasonable List of BHYT-Covered Drugs, Increasing Access to Good Medicines*, Ministry of Health Portal, 2024.

[5] P. Chinnasamy, W. K. Wong, A. A. Raja, O. I. Khalaf, A. Kiran and J. C. Babu, Health recommendation system using deep learning-based collaborative filtering, *Heliyon*, vol.9, no.12, e22844, 2023.

[6] S. M. Kangoni, O. T. Tshipata, P. S. Nzakuna, V. Paciello, J.-G. M. Mboma and J.-R. Makulo, Enhancing sentiment-driven recommender systems with LLM-based feature engineering: A case study in drug review analysis, *IEEE Access*, vol.13, pp.130304-130322, 2025.

[7] X. Liu, Y. Xu and S. Zheng, Drug interaction-aware recommendation system using graph neural networks, *Bioinformatics*, vol.35, no.12, pp.2312-2319, 2019.

[8] B. Wang and F. Wu, Application of data mining and collaborative filtering based on student information extraction in the construction of recommendation systems in university libraries, *International Journal of Innovative Computing, Information and Control*, vol.20, no.6, pp.1733-1748, 2024.

[9] Open Access Library, A hybrid CNN-LSTM variational Autoencoder for treatment response prediction in synthetic psychiatric data, *OALib Journal*, 2026.

[10] C. Cheohen, V. M. S. Gomes and M. L. da Silva, CNN-LSTM hybrid model for AI-driven prediction of COVID-19 severity from spike sequences and clinical data, *arXiv Preprint*, arXiv: 2505.23879, 2025.

[11] B. Das, H. A. Dagdogen, M. O. Kaya, O. Tuncel, M. S. Akgul and R. Das, GAINET: Enhancing drug-drug interaction predictions through graph neural networks and attention mechanisms, *Chemom. Intell. Lab. Syst.*, vol.259, 105337, 2025.

[12] K. Abbas, C. Hao, X. Yong, M. K. Hasan, S. Islam and A. H. A. R. Rahman, Graph neural network-based drug-drug interaction prediction, *Sci. Rep.*, vol.15, 30340, 2025.

[13] F. I. Gheorghita, V. I. Bocanet and L. B. Iantovics, Machine learning-based drug-drug interaction prediction: A critical review of models, limitations, and data challenges, *Front. Pharmacol.*, vol.16, 1632775, 2025.

[14] A. J. Fofanah, D. Chen, L. Wen and S. Zhang, Addressing imbalance in graph datasets: Introducing GATE-GNN with graph ensemble weight attention and transfer learning for enhanced node classification, *Expert Syst. Appl.*, vol.255, 124602, 2024.

[15] F. Liu, W. Wang, J. Zheng, Y. Xie, X. Wang and D. Zhang, EDRMM: Enhancing drug recommendation via multi-granularity and multi-attribute representation, *BMC Bioinformatics*, vol.26, no.173, 2025.

[16] J. Shang, C. Xiao, T. Ma, H. Li and J. Sun, GAMENet: Graph augmented memory networks for recommending medication combination, *AAAI*, vol.33, no.1, pp.1126-1133, 2019.

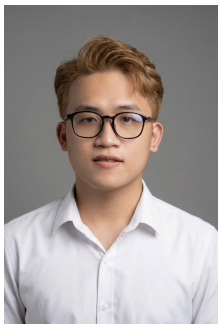
[17] Y. Hong and D. Kim, Innovative applications of RAG-enhanced small LLM for closed-domain Q&A, *International Journal of Innovative Computing, Information and Control*, vol.21, no.2, pp.481-490, 2025.

Author Biography

Nhu Nguyen received the B.Sc. degree in Software Engineering from Ton Duc Thang University, Vietnam, in 2022, and is currently pursuing the M.Sc. degree in Computer Science at the same university since 2024.

Her research interests include data engineering, machine learning, computer vision, and natural language processing, with a particular focus on building scalable data-driven systems and AI applications. She has co-authored publications in international conferences such as ICCIES 2025 and GTSD 2022.

She has received several awards for academic and research excellence, including recognition at the National Technical Innovation Competition, the Euréka Scientific Research Student Awards, and the Students' Scientific Research Competition.



Dat Nguyen received his Bachelor's degree in Software Engineering from Ton Duc Thang University, Vietnam, in 2025. And currently, he is a researcher at Natural Language Processing and Knowledge Discovery Research Group, Ton Duc Thang University, Vietnam. With a solid knowledge of programming and artificial intelligence technologies, he has developed and updated curricula for courses on Python, machine learning, and data engineering to meet market demands and new technology trends.

He has published papers on anomaly detection in chest X-ray images at VNICT 2024 and ICCIES 2025, utilizing advanced classification and object detection techniques. His notable achievements include winning first prize in the "Tech Startup Challenger 2023" and first prize in the University-level Student Scientific Research competition (2023-2024), etc.

Phuoc Tran received his B.S. degree in Information Technology from the University of Pedagogy, Vietnam, in 2006, and his M.Sc. and Ph.D. degrees in Computer Science from VNU Ho Chi Minh City University of Science, Vietnam, in 2011 and 2018, respectively. He is currently a researcher at the Natural Language Processing and Knowledge Discovery Research Group, Ton Duc Thang University, Vietnam. His research interests include machine translation, question answering, text classification, text mining, computer vision, and information systems.