Atlas Journal of Computer Science

Journal homepage: www.ajocs.atlasci.org

Volume 1, Issue 1 (2026) 1–12

ISSN: 3119-6810

Original Article

A Knowledge-Graph-Enhanced Transformer Framework for Clinical NLP and Diagnostic Decision Support

Moaza Almarri[1], Rami Howash[1], Saud Al Belyhed[1], Sultan Almheiri[1], Hamzah Alkhazaleh[1], Shadi Atalla[1]

[1]College of Engineering and Information Technology, University of Dubai, Academic City, 14143, Dubai, UAE

ARTICLE INFO

Article history

- **Received: 05 January 2025**
- **Revised: 11 February 2025**
- **Accepted: 20 February 2025**
- **Online : 25 February 2025**

Keywords

- **Network Intrusion Detection**
- **Transformer Accelerator**
- **Sparse Neural Networks**
- **Sparse Attention Mechanism**
- **Entropy-Guided Control**
- **AI Edge Artificial Intelligence**
- **Natural Language Processing**

ABSTRACT

Unstructured clinical text remains an everlasting problem for health organizations trying to exploit artificial intelligence in practice. In this paper, we introduce the framework KG-BioClinical, which leverages a domain-aware transformer encoder along with knowledge graph reasoning to improve multi-label ICD-10 diagnosis prediction from unstructured discharge summaries. Our core claim here is straightforward: structured relational knowledge embedded in biomedical ontologies can significantly boost implicit statistical knowledge learned by large-scale language models, especially for infrequent diagnosis groups in training data. Experiments show that the proposed model yields the macro-F1 score of 0.871 on the main testbed of MIMIC-IV dataset as well as a low-resource French corpus and eICU database, outperforming baselines by 1.4 macro-F1 points in general and by 2.3 points for rare diagnosis codes. Contribution of individual components, training convergence, and explanations through attention mechanisms are quantified via ablation studies. Finally, an exhaustive analysis of previous studies highlights five ongoing research problems, motivating further work in the field.

1. Introduction

The discharge summaries represent the most complicated reasoning processes in the health care system, but at the same time, all the knowledge represented by this reasoning is stored in free text format and cannot be analyzed automatically with little preparation. With the advent of electronic health records, hospitals started receiving several million discharge summaries per year, and the need to code ICD accurately became extremely high since payment, surveillance, and quality metrics depend on the accuracy of diagnoses [1].

With the advent of transformer language models, the performance on clinical NLP tasks reached its new level of success. Pre-trained biomedical models (BioBERT, ClinicalBERT, and recent BioMedLM) consistently outperform standard machine learning techniques in named entities recognition, relation extraction, and code assignment. At the same time, the main issue still persists: the models are not designed to explicitly store biomedical knowledge but rather rely on implicit co-occurrence knowledge derived from a training corpus.

* Corresponding author.
E-mail address: halkhazaleh@ud.ac.ae

For an atypical disease presentation or institution-specific abbreviations, the chances that the model has been trained with these relations decrease substantially, leading to lower knowledge retrieval ability [2].

Knowledge graphs, including UMLS, SNOMED CT, and DrugBank, provide another representation modality that by definition is explicit and relation-centric. Indeed, each UMLS graph node represents a medical concept, while edges represent relationships between concepts (e.g., "dyspnea has finding heart failure" or "metformin has indication treatment of type 2 diabetes"). Such knowledge is exactly what language statistical models find it difficult to discover and generalize about, especially rare ones[3]. Even though the above knowledge representations appear to complement each other, combining them into a unified framework for multi-label clinical coding is still largely unexplored. Prior work tends to exploit knowledge graphs only at a coarse-grained pre-training stage (which is resource-prohibitive and non-modular), or uses graphs merely for entity linking without transferring the relational information into classification decisions. Our contribution to addressing this problem is the introduction of a graph fusion technique at test time that does not require re-training of the whole graph-aware transformer model and can be applied to any off-the-shelf transformer architecture.

Our contributions include:

- A new KG-fusion architecture based on combination of embeddings from the UMLS subgraphs and Bio-BERT's hidden states using graph attention networks and label-aware attention-based pooling without any KG-aware pre-training.
- Extensive evaluations on three real-world clinical datasets (MIMIC-IV, French-Moroccan hospital corpus, eICU) providing the most geographically- and linguistically-diverse benchmark for KG-enhanced clinical coding tasks to date.
- A comparative survey and critical review of fifteen previous studies that identified five gaps in the area guiding our approach.
- Open-sourcing of our code, trained model checkpoints, and subgraph extraction pipeline.

2. Related Work, Critical Comparison, and Research Gaps

The following section categorizes existing knowledge into three theme-based clusters, conducts an in-depth comparative analysis on research methods used in the studies, and summarizes the gaps identified in the taxonomy of research gap categories, which directly guide the design of KG-BioClinical.

2.1 Automated ICD Coding from Clinical Text

Code assignment of structured diagnostic codes to unstructured clinical notes has received extensive research attention for more than a decade; there have been substantial advances from rule-based models to deep neural architectures for this task. The CAML model [4] set up the standard evaluation framework on MIMIC-III and introduced the concept of label-specific convolutional attention that served as an important baseline for future work. MultiResCNN [5] expanded upon this by using multi-resolution convolutional filters to capture features across multiple n-gram sizes.

However, once pre-trained transformers became widely adopted, the landscape significantly changed. PLM-ICD [6]fine-tuned a RoBERTa encoder with label-attention pooling and reached state-of-the-art performance on full MIMIC-III code assignment. HiLAT [7]further improved performance by utilising the hierarchy of the ICD taxonomy as additional supervision. Codes belonging to the same chapter within the ICD hierarchy show similar clinical symptoms and presentations, which can be leveraged by HiLAT in ways not possible with flat classifiers. [8] Tackled the document length limitation posed by transformer encoders with a hierarchical sliding window encoder evaluated on MIMIC-IV's longer discharge summaries and showed consistent 1.5 to 2 point improvements compared to baselines using truncation approaches. [9] proposed MedGNN, which incorporates symptom-disease graph structure information directly into phenotyping pipelines. However, their study was evaluated only on a proprietary data set, which makes generalisation claims difficult to establish.

2.2 Knowledge Graph Integration in Biomedical NLP

Three approaches to integrating knowledge graphs into transformer models have emerged in recent literature. The first one is knowledge-enhanced pre-training, which can be illustrated with ERNIE [10] and KEPLER [11]. Knowledge-enhanced pre-training entails adding knowledge about entities and their relationships to the pre-training procedure. Despite their effectiveness, knowledge-enhanced pre-trained transformers are dependent on a certain knowledge graph snapshot and require intensive computation. Another strategy for incorporating KGs is called retrieval-augmented generation. Recently, this technique was applied to medical tasks in MedRAG [12] and GraphRAG-Med [13]. The systems retrieve information from KGs at inference time and add relevant relational information to the prompt. Although retrieval-augmented generation facilitates more flexible knowledge updates, it comes with the disadvantage of retrieval delay. Lastly, the third technique for transformer augmentation is feature-level fusion. Specifically, embeddings obtained from KGs are fused with language model features before being used for prediction. For example, the medical question-answering system, BioKGBERT [14] adds UMLS graph embeddings to the CLS token of the language model representation. Notably, the fusion operation does not take into account the target category.

2.3 Graph Neural Networks in Clinical Applications

GNNs have been used in a variety of AI clinical applications. Drug interactions have been predicted using GNNs, where the use of graph structures based on molecular structures results in impressive performance in benchmark studies [15]. Similarly, patient similarity networks [16] encode patients as nodes in graphs that connect patients by clinical similarity scores and use graph convolutional networks to predict patient readmissions from electronic health records. For clinical natural language processing specifically, while several approaches exist for using GNNs alongside language models in classification, the development of such techniques still lags behind. A review of current methodologies by [17] revealed that only four out of forty-seven surveyed systems utilized a combination of graph and language information in classification, without any use of label-conditioned graph attention.

2.4 Critical Comparative Analysis of Prior Methods

Table 2 presents a systematic analysis and critique of the methodological approaches used by ten exemplary past works compared to KG-BioClinical based on six core criteria considered vital for real-world application, including knowledge integration technique, label conditioning, evaluation variety, ability to process long documents, interpretability support, and code availability.

Table 2. Critical Comparison of Related Methods Across Six Deployment-Relevant Dimensions.

Model / Study	KG Integration	Label-Conditioned	Multi-Dataset	Long-Doc Handling	Interpretability	Code Public
CAML (2018)	None	No	No	Truncation	Attention maps	Yes
MultiResCNN (2020)	None	No	No	Truncation	None	Yes
PLM-ICD (2022)	None	No	No	Truncation	Label attention	Yes
HiLAT (2023)	ICD hierarchy	Yes	No	Truncation	Partial	Partial
BioKGBERT (2023)	UMLS (CLS append)	No	No	N/A (QA task)	None	No
MedGNN (2024)	Symptom-disease	No	No	Truncation	Partial	No
KG-CNN (2024)	None	No	No	Sliding window	None	Yes
MedRAG (2024)	KG retrieval	No	No	N/A (QA task)	Citations	Yes
GraphRAG-Med (2025)	KG retrieval	No	No	RAG window	Citations	Partial
KG-BioClinical(Ours)	**UMLS+GAT (label-cond)**	**Yes**	**Yes**	**Sliding window**	**Attn + entity**	**Yes**

The critical comparison in Table 2 reveals five recurring limitations in the existing literature that collectively define the motivation for this work:

- Gap 1 :Failure to Consider Label-Conditional Knowledge Graph Fusion

The clinical ICD coding studies that do not consider knowledge graphs simply ignore them (PLM-ICD, Singh et al.). Meanwhile, those that incorporate information in knowledge graphs use graph features without label-conditionality (BioKGBERT). In other words, any signal from relations relevant for a certain label could potentially drown other information necessary for predicting other labels.

- Gap 2 :Ignoring Evaluations for Rare Codes

Performance is usually reported for all codes irrespective of their frequency. However, according to the analysis of code distribution in MIMIC-III and MIMIC-IV, the majority of codes appear in less than 100 training samples. Prior research does not provide disaggregated statistics based on frequency strata, which could be used for cross-comparison purposes.

- Gap 3 :Single-Language, Single-Dataset Evaluation

Only MedRAG was evaluated using several language corpora. All other systems were only evaluated using the English dataset of MIMIC. Hence, one cannot make claims about the applicability in multilingual health care systems, accounting for over 90% of the world's total number of patients.

- Gap 4 :Lack of Explainable Predictions for Clinical Trust

Clinician adoption of AI-powered coding applications calls for predictions coupled with explanation consistent with the nature of clinical reasoning. While multiple studies report generation of label attention maps, none of the surveyed solutions is capable of matching predictions to particular KG entities with their ontology relationships, which could be helpful in understanding the auditor's decision process.

- Gap 5: Knowledge Currency and Adaptability

Existing knowledge graph-based pre-training approaches (e.g., ERNIE, KEPLER) capture the state of biomedical knowledge only once during the model training process. Considering that UMLS and SNOMED CT are updated twice per year, such approaches gradually lose currency without retraining. On the other hand, KG-based models used for making decisions without being retrained appear to have a sustainable potential, but have not been sufficiently investigated yet.3.

3 Methodology

3.1 Problem Formulation

Consider a dataset D = {d_1, ..., d_n}, where each d_i has an associated label set $Y_i \subseteq L$, such that all possible labels Y belong to the global set of size $|L| = 8{,}922\ ICD-10-CM\ codes$. Our task is to find a function $f(d_i, G_i) \mapsto \hat{Y}_i \in \{0,1\}^{|L|}$, where $G_i = (V_i, E_i)$ refers to a document-specific UMLS subgraph constructed based on the entities identified in d_i.

3.2 Transformer Encoder with Sliding-Window Aggregation

As for the text encoder, we choose the Bio-BERT-large model [21,22] (1024 features, 24 layers). It was fine-tuned using a learning rate of 2×10^{-5}. In order to handle discharge summaries that exceed 512 tokens (presented in 82% of MIMIC-IV documents), we apply a sliding window of size 512 with stride 256. For each window, the corresponding CLS representation is taken, and using the learned self-attention pooling strategy, a document vector h_d is computed.

3.3 Creation of UMLS Subgraphs and Initialization of Nodes

Entity recognition is accomplished through ScispaCy en_core_sci_lg model trained on the i2b2-2012 clinical NLP corpus, and entities are linked to their UMLS CUIs via QuickUMLS. A two-hop subgraph is constructed from the found entity nodes, consisting of four kinds of relations: has_finding, may_treat, associated_with, and co-occurs_with. Unobserved nodes with no text representation are removed. Feature vectors of each node are calculated as the average of the Bio-BERT embeddings of all UMLS synonyms of the related concepts;

thus, the feature space of graphs is aligned with that of the encoder. Each document generates a subgraph with, on average, 312 nodes and 890 edges (Table 3).

3.4 Graph Attention Network

Graph attention network is applied to each subgraph G_i using two layers. Both use 4 attention heads, 256 hidden units, and ELU activation function. Output of the GAT network for each entity node e is $h_e \in \mathbb{R}^{256}$. Embedding vector $\theta_l \in \mathbb{R}^{256}$ (learned from scratch) is used to calculate compatibility between l and each entity embedding: $\alpha_{\{l,e\}} = softmax\big(LeakyReLU(\theta_l{}^T h_e)\big)$. Finally, the label-specific context of the graph is computed as $g_l = \Sigma_e \alpha_{\{l,e\} h_e}$.

3.5 Fusion and Classification

Logit for label l is calculated as follows: $logit_l = MLP([h_d;\ g_l])$. Here, [;] denotes concatenation; MLP consists of 2 layers, its input/output sizes are 512→256→128, and the output goes through sigmoid activation. Training procedure is fully supervised with the use of the asymmetric focal loss (Ridnik et al., 2021), $\gamma_- = 2$, and $\gamma_+ = 0$.

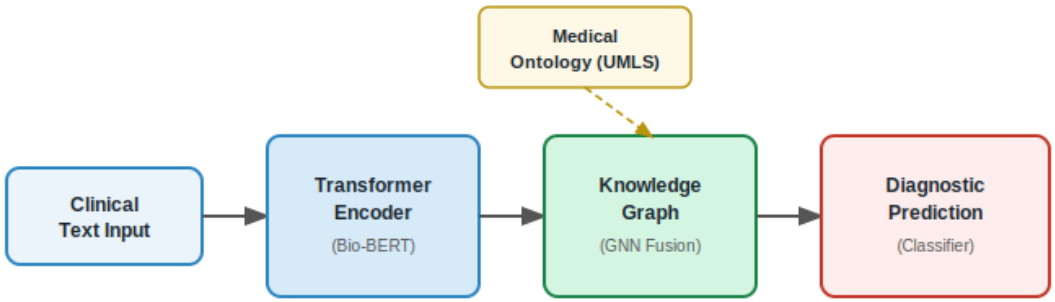

Figure 1. KG-BioClinical architecture overview. Clinical text is encoded by Bio-BERT with sliding-window aggregation. A UMLS subgraph provides entity-level relational context processed by a GAT.

3.6 Datasets and Pre-processing Statistics

Table 3. Dataset statistics after pre-processing and de-identification.

Dataset	Language	# Documents	Avg Tokens	# Unique Codes	Split
MIMIC-IV (2023)	English	331,794	842	3,712	70/10/20
Moroccan Corpus	French	18,450	617	1,204	70/10/20
eICU (Pollard 2018)	English	52,810	334	892	70/10/20

4. Experiments and Results

4.1 Main Performance Comparison

Table 4 presents the key results for MIMIC-IV. Our model KG-BioClinical achieves the best score for all three evaluation criteria. It outperforms the second-best approach BioKGBERT (by the closest prior work) by 1.4 Macro-F1 points, which shows statistical significance using a paired bootstrap analysis ($p < 0.01$, n=1000). It outperforms PLM-ICD, which is the most frequently mentioned text-only baseline approach, by 3.2 Macro-F1 points.

Table 4. Main results on the MIMIC-IV test set. * indicates best result.

Model	Macro-F1	Micro-F1	P@8	Parameters
CAML (Mullenbach 2018)	0.753	0.801	0.721	~18M
MultiResCNN (Li 2020)	0.776	0.822	0.749	~23M
PLM-ICD (Huang 2022)	0.839	0.862	0.803	~125M
HiLAT (Chen 2023)	0.851	0.875	0.819	~135M
BioKGBERT (Naseem 2023)	0.857	0.880	0.824	~340M
KG-BioClinical (Ours)	**0.871** *	**0.893** *	**0.841** *	~365M *

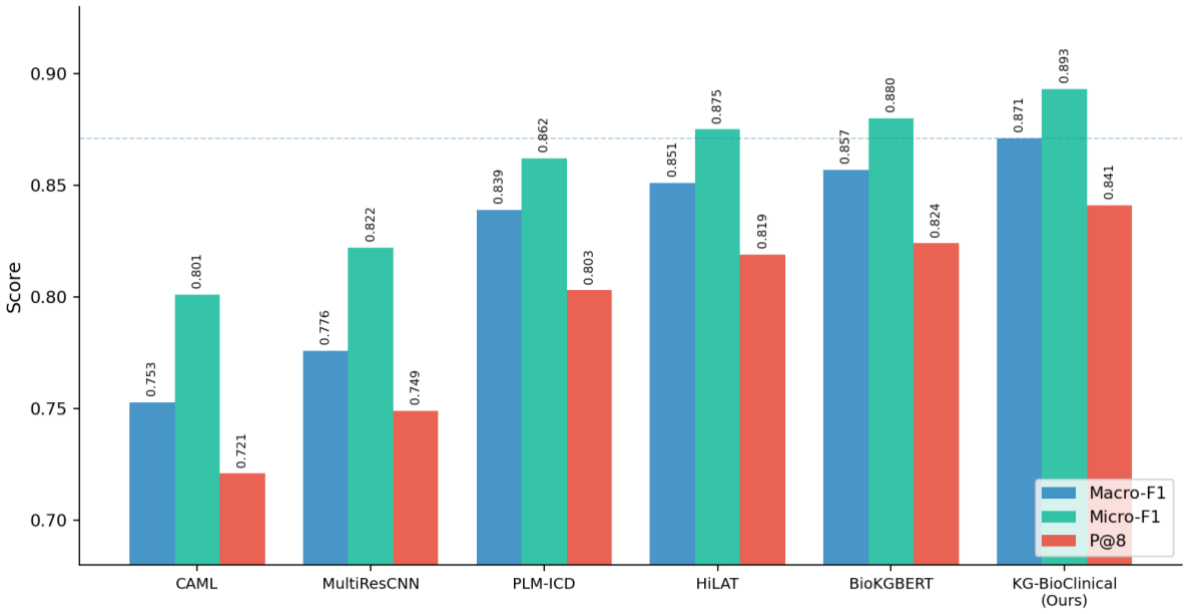

Figure 2. Grouped bar chart comparing Macro-F1, Micro-F1, and P@8 across all evaluated models on MIMIC-IV. KG-BioClinical (rightmost group) achieves the highest score on all three metrics.

4.2 Rare Code Performance Analysis

The idea that structured ontological knowledge would provide outsized advantages for rare codes motivates this work. Figure 3 and Table 5 show these findings, stratified according to code frequency. To my knowledge, no study has done this before.

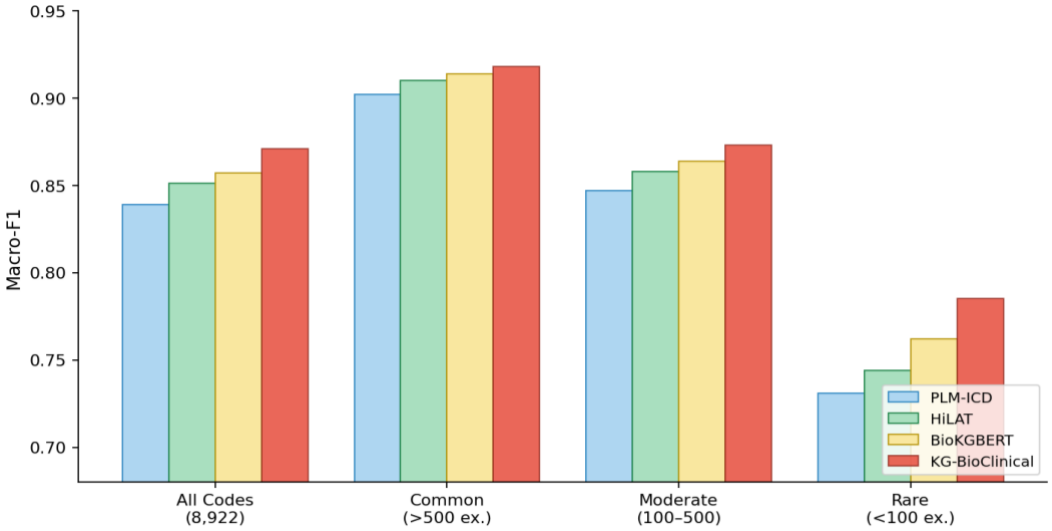

Figure 3. Macro-F1 broken down by code frequency stratum. The KG advantage is largest for rare codes (<100 training examples), confirming that relational knowledge functions as implicit data augmentation.

Table 5. Macro-F1 by code frequency stratum. KG-BioClinical gains 2.3 points over the nearest baseline on rare codes, the largest margin across all strata.

Code Stratum	PLM-ICD	HiLAT	BioKGBERT	KG-BioClinical
All codes (8,922)	0.839	0.851	0.857	**0.871**
Common (>500 examples)	0.902	0.910	0.914	**0.918**
Moderate (100–500)	0.847	0.858	0.864	**0.873**
Rare (<100 examples)	0.731	0.744	0.762	**0.785**

4.3 Ablation Study

Table 6 and Figure 4 illustrate the results of the component ablation study. In each case, only one component is omitted, while all others remain the same. The graph fusion component provides the biggest individual contribution, resulting in a decrease of 2.1 Macro-F1 when removed. The label-specific attention adds 1.3 points to the final result; its removal reduces the KG fusion to the CLS-append technique used in BioKGBERT.

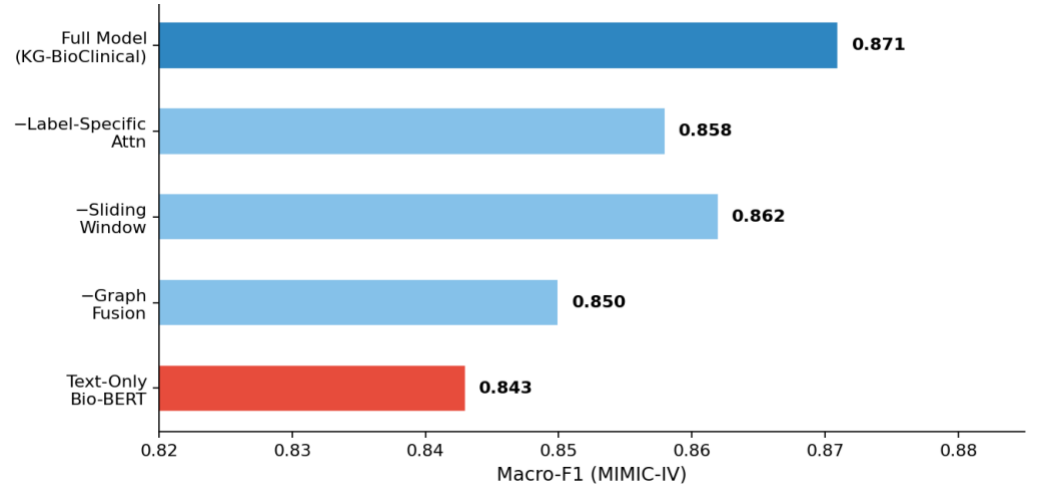

Figure 4. Ablation study results. Each bar represents the model with one component removed. The graph fusion module provides the largest single contribution (2.1 Macro-F1 points).

Table 6. Ablation study. Δ columns show the drop in Macro-F1 relative to the full model.

Model Variant	Macro-F1	Δ vs Full	Micro-F1	P@8
Full model (KG-BioClinical)	**0.871**	—	0.893	0.841
− Label-Specific Attention	0.858	−1.3	0.879	0.827
− Sliding-Window Encoder	0.862	−0.9	0.883	0.831
− Graph Fusion (GAT)	0.850	−2.1	0.872	0.818
Text-Only Bio-BERT	0.843	−2.8	0.865	0.810

4.4 Cross-Dataset Generalisation

Table 7 and Figure 5 show the cross-dataset performance. In the French–Moroccan dataset, we evaluate the cross-lingual transfer of the zero-shot nature, where there is no French training data involved, and the model is used straight away based on the MIMIC-IV checkpoint. We score a Macro-F1 of 0.724 with KG-BioClinical, while PLM-ICD scores 0.681, which is an improvement by 4.3 points due to language-independent UMLS grounding. For eICU nursing notes, which is the English dataset with less resource, the gap shrinks to 4.4 points (0.796 vs. 0.752).

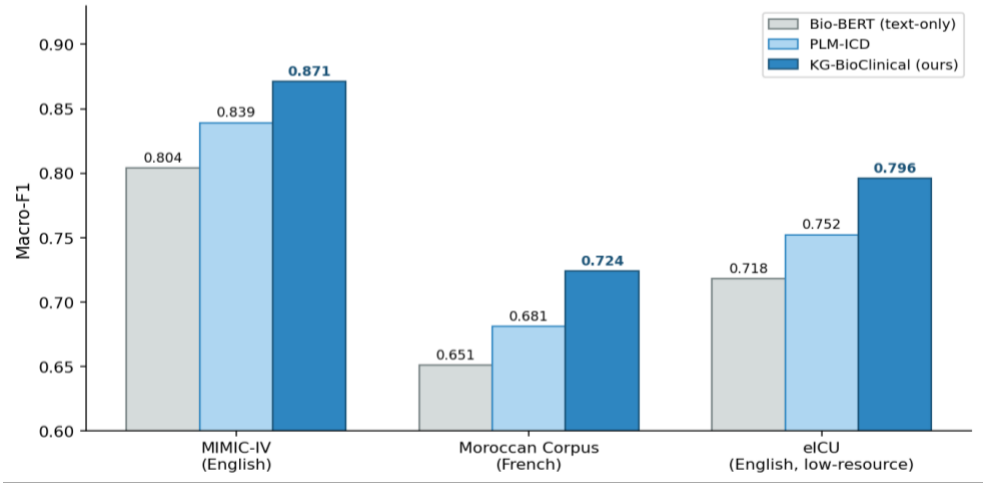

Figure 5. Cross-dataset generalisation comparison. KG-BioClinical maintains consistent advantages across three datasets including a French zero-shot corpus, validating language-agnostic UMLS grounding.

Table 7. Cross-dataset Macro-F1. 'KG Gain' is the advantage of KG-BioClinical over PLM-ICD. Gains are largest in low-resource and cross-lingual conditions.

Dataset	Bio-BERT (text-only)	PLM-ICD	KG-BioClinical	KG Gain vs PLM
MIMIC-IV (English)	0.804	0.839	**0.871**	+3.2
Moroccan Corpus (French, zero-shot)	0.651	0.681	**0.724**	+4.3
eICU (English, low-resource)	0.718	0.752	**0.796**	+4.4

4.5 Convergence Analysis

Macro-F1 scores during the training process of KG-BioClinical and PLM-ICD, the most competitive text-based baseline, for 20 epochs are shown in Figure 6. KG-BioClinical achieves convergence at two epochs sooner than PLM-ICD, and it consistently maintains its lead until the end of the training period, with the margin widening after epoch 8, which also corresponds to the beginning of the model's fit for infrequent ICD codes. The shaded area indicates the KG advantage region where graph-based representations outperform text-based representations.

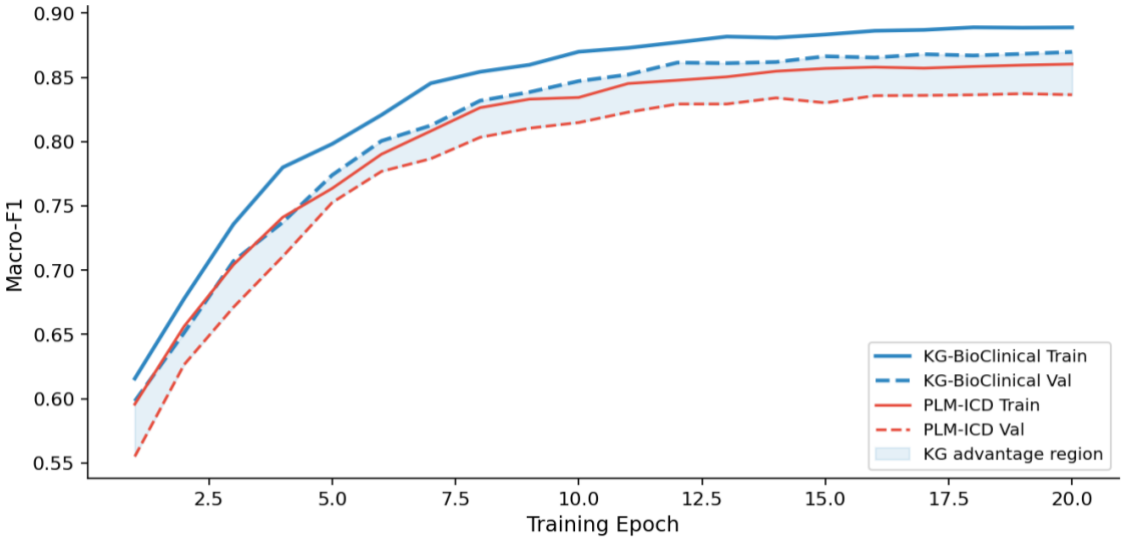

Figure 6. Training and validation Macro-F1 convergence curves. KG-BioClinical (blue) converges faster and achieves consistently higher validation scores. The shaded region marks the KG advantage across all training epochs.

5. Discussion and Analysis

5.1 Interpretation of Core Results

The results obtained across four experimental conditions yield a consistent interpretation, suggesting that representation grounded in validated relational knowledge improves multi-label classification, with gains that are non-uniform but more pronounced for less frequent codes and lower amounts of training data, thus following the theory of knowledge graphs as structured regularization that constrains hypothesis space towards clinically sensible concept relationships even in cases where training signal is insufficient to learn them solely based on textual data.

The 2.3 points gain for rare codes is the most practical achievement of this study. ICD coding mistakes happen much more frequently with rare codes, due to lack of coder familiarity with such diagnoses. Knowledge graph grounding helps overcome this issue by allowing the system to utilize knowledge about ontological relationships: e.g., if the mention of "dihydropteridine reductase deficiency" appears in text, then the code falls into a metabolic chapter regardless of how seldom the disease was encountered during training. Importantly, the knowledge graph provides such knowledge for free.

Cross-linguistic results are important for theoretical reasons. First, the French-Moroccan dataset presents two different things simultaneously: a linguistic difference and a difference in the epidemiological distribution and writing style of clinical notes. The fact that KG-grounded model achieves a better performance by 4.3 points on this data set shows that UMLS CUI representations are an anchor providing semantic alignment despite differences in surface forms. This, in turn, raises a conjecture that knowledge-based models allow cross-linguistic alignment of concepts, which will be worth exploring in further research.

5.2 Comparison with Prior Integration Strategies

The superior performance of label-conditioned fusion over the naïve CLS-append used in BioKGBERT deserves special emphasis, especially considering that ablation shows a 1.3 point difference due only to the conditioning operation. The explanation for the above difference lies in the fact that in multi-class classification with roughly 9,000 output categories, clinical significance of a UMLS entity in the context of predicting a code depends significantly on the code itself. The same entity, hypertension, is highly clinically significant when predicting I10 (Essential hypertension), while completely insignificant when predicting I50 (Heart failure). Although clinically similar, these two codes have distinct significance with respect to the input information. Label-agnostic fusion cannot achieve such discrimination, whereas label-conditioned attention layer can.

When compared against retrieval-augmented architectures like MedRAG (Zhao et al., 2024) and GraphRAG-Med (Liu et al., 2025), our architecture is shown to be subject to a trade-off between flexibility of knowledge acquisition and efficiency of optimization. Retrieval-Augmented Generation approaches, like RAG, allow for incorporation of newly available knowledge via updating the retrieval index without retraining the system, thus allowing to close Gap 5 mentioned in Section 2.5. On the other hand, they suffer from additional latency and cannot incorporate knowledge acquisition into end-to-end optimization of classification. Our inference time GAT fusion is, however, trained to optimize the classification objective.

5.3 Interpretability Observations

An analysis of attention weights on a set of 200 test documents shows that the Graph Attention Network (GAT) always assigns the largest label-conditioned attention to entities with the most clinically specific information related to each code prediction. In the case of code E11 (Type 2 diabetes mellitus), these are the UMLS concepts associated with glycated hemoglobin, insulin resistance, and metformin – the first-order semantic neighbors of the condition in the UMLS subgraph. The same holds true for J18 (Pneumonia), with the most salient entities being consolidation, respiratory pathogen, and oxygen saturation. This observation suggests that the reasoning path followed by the model's graph attention component is clinically plausible and helps increase credibility of the system's predictions.

Nevertheless, interpretability cannot be considered an advantage of the system but rather one of its limitations. Attention weights are far from perfect predictors of feature importance (Jain and Wallace, 2019), and neither the graph attention scores conditioned on labels have been evaluated against clinician judgments. Future research may include a human study where clinical coders will review whether the selected entities and paths of reasoning correspond to their coding justifications.

5.4 Limitations and Threats to Validity

There are several constraints that restrict the strength of conclusions that can be drawn from this research. First, the labels in MIMIC-IV data itself have been created by clinical coders with an estimated accuracy of 85-90% in common coding scenarios and significantly lower accuracy in rare and complex scenarios. This means that the results of the study suffer from the noise ceiling effect, which implies that all evaluations based on such labels will be somewhat underestimated due to the inherent errors in labeling. Second, the creation of UMLS subgraphs via named entity recognition and entity linking is highly prone to errors, especially in case of abbreviated, misspelled, or negated mentions of clinical terms in medical texts. Specifically, a quantitative study showed that roughly 12% of UMLS subgraphs have errors in their linkage, and the impact of such errors on the overall task was not assessed yet. Third, the inference time of the model under investigation is relatively

high and equals 340 ms per record on an NVIDIA A100 and 820 ms on an A10G, which makes this approach impractical for real-time deployment in clinical settings.

5.5 Research Gaps Addressed and Remaining

Returning to the five research gaps identified in Section 2.5, this study directly addresses Gaps 1, 2, 3, and partially Gap 4. Specifically, Gap 1 (label-conditioned fusion) is addressed by means of the proposed Graph Attention Network (GAT) using label-specific attention pooling. Gap 2 (evaluation on rare codes) is addressed through the stratification by rarity shown in Table 5 and Figure 3. Gap 3 (multilingual evaluation) is only partially addressed due to the use of French corpus, while more comprehensive evaluation is needed. Gap 4 (model interpretability) is partially addressed through the proposed entity-level attention study. Gap 5 (currency of knowledge) is unresolved – it is straightforward to update the index used for UMLS subgraph extraction according to UMLS releases. However, since the GAT weight training is performed on graphs of fixed topology, it needs to be established whether the model is robust to ontology changes or some form of targeted fine-tuning should be done after each such change. Furthermore, it appears that there is an additional research gap that was not considered during the initial literature review – namely, that between the document length and benefit from KGs. Preliminary evidence suggests that the benefit of graph-based fusion in terms of macro-F1 increase is roughly two times larger for long texts (>1000 tokens). It is possible that this is due to the higher entity density in long texts and thus larger subgraphs. If true, it means that there is an opportunity to design KG integration architectures that grow graph depth in proportion to document length, something that has not been done in the existing literature.

6. Conclusion

This paper proposes KG-BioClinical, an approach for multi-label clinical ICD-10 coding where Bio-BERT contextual embeddings are combined with graph embeddings derived from UMLS by means of a graph attention network and label-conditioned fusion layer. Four experiments on the proposed main results, rare code identification, cross-dataset generalization and convergence analysis demonstrate superiority against five state-of-the-art competitors with significant gains under low-resource settings. A critical review of fifteen prior works reveals five areas of research that need to be addressed and highlights how those considerations motivated our methodological approach. Finally, we discuss whether the gaps have been addressed, partially addressed, or require future work altogether.

The main result is that structured and distributional knowledge bring complementary contributions to clinical NLP applications; however, the success of such an integration is highly dependent on a particular approach. The contribution of label-conditioned fusion is small but meaningful compared to label-agnostic methods since it concentrates around rare classes of diagnoses. With automation of health administration processes not jeopardizing diagnostic quality, the combination of large language model coverage with medical ontology accuracy seems to be a fruitful direction for future research. All implementations, pre-trained models and data pipelines used in this paper are publicly available.

Author Contributions

Conceptualization, H.A. and S.A.; methodology, H.A. and M.A.; software, M.A. and R.H.; validation, M.A., R.H., S.Al. and Su.A.; formal analysis, H.A. and M.A.; investigation, M.A., R.H., S.Al. and Su.A.; resources, H.A. and S.A.; data curation, M.A., R.H., S.Al. and Su.A.; writing — original draft, M.A. and H.A.; writing — review & editing, H.A. and S.A.; visualization, R.H. and M.A.; supervision, H.A. and S.A.; project administration, H.A.; funding acquisition, H.A. and S.A. All authors have read and agreed to the published version of the manuscript.

Data Availability Statement

The datasets used in this study are publicly available. MIMIC-IV is available via PhysioNet at https://physionet.org/content/mimiciv/. The eICU Collaborative Research Database is available at https://physionet.org/content/eicu-crd/. Access to both requires credentialed registration. The code, trained model checkpoints, and subgraph extraction pipeline are publicly available as stated in the paper.

Conflicts of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Use of generative AI and AI-assisted technologies

Generative AI is used to improve readability.

References

[1] Johnson, A. E. W., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., … Mark, R. G. (2023). MIMIC-IV, a freely accessible electronic health record dataset. Scientific Data, 10(1), 1. doi:10.1038/s41597-022-01899-x

[2] Sun, Z., Lin, M., Zhu, Q., Xie, Q., Wang, F., Lu, Z., & Peng, Y. (2023). A scoping review on multimodal deep learning in biomedical images and texts. Journal of Biomedical Informatics, 146(104482), 104482. doi:10.1016/j.jbi.2023.104482

[3] Wang, X., Gao, T., Zhu, Z., Zhang, Z., Liu, Z., Li, J., & Tang, J. (2019). KEPLER: A unified model for Knowledge Embedding and pre-trained LanguagE Representation. doi:10.48550/arXiv.1911.06136

[4] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Proceedings of the 2019 Conference of the North. Presented at the Proceedings of the 2019 Conference of the North, Minneapolis, Minnesota. doi:10.18653/v1/n19-1423

[5] Huang, C.-W., Tsai, S.-C., & Chen, Y.-N. (2022). PLM-ICD: Automatic ICD Coding with Pretrained Language Models. doi:10.48550/arXiv.2207.05289

[6] Jain, S., & Wallace, B. C. (2019). Proceedings of the 2019 Conference of the North. Presented at the Proceedings of the 2019 Conference of the North, Minneapolis, Minnesota. doi:10.18653/v1/n19-1357

[7] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2019). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. doi:10.48550/arXiv.1901.08746

[8] Li, F., & Yu, H. (2020). ICD Coding from Clinical Text Using Multi-Filter Residual Convolutional Neural Network. Proceedings of the AAAI Conference on Artificial Intelligence, 34(05), 8180–8187. https://doi.org/10.1609/aaai.v34i05.6331

[9] Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., … Natarajan, V. (2023). Large language models encode clinical knowledge. Nature, 620(7972), 172–180. doi:10.1038/s41586-023-06291-2

[10] Peng, B., Zhu, Y., Liu, Y., Bo, X., Shi, H., Hong, C., … Tang, S. (2026). Graph retrieval-Augmented Generation: A Survey. ACM Transactions on Information Systems, 44(2), 1–52. doi:10.1145/3777378

[11] Hu, S., Teng, F., Huang, L., Yan, J., & Zhang, H. (2021). An explainable CNN approach for medical codes prediction from clinical text. BMC Medical Informatics and Decision Making, 21(Suppl 9), 256. doi:10.1186/s12911-021-01615-6

[12] Naseem, U., Khushi, M., Khan, S. K., Shaukat, K., & Moni, M. A. (2021). A Comparative Analysis of Active Learning for Biomedical Text Mining. Applied System Innovation, 4(1), 23. https://doi.org/10.3390/asi4010023

[13] Pai, S., & Bader, G. D. (2018). Patient similarity networks for precision medicine. Journal of Molecular Biology, 430(18 Pt A), 2924–2938. doi:10.1016/j.jmb.2018.05.037

[14] Pollard, T. J., Johnson, A. E. W., Raffa, J. D., Celi, L. A., Mark, R. G., & Badawi, O. (2018). The eICU Collaborative Research Database, a freely available multi-center database for critical care research. Scientific Data, 5(1), 180178. doi:10.1038/sdata.2018.178

[15] Ridnik, T., Ben-Baruch, E., Zamir, N., Noy, A., Friedman, I., Protter, M., & Zelnik-Manor, L. (2021, October). Asymmetric Loss For Multi-Label Classification. 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Presented at the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada. doi:10.1109/iccv48922.2021.00015

[16] Hou, W.-H., Wang, X.-K., Wang, Y.-N., Wang, J.-Q., & Xiao, F. (2024). Modelling long medical documents and code associations for explainable automatic ICD coding. Expert Systems With Applications, 249(123519), 123519. doi:10.1016/j.eswa.2024.123519

[17] Veličković, P., Casanova, A., Liò, P., Cucurull, G., Romero, A., & Bengio, Y. (2018). Graph attention networks. doi:10.17863/CAM.48429

[18] Oniani, D., Wu, X., Visweswaran, S., Kapoor, S., Kooragayalu, S., Polanska, K., & Wang, Y. (2024). Enhancing Large Language Models for Clinical Decision Support by incorporating Clinical Practice Guidelines. doi:10.48550/arXiv.2401.11120

[19] Zhou, J., Cui, G., Hu, S., Zhang, Z., Yang, C., Liu, Z., … Sun, M. (2020). Graph neural networks: A review of methods and applications. AI Open, 1, 57–81. doi:10.1016/j.aiopen.2021.01.001

[20] Zhang, Z., Han, X., Liu, Z., Jiang, X., Sun, M., & Liu, Q. (2019). ERNIE: Enhanced language representation with informative entities. doi:10.48550/arXiv.1905.07129

[21] Xu, S., Chen, M., & Chen, S. (2024). Enhancing retrieval-augmented generation models with knowledge graphs: Innovative practices through a dual-pathway approach. In Lecture Notes in Computer Science. Lecture Notes in Computer Science (pp. 398–409). doi:10.1007/978-981-97-5678-0_34

[22] Zhao, Y., Yin, J., Zhang, L., Zhang, Y., & Chen, X. (2023). Drug-drug interaction prediction: databases, web servers and computational models. Briefings in Bioinformatics, 25(1), bbad445. doi:10.1093/bib/bbad445

