Towards Understanding Attention-based Reasoning through Graph Structures in Medical Codes Classification

Noon Pokaratsiri Goldstein[1] **Saadullah Amin**[2] **Günter Neumann**[1]

[1]German Research Center for Artificial Intelligence (DFKI), D3.2
[2]Department of Language Science and Technology, A2.2
[1,2]Saarland Informatics Campus, Saarland University, Saarbrücken, Germany
{noon.pokaratsiri, guenter.neumann}@dfki.de, saam00002@stud.uni-saarland.de

Abstract

A common approach to automatically assigning diagnostic and procedural clinical codes to health records is to solve the task as a multi-label classification problem. Difficulties associated with this task stem from domain knowledge requirements, long document texts, large and imbalanced label space, reflecting the breadth and dependencies between medical diagnoses and procedures. Decisions in the healthcare domain also need to demonstrate sound reasoning, both when they are correct and when they are erroneous. Existing works address some of these challenges by incorporating external knowledge, which can be encoded into a graph-structured format. Incorporating graph structures on the output label space or between the input document and output label spaces have shown promising results in medical codes classification. Limited focus has been put on utilizing graph-based representation on the input document space. To partially bridge this gap, we represent clinical texts as graph-structured data through the UMLS Metathesaurus; we explore implicit graph representation through pre-trained knowledge graph embeddings and explicit domain-knowledge guided encoding of document concepts and relational information through graph neural networks. Our findings highlight the benefits of pre-trained knowledge graph embeddings in understanding model's attention-based reasoning. In contrast, transparent domain knowledge guidance in graph encoder approaches is overshadowed by performance loss. Our qualitative analysis identifies limitations that contribute to prediction errors.

1 Introduction

The codification of clinical texts by assigning the International Classification of Diseases (ICD) codes for the purpose of streamlining research, insurance billing, and other workflow standardization is a necessary task in healthcare settings. To assign an accurate and complete set of ICD codes to a clinical text, both a knowledge of institutional guidelines and understanding of medical terminology are crucial. Consequently, it is time and cost intensive. Solving the task as a multi-label classification (MLC) problem is one of the common top-performing deep learning approaches to automating this task.

In addition to challenges stemming from the extensive domain knowledge requirements, clinical notes are often over 3,000 words long; due to computational time and memory limitations, models often have to truncate these documents to a smaller size (Moons et al., 2020; Kaur et al., 2021), risking information loss that could be helpful in predictions. Many pre-trained language models such as BERT (Devlin et al., 2019) and its variants, for instance, can only take inputs up to 512 tokens.

External knowledge resources such as the UMLS Metathesaurus (Bodenreider, 2004) for medical concepts and relational information have shown promising results in named entity recognition (NER) (Liang et al., 2023) and automatic ICD coding (Yuan et al., 2022). While attention mechanism (Bahdanau et al., 2015) in combination with knowledge graphs (KG) and graph neural networks (GNN) have been shown to be beneficial when applied to relational information from the output (label) space in this task, the effects of graph representation on the input (document) space are not yet extensively studied.

We are motivated by the applications of this work in modeling other clinical tasks that can also be set up as an MLC problem, e.g. inpatient documentation from multi-modal or non-text input data[1]. It is crucial in critical and highly-regulated fields that human domain experts can understand what con-

[1]Real-time charting in electronic health records (EHR) for clinicians in some settings involves selecting corresponding options from a fixed menu with optional unstructured texts, similar to data entries in a spreadsheet.

Proceedings of TextGraphs-17: Graph-based Methods for Natural Language Processing, pages 78–92
August 15, 2024

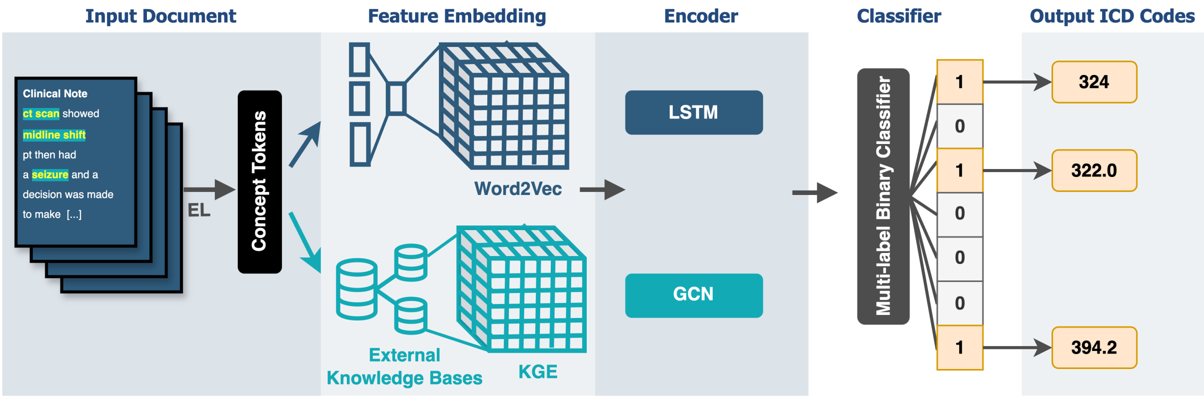

Figure 1: Overview of MLC pipeline: a) concept-based tokens are extracted to represent the input documents, b) tokens are represented by pre-trained feature embeddings (Word2Vec or KGE), c) encoding step transforms input features into latent representations (LSTM or GCN output) and d) binary classifiers determine whether the output representations belong to specific labels.

tribute to correct and incorrect predictions when incorporating automated systems' outputs in their workflow. These considerations influence our decision to investigate concept-based features and verify model's attention-based interpretability through qualitative analysis.

We investigate the impact of implicit graph structures in the form of knowledge graph embeddings (KGE) concept features representation and explicit domain-knowledge guided encoding of input document concepts and their relational information using GNN. Our contributions can be summarized as follows: 1) we highlight the benefits of domain knowledge injection through KGE over traditional contextualized embeddings in representing concept-based features and facilitating clinically intuitive attention-based reasoning, 2) we demonstrate the limitations of GNN encoding architecture, and 3) we identify challenges that contribute to attention-based reasoning errors.

2 Related Works

Knowledge Graph Embeddings: Teng et al. (2020) incorporate knowledge graph embeddings (KGE) as a supplement to text representations to simulate the human reasoning process of deriving ICD codes from a medical knowledge base and to make results more interpretable when combined with the attention mechanism. Chang et al. (2020) demonstrate that KGE are effective at leveraging relational information and representing biomedical domain knowledge; e.g. TransE (Bordes et al., 2013) and RotatE (Sun et al., 2018) are able to retain semantic group and type information inherent in the source knowledge base ontology e.g. SNOMED CT in the UMLS. Combining KG represented entities with input document representations also shows promising improvements in relation extractions (Matsubara et al., 2023). Beyond these works in the biomedical domain, to date, methods involving KGE in automatic ICD coding have been limited.

Graph Neural Networks: EHR data often contain information regarding diagnoses, lab values, encounters, and the patients organized in a graph-like structure to reflect clinical decisions process (Choi et al., 2020). These observations suggest that the features in an EHR encounter and clinical notes have structural relationships. GNN architectures are known to be effective at representing relational information, making them suitable for capturing dependencies among ICD codes and medical concepts. Choi et al. (2020) posit that Graph Convolutional Networks (GCN) represent a special case of Transformer (Vaswani et al., 2017) and propose Graph Convolutional Transformer (GCT) to structurally represent key components in an EHR document. Qiu et al. (2019), Zong and Sun (2020), and Cao et al. (2020) use GCN to model ICD code and/or concept co-occurrence to address the class imbalance problem in the output (label) space.

Attention Mechanism: To provide human-interpretable results, Mullenbach et al. (2018) and Teng et al. (2020) utilize attention mechanism (Bahdanau et al., 2015; Luong et al., 2015) to verify that relevant text spans are clinically informative. Teng et al. (2020), Vu et al. (2020), Saini et al. (2021), and Yuan et al. (2022) use the *softmax* operation to calculate label-wise attention weights from the encoder's output to create label-specific vectors representing the input document. Many high-performing models incorporate variations of the attention mechanism. In combination with domain knowledge implicitly represented through KGE, the attention mechanism helps the model focus on parts of the input document relevant to the predicted labels, resembling how a human medical coder concentrates on relevant parts of the document to determine the corresponding ICD codes based on their domain knowledge expertise. We refer to this process as attention-based reasoning in this work.

3 Methodology

When ICD coding is set up as an MLC task, as shown in Figure 1, a document D is represented as a sequence $\mathcal{X} = [x_1, x_2, x_3, ...x_n]$, where n represents the number of words or extracted concepts in $\mathcal{X}$. The classification model's learning task is to output a label vector $\mathcal{Y} = [y_1, y_2, ..., y_L]$, where L is the total number of codes from a label set L and each $y_i \in \{0, 1\}$. 1 denotes the document contains code i and 0 otherwise. A common training objective is to minimize the binary cross entropy (BCE) loss function between the predicted labels $\tilde{y}_i$ and the true labels y_i.

All experiments are conducted on the Multi-parameter Intelligent Monitoring in Intensive Care-III (**MIMIC-III**) dataset (Johnson et al., 2016). We focus on the discharge summaries and their assigned International Classification of Diseases, 9th Edition, (ICD-9) codes[2]. We follow pre-processing steps and measure results using the same evaluation metrics in Mullenbach et al. (2018) and Vu et al. (2020).

3.1 Concept Features Tokenization

Using text input as a baseline reference, we represent a document as a sequence of medical concepts. Exploiting mapping between medical terms and their textual descriptions in large ontological databases, e.g. the Unified Medical Language System (UMLS) (Bodenreider, 2004), identifying concepts in the input documents can be viewed as an entity linking (EL) task. Within the UMLS, terms across vocabularies are assigned Concept Unique Identifiers (CUIs). Additional attributes such as semantic types, relations, and hierarchical information are also available across CUIs. Since ICD codes are a subset of concepts within the UMLS, using concept (CUI) tokens also provides a way to incorporate additional external knowledge into the model.

3.1.1 Concepts Extraction

We use ScispaCy UMLS entity linking (EL) tool (Neumann et al., 2019) to extract CUIs from the original discharge summaries. We select only CUIs with at least 0.7 confidence scores. Choosing a higher score of 0.8 does not empirically improve results in our experiments (see Appendix A.4). Analogous to the pruning steps in a text pre-processing pipeline, we also prune out rare and frequent CUIs. Using analogous thresholds as in Mullenbach et al. (2018) and Vu et al. (2020), we determine the minimum and maximum frequency thresholds for CUI tokens as follows:

- **frequent**: normalized frequencies exceeding 1500 times per million tokens.
- **rare**: normalized frequencies less than 0.1 times per million tokens.

We also discard CUIs that do not belong to the semantic types of the MIMIC-III dataset ICD-9 codes as well as zero-shot CUIs.[3]

The resultant vocabulary size of the dataset is 26,485 unique CUI tokens. As seen in Table 1, the average input sequence lengths across partitions are well within the typical truncated input lengths of existing state-of-the-art models.

Version	Partition	Min	Mean	Max
Full	Train	9 (55)	696 (1,731)	4,560 (11,940)
	Validation	103 (244)	819 (2,049)	3,038 (7,247)
	Test	90 (252)	825 (2,057)	4,725 (8,209)
Top-50	Train	62 (117)	715 (1,782)	3,665 (8,387)
	Validation	102 (244)	826 (2,066)	3,036 (7,247)
	Test	108 (259)	841 (2,095)	3,061 (7,128)

Table 1: Minimum, mean, and maximum CUI and text tokens (in parentheses) per document for the Full and Top-50 MIMIC-III dataset partitions after pre-processing.

3.2 Feature Representation

Contextualized Representation: Word2Vec (W2V) embeddings for CUIs serve as a comparative baseline against KGE in our experiments due

[2]Multiple editions of ICD codes exist; for simplification, ICD and ICD-9 codes are used interchangeably in this work unless otherwise indicated.

[3]Zero-shot CUIs are defined as CUIs in the validation or test partition not seen in the train set.

EHR Feature	UMLS Semantic Group (SG) or Type (TUI)
Diagnosis	DISO - Disorders ANAT - Anatomy PHYS - Physiology PHEN - Phenomena LIVB - Living Beings
Procedure	PROC - Procedures DEVI - Devices ACTI - Activities & Behaviors
Lab Result	CHEM - Chemicals & Drugs T034 - Laboratory or Test Result T059 - Laboratory Procedure
Concept	CONC - Concepts & Ideas

Table 2: UMLS Semantic Groups (SG) and Semantic Type Information (TUI) and their corresponding EHR structural features: Diagnosis, Procedure, Lab Result, and Concept; features are identified based on our observations and findings in Choi et al. (2020).

to its usage in existing top-performing models for the text input type. The reference results with text features in Table 3 also use W2V embeddings. Using the same parameters as in Mullenbach et al. (2018) and Vu et al. (2020), we train W2V embeddings for CUI tokens with CBOW (Mikolov et al., 2013) algorithm. We use Gensim (Řehůřek and Sojka, 2010) W2V implementation.[4]

Knowledge Representation: We use TransE Bordes et al. (2013) KGE trained on pre-processed data of the UMLS 2019AB released publicly by Chang et al. (2020)[5]. Since both TransE (Bordes et al., 2013) and RotatE (Sun et al., 2018) achieve comparable results on semantic classification tasks and capture similar semantic information as investigated in Chang et al. (2020), experiments comparing performance between different types of KGE are beyond the focus of this works and are left for future works. We use DGL-KE (Zheng et al., 2020) implementation of TransE for training according to steps described in Chang et al. (2020).[6]

3.3 Encoders

Label Attention Encoder (LAAT): The LAAT model introduced by Vu et al. (2020) follows an MLC pipeline as shown in Figure 1. It consists of an embedding layer where pre-trained W2V embeddings are used to represent document input tokens. The encoder is a bidirectional Long Short Term Memory (LSTM) network whose output provides latent feature representations for the input tokens up to a specified number; this is represented as a vector $\mathbf{H}$ where $\mathbf{H} \in \mathbb{R}^{2u \times n}$. n refers to the number of input tokens and u is the LSTM hidden size. The attention layer $\mathbf{A} \in \mathbb{R}^{|L| \times n}$ transforms the feature representations $\mathbf{H}$ into label-specific vectors as shown in Eq. 1 to 3. $\mathbf{W} \in \mathbb{R}^{d_a \times 2u}$ and $\mathbf{U} \in \mathbb{R}^{|L| \times d_a}$ matrices are learnable parameters. u and d_a are tunable hyper-parameters. The output of the label-specific layer $\mathbf{V} \in \mathbb{R}^{2u \times |L|}$ is the representation of the input document where each i^{th} column in $\mathbf{V}$ corresponds to the i^{th} label in L. The last layer is a feed-forward neural network followed by a sigmoid activation function, which predicts whether a specific ICD code is assigned to the input document or not.

$$\mathbf{Z} = \tanh(\mathbf{WH}) \tag{1}$$
$$\mathbf{A} = \text{softmax}(\mathbf{UZ}) \tag{2}$$
$$\mathbf{V} = \mathbf{HA}^{\text{T}} \tag{3}$$

We re-implement the model to accommodate concept-based tokens using PyTorch (Paszke et al., 2017). We follow implementation details such as optimal hyper-parameters, learning rate, batch size, number of epochs, dropout probability, AdamW (Loshchilov and Hutter, 2018) optimization, and learning rate scheduler as implemented by Vu et al. (2020). In lieu of early stopping, we save the model with the highest validation F1_{micro} for evaluation against the test partition. See Appendix A.2.1 for implementation details. We consider this model a high-performing non-GNN baseline encoder.[7]

GNN Encoder: We use 2-layer Graph Convolution Networks (GCN) (Kipf and Welling, 2017) as a representative GNN encoder for experiments investigating GNN domain knowledge encoding. Choi et al. (2020) demonstrates the correspondence between normalized adjacency matrix calculations in GCN and the attention equation in the Transformer (Vaswani et al., 2017) architecture. Similar to how LAAT utilizes attention mechanism to focus on relevant parts of the input data (represented

[4]Other types of corpus-based embeddings have been proposed to represent concepts in the UMLS, notably Cui2Vec (Beam et al., 2020) and Med2Vec (Choi et al., 2016). However, Chang et al. (2020) observe that these approaches have limitations due to data inaccessibility, high computational requirements, and low coverage, which make their usability for downstream tasks limited.

[5]The link to the data files is published through the SNOMED CT Knowledge Graph Embeddings Git repository: https://github.com/dchang56/snomed_kge

[6]See Appendix A.2.2 for KGE training hyperparameters.

[7]Higher performing encoders have since been proposed and our study can be extrapolated to them; however, for simplicity and discussion, we designate LAAT as a strong non-GNN baseline for this task.

Encoder	Embedding	Precision		Recall		F1		AUC		P@5
		macro	micro	macro	micro	macro	micro	macro	micro	
LAAT (50)	W2V	59.11	64.90	48.92	55.03	53.53	59.56	86.07	89.41	58.06
	KGE	**64.11**	**68.46**	**54.55**	**59.02**	**58.94**	**63.39**	**88.22**	**91.14**	**60.69**
	Text	72.04	75.60	61.84	66.95	66.55	71.01	92.79	94.60	67.28
LAAT (Full)	W2V	7.26	**65.78**	4.70	35.44	5.70	46.07	84.92	97.77	73.31
	KGE	**7.86**	64.78	**5.47**	**37.80**	**6.45**	**47.74**	**86.62**	**98.05**	**74.41**
	Text	10.65	65.70	9.19	50.64	9.87	57.20	89.84	98.56	80.91
GCN_{EHR} (50)	W2V	54.81	65.04	34.75	44.15	42.53	52.60	83.96	87.02	54.49
	KGE	**58.75**	**65.24**	**41.61**	**48.41**	**48.71**	**55.58**	**84.72**	**87.72**	**56.23**
GCN_{EHR} (Full)	W2V	3.53	60.19	1.55	18.17	2.16	27.92	75.31	96.28	58.61
	KGE	**3.89**	**60.69**	**1.56**	**18.91**	**2.23**	**28.84**	**76.10**	**96.40**	**59.32**

Table 3: Results from experiments on the LAAT and GCN models with the MIMIC-III Top-50 and Full test sets comparing KGE and W2V CUI embedding types. Text input results are included as a reference as it is the input type in Vu et al. (2020). Underlined scores are highest across input types; **bold** ones are the highest within CUI input.

Version	Model	Precision		Recall		F1		AUC		P@5
		macro	micro	macro	micro	macro	micro	macro	micro	
Top-50	GCN_{BASE}	**62.12**	**67.81**	38.22	45.02	47.33	54.11	84.54	87.40	56.00
	GCN_{EHR}	58.76	65.24	**41.61**	**48.41**	**48.72**	**55.58**	**84.72**	**87.72**	**56.23**
Full	GCN_{BASE}	2.86	55.53	1.31	17.81	1.80	26.97	**77.19**	96.07	55.60
	GCN_{EHR}	**3.89**	**60.69**	**1.56**	**18.91**	**2.23**	**28.84**	76.10	**96.40**	**59.32**

Table 4: Results from GCN experiment comparing different edge connection approaches; all models use KGE node embeddings to represent CUIs.

as an output of an LSTM encoder), GCN encoder and the readout function output a graph-level representation of the input document that focuses on relevant concept nodes in the graph.

Each input document that has been processed into a sequence of CUIs is represented as a graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, where $\mathcal{V}$ and $\mathcal{E}$ are nodes and edges. Each node in $\mathcal{V}$ represents a unique CUI in a document. An edge in $\mathcal{E}$ represents a connection or relation between CUIs (nodes) as determined by different graph construction methods. To obtain a document-level representation for classification, we specify a sum pooling readout function as it has been shown to be optimal for graph classification tasks (Xu et al., 2018). A readout function can be a simple sum, mean, or max pooling function or more complex (Xu et al., 2018; Ying et al., 2018; Zheng et al., 2020); however, this is beyond the focus of this work.

3.4 Experiment Settings

Implicit Graph Structures with KGE: We compare performance between KGE and W2V embeddings on the LAAT model for the CUI-represented input and on the GCN encoder model. We investigate if KGE pre-training and the implicit relational information from the external UMLS knowledge base improve ICD-9 classification.

Explicit Graph Structures with GNN: We compare a graph edge construction method that explicitly follows clinical reasoning steps as reflected in CUIs co-occurrences against a baseline approach guided by relations in the UMLS KG. As observed in Choi et al. (2020) and our manual annotation (see Appendix A.3), there is a relationship between diagnostic information and treatments that is also reflected in EHR structural features as shown in Table 2. In this work, we refer to the process of relating treatments or procedures to diagnostic information as clinical reasoning. Since ICD codes encompass health-related phenomena (e.g. signs and symptoms, findings, complaints, social factors etc.) and treatment concepts, we investigate if the explicit relational information encoding following a domain-knowledge guided approach improves ICD-9 classification.

1. **Baseline** (GCN_{BASE}) Nodes representing CUIs in a document have edges between them if both nodes (CUIs) are related in the UMLS KG used in pre-training KGE.

2. **Domain-Knowledge Guided** (GCN_{EHR})

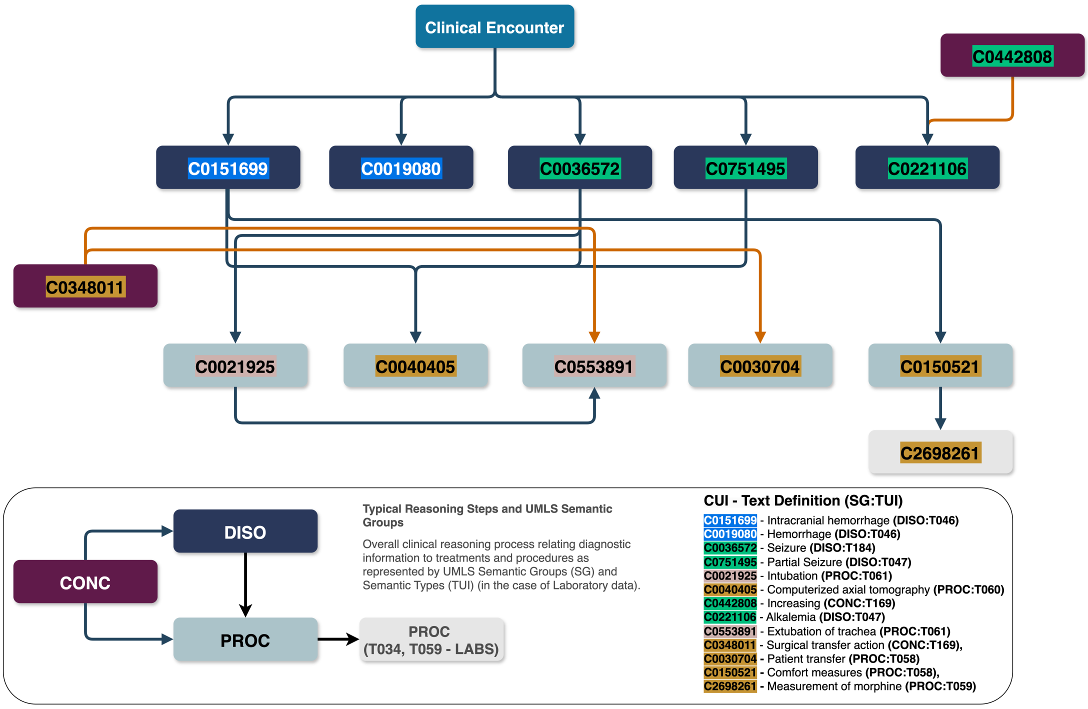

Figure 2: The clinical reasoning steps relating distantly mentioned CUIs in the manual annotation example shown in Figure 5 of Appendix A.3 are demonstrated in this flow chart. The CUIs are color-coded by their UMLS Semantic Group (SG) and are organized into EHR structural features described in Table 2. The arrows demonstrate how Diagnostic (DISO) CUIs are related to Procedural (PROC) and Lab (CHEM, T034, T059) CUIs and how Concept (CONC) CUIs are associated with both Diagnostic and Procedural CUIs.

From manual annotation of 5 randomly selected samples in the Top-50 version of the training dataset[8], we observe co-occurrences between CUIs that also follow typical clinical presentations. For instance, CUIs describing diagnoses are present along with CUIs for certain procedures. As shown in Figure 2, it is possible to group CUIs corresponding to EHR feature types such as diagnoses, procedures, concepts, and lab data based on the UMLS semantic information. While a domain expert with clinical experience can easily relate diagnostic concepts and commonly associated treatment procedures, conditional probabilities between CUIs of different semantic groups can provide a useful edge connection guidance that follows clinical reasoning as proposed in Choi et al. (2020). The steps are summarized as follows:

(a) CUIs are grouped by their UMLS Semantic Group (SG) and EHR feature type described in Table 2.

(b) Conditional probabilities of the co-occurrences of CUIs across these groups are calculated from the training partition as in Choi et al. (2020).

(c) Edges are present between CUIs if their conditional probability exceeds a specified threshold: 0.3, 0.5, 0.7, 0.8.

3.5 Attention-based Reasoning Evaluation

To evaluate the attention-based reasoning interpretability, we analyze input text and concept tokens from the Top-50 LAAT experiments. After filtering out test partition samples with no predicted labels, we randomly select 10 samples that contain predictions of the most commonly occurring labels in the test partition. We extract tokens with normalized activation weights from LAAT Attention Layer **A** (Eq. 2) of at least 0.5 of the maximum attention weight (for each predicted label) and compare them to tokens annotated by an intensive care clinician[9] as relevant. We choose 0.5 as results in Teng et al. (2020) comparing interpretability evaluation of text segments extracted from higher attention weights (0.8 threshold) show lower accuracy

[8] See Appendix A.3 for an annotated example.

[9] We use the definition of clinician as explained in Institute of Medicine (US) Committee on the Future of Primary Care (1994).

than those from lower weights; their findings suggest lower weight ranges may identify potentially informative tokens.

4 Results

Results in Table 3 demonstrate the benefits of implicit graph-representation in the form of KGE on both LAAT and GCN encoders over corpus-based CUI embeddings. KGE shows improvement over W2V CUI embeddings across all metrics on the LAAT model in the Top-50 and Full versions, with an exception of the Precision$_{micro}$ where W2V performance is higher. On GCN$_{EHR}$ model, KGE shows slightly higher performance across all metrics over W2V embeddings. Our findings support observations noted in Chang et al. (2020) and Teng et al. (2020) that KGEs improve domain knowledge representation on the input document space in leveraging relational information. However, with the exception of Precision$_{\text{micro}}$ and AUC$_{\text{micro}}$ metrics in the Full version where CUI results are comparable to text-input baseline, concept features result in lower performance than text features. For critical-domain applications, the interpretability advantage of concept-based features over text-based input type as demonstrated in Section 4.1 may justify some performance trade-offs.

Table 4 shows the impact of graph edge construction approaches on GCN performance. Across most of the metrics, a graph construction method that incorporates clinical reasoning and EHR structure offers some benefits over baseline, where edges are connected based on KG relations. An exception is observed in the Top-50 Precision, where the baseline KG-guided construction outperforms the EHR-guided approach. The more noticeable difference in the Full version can be attributed to a larger code base exceeding KG coverage, thus, contributing to a lower Recall in the GCN$_{BASE}$ approach. While GCN as a standalone encoder provides an ability to explicitly encode relational information that reflects clinical reasoning and EHR structural features in the graph construction methods, possibly improving model's interpretability by domain experts, this contribution is limited due to much lower performance across all metrics in comparison to LAAT model.

EHR Conditional Probability Threshold: Among the GCN$_{EHR}$ approaches, performance varies according to minimum co-occurrence conditional probability threshold between EHR structural feature groups. As shown in Figure 3, this variability is more noticeable in the Top-50 than in the Full version. Based on fine-tuning for the highest F1$_{micro}$ among GCN$_{EHR}$ experiments over different thresholds, the optimal minimum probabilities for the GCN$_{EHR}$ are 0.7 and 0.5 for the Top-50 and Full version respectively. GCN$_{EHR}$ results reported in Table 4 are based on these thresholds for their respective version.

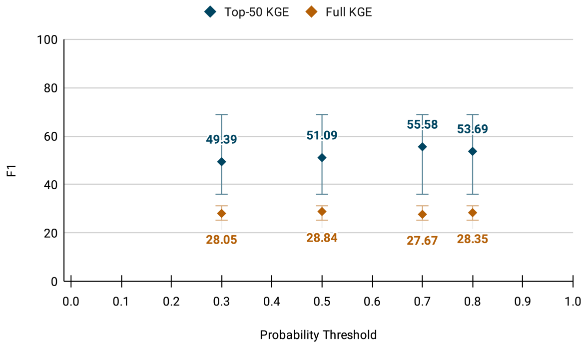

Figure 3: F1$_{micro}$ score in relation to minimum conditional probability threshold in the Top-50 & Full versions of GCN$_{EHR}$ model. Error bars indicate the standard deviation from the mean F1 scores of each group; boundaries are shown at 6 times the standard deviation for clearer visualization.

4.1 Attention-based Reasoning Interpretability

Examples in Table 5 demonstrate the impact of different input feature types on the model's attention mechanism. Clinician-annotated text and CUI tokens are shown as a reference. Our goal is to verify that label predictions are made following clinically informative attention-based reasoning. A false positive example ("401.9") is included to illustrate if erroneous predictions are avoidable, i.e. given the available information (in the form of document text or CUI tokens), would a clinician make similarly incorrect label predictions?

In the example where the ICD label, its CUI, and description match with the input CUI or their text description, KGE and W2V concept features are equally informative as in the example of "427.31:Atrial Fibrillation". Both concept embedding types are more precise than the highlighted text tokens ("fib" and "fibrillation"), possibly due to exact CUI matching. Dropping "a" from "a fib" suggests that the attention mechanism may potentially associate the same text token for both "a fib", "v fib" (ventricular fibrillation), or other terms that are partially similar in the text-input model.

ICD-9:Description (CUI)	Feature Type	Attention Weight $\geq$ 0.5 % of Max
427.31:Atrial Fibrillation (C0004238)	Text	fib, **fibrillation**
	KGE	C0004238 - Atrial fibrillation
	W2V	C0004238 - Atrial Fibrillation C0344434 - ECG: atrial fibrillation
	$\text{Text}_{\text{human}}$	a fib, atrial fibrillation
	$\text{CUI}_{\text{human}}$	C0004238-Atrial fibrillation
038.9:Septicemia (C0036690)	Text	septic
	KGE	C0349410 - Single organ dysfunction (2:0.9-1.0) C0026766 - Multiple organ failure (5:0.6-0.9) C0277524 - Infectious colitis C1457868 - Worse
	W2V	C0349410 - Single organ dysfunction C0004030 - Aspergillosis
	$\text{Text}_{\text{human}}$	drop in blood pressure, iv fluids, pressors, hyperdynamic left ventricle presumed to be septic, samples grew mold
	$\text{CUI}_{\text{human}}$	C0020649 - Low blood pressure C0349410 - Single organ dysfunction C0948268 - Hemodynamic instability C0009450 - Disorder due to infection
995.92:Severe Sepsis (C1719672)	Text	**septic**, pressors, central
	KGE	C0026766 - Multiple organ failure (4:0.5-1.0) C0349410 - Single organ dysfunction (2:0.8) C1457868 - Worse C0004030 - Aspergillosis
	W2V	C0349410 - Single organ dysfunction C0004030 - Aspergillosis
	$\text{Text}_{\text{human}}$	drop in blood pressure, iv fluids, pressors, hyperdynamic left ventricle presumed to be septic, multisystem organ failure worsened, hemodynamic status worsened
	$\text{CUI}_{\text{human}}$	C0020649 - Low blood pressure C0026766 - Multiple organ failure C0948268 - Hemodynamic instability C0009450 - Disorder due to infection C0443343 - Unstable status
96.72:Continuous invasive mechanical ventilation for 96 consecutive hours or more (C2349745)	Text	Intubated, mold, **which**, aspergillis
	KGE	C0553891 - Extubation of trachea C0011065 - Death (2:0.65-0.9) C0425043 - Death of relative C0205463 - Physiologic
	W2V	C0011065 - Death C0278060 - Mental state
	$\text{Text}_{\text{human}}$	Intubation, remained intubated, over the next several days, extubation
	$\text{CUI}_{\text{human}}$	C0021925 - Intubation C0553891 - Extubation of trachea
401.9:Essential Hypertension (C0085580)*	Text	hypertension
	KGE	C0020538 - Hypertensive disorder (4:0.6-1.0) C0020473 - Hyperlipidemia C0221155 - Systolic hypertension (3:0.5-0.7) C0235222 - Diastolic hypertension (3:0.5-0.6)
	W2V	C0428465 - Serum lipids high C0221155 - Systolic hypertension (3:0.7-0.8) C0235222 - Diastolic hypertension (4:0.6-0.7) C1696708 - Prehypertension (2:0.7) C0019099 - Congo-Crimean hemorrhagic fever C0020538 - Hypertensive disorder C0020473 - Hyperlipidemia
	$\text{Text}_{\text{human}}$	no† prior history of htn, hypertension, due to pain† post procedure or undiagnosed† htn
	$\text{CUI}_{\text{human}}$	C0030193 - Pain† C0262534 - Labile hypertension due to being in a clinical environment†

Table 5: Comparison of tokens with attention weights $\geq$ 0.5 of the highest attention weight across feature types. $\star$ indicates a false positive label example. **Bold** font indicates text tokens with highest attention weights. †indicates tokens are crucial to preventing false predictions. CUI tokens are ordered from highest to lowest weights with number of occurrences and attention weight % range in parentheses.

When the label CUIs are not present in the input document, as in "038.9:Septicemia", "995.92:Severe Sepsis", and "96.72:Continuous invasive mechanical ventilation for 96 consecutive hours or more" examples, the model's attention mechanism identifies more clinically informative CUIs in the KGE model than in the W2V model. Slightly different KGE CUIs and attention weight distributions are associated with "038.9" and "995.92" labels. In contrast, the exact same W2V CUIs and almost identical attention distributions are associated with both labels. In the case of label "96.72", KGE model does identify one of the relevant tokens (C0553891 - Extubation of trachea, which implies prior intubation and continuous invasive mechanical ventilation), while W2V model does not identify either of them. While both models predict equally correct labels, the external knowledge implicitly represented in KGE helps facilitate more clinically intuitive attention-based reasoning compared to W2V embeddings.

Both KGE and W2V attentions include neighboring tokens and their synonyms, e.g. C0011065, C0425043, C0278060, C0019099 for "96.72" and "401.9". The presence of extraneous CUIs that are similar concept-wise to relevant collocate CUIs such as "C0019099 - Congo-Crimean hemorrhagic fever" and "C0425043 - Death of relative" highlights the importance of optimizing the concept extraction accuracy in concept-based models. While the extraneous CUI tokens can be clinically associated with the relevant tokens or the predicted label concept-wise, the text-input attention mechanism can identify tokens that have no clinical importance as being most associated with a correctly predicted label, e.g. "which" has the highest attention weight for the label "96.72". Making correct predictions based on unjustifiable reasoning is undesirable as it raises concerns over the model's trustworthiness.

Regardless of feature type, the attention mechanism ignores negation. Negated mentions are common in EHR as clinicians document their assessments, noting findings as absent as opposed to not mentioning them at all; the latter may lead to the undesirable assumption of not having made an assessment. As seen in the "401.9" example, "no" or "undiagnosed" are not considered relevant, as indicated by the tokens being omitted by attention weights, leading to a false prediction. In contrast, the clinician-annotated example shows these negation tokens are relevant for excluding the false positive label. As there are no CUIs indicating negation or a diagnostic absence in the input document, it appears that negation in the text input is filtered out during the concepts extraction step. Despite the absence of negation-like CUIs in the input documents, clinician-annotated CUIs include concepts that can prevent the false "401.9" prediction: "C0262534 - hypertension due to being in a clinical environment" in conjunction with "C0030193 - pain". This observation regarding negation-related errors aligns with findings in Hossain et al. (2020) (despite their analysis being with respect to machine translation systems), indicating that the presence of negation can significantly lower downstream output quality. The presence of CUIs that can lead to excluding negation-related false positive labels without needing to encode negation as a concept suggests a potential alternative for future works in addressing this challenge.

4.2 Limitations & Future Works

UMLS KG covers broad medical concepts and relations that may not overlap with rules in the ICD-9 coding guidelines that are periodically updated. While our results suggest that GCN performance is impacted by graph construction approaches, heuristics based on clinical reasoning may not be as useful for ICD coding, particularly if the intended purpose is non-clinical. Future works on ICD-9 coding on this dataset should explore KG construction from concepts and relations according to rules in the dataset's edition of ICD coding guidelines.

Our qualitative analysis is based on a small sample size and one clinician's annotation; future works with more resources should expand the sample size and include analysis by multiple experts from the intended application domain. To maintain a defined scope of our study with respect to existing reference models results, our experiments are conducted only on one dataset and one version of ICD-9 codes, excluding ICD-10. A more recent dataset, MIMIC-IV (Johnson et al., 2023), has been released since the time of our experiments. Additionally, a recent study by Edin et al. (2023) comparing benchmark models on both MIMIC-III (Johnson et al., 2016) and MIMIC-IV (Johnson et al., 2023) datasets with results on both ICD-9 and ICD-10 codes should facilitate the extrapolation of our approach to broader datasets.

As shown in Table 1, documents represented as concept-based (CUI) tokens are 1/3 in length of those represented as text-based tokens. The shorter input documents enable future experiments on larger models previously deemed incompatible. Since text-based models still lead in performance, utilizing CUI descriptions instead of the CUI themselves as features is worth exploring. CUI and ICD codes have meanings through their corresponding descriptions. Considering KGE's low concept coverage and recent works involving domain-knowledge-augmented (UMLS) BERT (Michalopoulos et al., 2021), future research directions may include leveraging generative models in KG expansion and using concept-based KGE or GCN encoded relational information to augment text-based features.

Standard MLC evaluation metrics, which consider all label classes to be independent (Kosmopoulos et al., 2015), can be problematic as a model predicting more generalized labels, e.g. parent labels encompassing the ground truths, or sibling labels in the ICD code structure, would be considered as low-performing as a model predicting completely unrelated labels. Depending on downstream applications, hierarchical evaluation metrics that are more suitable for MLC of dependent label classes should also be considered for automatic ICD coding evaluation.

5 Conclusion

Our investigation into implicit graph representation in the input space highlights the benefits of KGE over corpus-based concept-feature embeddings in improving the model's attention-based reasoning interpretability. The experiments involving explicit relational information representation through graph construction approaches demonstrate the limitations of GCN as a standalone encoder in ICD coding task. The qualitative analysis of the attention-based reasoning identifies challenges that contribute to erroneous predictions and provides insight into how KG construction may be improved in future works. Our contributions underscore the potential for graph concept-based features while addressing the difficulties associated with medical codes classification as an MLC problem from long input documents, domain knowledge requirements, and interpretability.

Acknowledgements

We thank the anonymous reviewers, Josef van Genabith, and Tanja Bäumel for their constructive feedback. The work was partially funded by the European Union (EU) through the project PERKS (ID: 101120323) under the "Digital, Industry, and Space" funding program and the German Federal Ministry of Education and Research (BMBF) through the project XAINES (ID: 01IW20005).

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Andrew L. Beam, Benjamin Kompa, Allen Schmaltz, Inbar Fried, Griffin Weber, Nathan Palmer, Xu Shi, Tianxi Cai, and Isaac S. Kohane. 2020. Clinical Concept Embeddings Learned from Massive Sources of Multimodal Medical Data. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 25:295–306.

Olivier Bodenreider. 2004. The Unified Medical Language System (UMLS): Integrating Biomedical Terminology. *Nucleic Acids Research*, 32(Database issue):D267–D270.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating Embeddings for Modeling Multi-relational Data. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Pengfei Cao, Yubo Chen, Kang Liu, Jun Zhao, Shengping Liu, and Weifeng Chong. 2020. HyperCore: Hyperbolic and Co-graph Representation for Automatic ICD Coding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3105–3114, Online. Association for Computational Linguistics.

David Chang, Ivana Balažević, Carl Allen, Daniel Chawla, Cynthia Brandt, and Andrew Taylor. 2020. Benchmark and Best Practices for Biomedical Knowledge Graph Embeddings. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, pages 167–176, Online. Association for Computational Linguistics.

Edward Choi, Mohammad Taha Bahadori, Elizabeth Searles, Catherine Coffey, Michael Thompson, James Bost, Javier Tejedor-Sojo, and Jimeng Sun. 2016. Multi-layer Representation Learning for Medical Concepts. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1495–1504, San Francisco California USA. ACM.

Edward Choi, Zhen Xu, Yujia Li, Michael Dusenberry, Gerardo Flores, Emily Xue, and Andrew Dai. 2020. Learning the Graphical Structure of Electronic Health Records with Graph Convolutional Transformer. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(01):606–613.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Joakim Edin, Alexander Junge, Jakob D. Havtorn, Lasse Borgholt, Maria Maistro, Tuukka Ruotsalo, and Lars Maaløe. 2023. Automated Medical Coding on MIMIC-III and MIMIC-IV: A Critical Review and Replicability Study. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, pages 2572–2582. Association for Computing Machinery.

Md Mosharaf Hossain, Antonios Anastasopoulos, Eduardo Blanco, and Alexis Palmer. 2020. It's not a Non-Issue: Negation as a Source of Error in Machine Translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3869–3885. Association for Computational Linguistics.

Institute of Medicine (US) Committee on the Future of Primary Care. 1994. *Defining Primary Care: An Interim Report*. National Academies Press (US).

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li-wei H. Lehman, Leo A. Celi, and Roger G. Mark. 2023. MIMIC-IV, a Freely Accessible Electronic Health Record Dataset. 10(1):1. Publisher: Nature Publishing Group.

Alistair E.W. 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*, 3:160035.

Rajvir Kaur, Jeewani Anupama Ginige, and Oliver Obst. 2021. A Systematic Literature Review of Automated ICD Coding and Classification Systems using Discharge Summaries. *arXiv:2107.10652 [cs]*.

Thomas N. Kipf and Max Welling. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *Proceedings of the 5th International Conference on Learning Representations*, ICLR '17, Palais des Congrès Neptune, Toulon, France.

Aris Kosmopoulos, Ioannis Partalas, Eric Gaussier, Georgios Paliouras, and Ion Androutsopoulos. 2015. Evaluation Measures for Hierarchical Classification: A Unified View and Novel Approaches. *Data Mining and Knowledge Discovery*, 29(3):820–865.

Siting Liang, Mareike Hartmann, and Daniel Sonntag. 2023. Cross-domain German medical named entity recognition using a pre-trained language model and unified medical semantic types. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 259–271, Toronto, Canada. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective Approaches to Attention-based Neural Machine Translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal. Association for Computational Linguistics.

Hehuan Ma, Yu Rong, and Junzhou Huang. 2022. Graph Neural Networks: Scalability. In Lingfei Wu, Peng Cui, Jian Pei, and Liang Zhao, editors, *Graph Neural Networks: Foundations, Frontiers, and Applications*, pages 99–119. Springer Nature Singapore, Singapore.

Takuma Matsubara, Makoto Miwa, and Yutaka Sasaki. 2023. Distantly Supervised Document-Level Biomedical Relation Extraction with Neighborhood Knowledge Graphs. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 363–368, Toronto, Canada. Association for Computational Linguistics.

George Michalopoulos, Yuanxin Wang, Hussam Kaka, Helen Chen, and Alexander Wong. 2021. Umls-BERT: Clinical Domain Knowledge Augmentation of Contextual Embeddings Using the Unified Medical Language System Metathesaurus. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1744–1753, Online. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *CoRR*, abs/1301.3781.

Elias Moons, Aditya Khanna, Abbas Akkasi, and Marie-Francine Moens. 2020. A Comparison of Deep Learning Methods for ICD Coding of Clinical Records. *Applied Sciences*, 10(15):5262.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable Prediction of Medical Codes from Clinical Text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, arXiv:1802.05695, pages 1101–1111, New Orleans, Louisiana. Association for Computational Linguistics.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch. *NIPS-W*.

Lin Qiu, Yunxuan Xiao, Yanru Qu, Hao Zhou, Lei Li, Weinan Zhang, and Yong Yu. 2019. Dynamically Fused Graph Network for Multi-hop Reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6140–6150, Florence, Italy. Association for Computational Linguistics.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA.

Deepak Saini, Arnav Kumar Jain, Kushal Dave, Jian Jiao, Amit Singh, Ruofei Zhang, and Manik Varma. 2021. GalaXC: Graph Neural Networks with Labelwise Attention for Extreme Classification. In *Proceedings of the Web Conference 2021*, pages 3733–3744, Ljubljana Slovenia. ACM.

Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. 2018. RotatE: Knowledge Graph Embedding by Relational Rotation in Complex Space. In *International Conference on Learning Representations*.

Fei Teng, Wei Yang, Li Chen, LuFei Huang, and Qiang Xu. 2020. Explainable Prediction of Medical Codes With Knowledge Graphs. *Frontiers in Bioengineering and Biotechnology*, 8.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Thanh Vu, Dat Quoc Nguyen, and Anthony Nguyen. 2020. A Label Attention Model for ICD Coding from Clinical Text. In *Twenty-Ninth International Joint Conference on Artificial Intelligence*, volume 4, pages 3335–3341.

Minjie Wang, Da Zheng, Zihao Ye, Quan Gan, Mufei Li, Xiang Song, Jinjing Zhou, Chao Ma, Lingfan Yu, Yu Gai, Tianjun Xiao, Tong He, George Karypis, Jinyang Li, and Zheng Zhang. 2020. Deep Graph Library: A Graph-Centric, Highly-Performant Package for Graph Neural Networks.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. 2018. How Powerful are Graph Neural Networks? *CoRR*, abs/1810.00826.

Zhitao Ying, Jiaxuan You, Christopher Morris, Xiang Ren, Will Hamilton, and Jure Leskovec. 2018. Hierarchical Graph Representation Learning with Differentiable Pooling. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Zheng Yuan, Chuanqi Tan, and Songfang Huang. 2022. Code Synonyms Do Matter: Multiple Synonyms Matching Network for Automatic ICD Coding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2:*

Short Papers), pages 808–814, Dublin, Ireland. Association for Computational Linguistics.

Da Zheng, Xiang Song, Chao Ma, Zeyuan Tan, Zihao Ye, Jin Dong, Hao Xiong, Zheng Zhang, and George Karypis. 2020. DGL-KE: Training Knowledge Graph Embeddings at Scale. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, pages 739–748, New York, NY, USA. Association for Computing Machinery.

Daoming Zong and Shiliang Sun. 2020. GNN-XML: Graph Neural Networks for Extreme Multi-label Text Classification. *arXiv:2012.05860 [cs]*.

A Supplementary Material

Additional information regarding the UMLS and ICD-9 codes are explained in the following sections. Implementation details including hyper-parameters specified in our experiments are provided for reproducibility. Our Git repository[10] also contains further implementation details and code to reproduce our experiments. Additional experiment results not part of the main contributions are also included.

A.1 ICD-9 Code Structure

Moons et al. (2020) describes the structure of ICD-9 codes as consisting of at most five numbers: the first three represent a disease category, a fourth number narrows down to specific diseases, and a fifth number differentiates between specific disease variants. This structure creates a hierarchical taxonomy with up to 4 layers (L1 - L4) from the root level as shown in Figure 4.

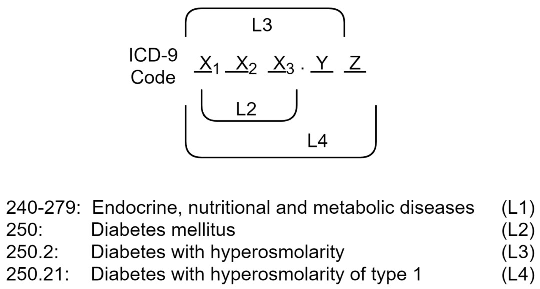

Figure 4: An example of ICD-9 codes with code descriptions illustrating the hierarchical layers. The example here shows how diabetes mellitus and its specific variants are represented in the ICD-9 code taxonomy. Illustration is reproduced from Moons et al. (2020).

Being a subset of the UMLS Knowledge Bases (Bodenreider, 2004), ICD-9 codes have corresponding Concept Unique Identifiers (CUIs) in the UMLS, which also contains Semantic Type Information (TUI); examples from the Top-50 ICD-9 codes of the MIMIC-III dataset and their UMLS information are shown in Table 6. Within the UMLS, high-level grouping based on TUI is noted among the codes in Table 6; both C0176511 and C0189898 share the same TUI as they both describe diagnostic procedures. The grouping in the UMLS does not always correspond to the same hierarchy in the ICD-9 taxonomy as noted by the mentioned codes being under two distinct L2-level numbers.

ICD-9	CUI	TUI	Description
33.24	C0176511	T060	Closed [endoscopic] biopsy of bronchus
37.23	C0189898	T060	Catheterization of both left and right heart
38.91	C0007431	T061	Arterial catheterization
38.93	C0162203	T058	Venous catheterization, not elsewhere classified

Table 6: Examples of ICD-9 codes and their corresponding UMLS CUI, TUI, and descriptions from the Top-50 ICD-9 code labels of the MIMIC-III dataset (Johnson et al., 2016).

A.2 Implementation Details

The following sections describe hyper-parameters used in our experiments. We do not fine-tune hyper-parameters for our specific dataset training; we prioritize keeping hyper-parameters as close as possible to those reported as optimal by Vu et al. (2020) for the LAAT model.

A.2.1 LAAT

As in Vu et al. (2020), we train for 50 epochs, using a batch size of 8, with AdamW (Loshchilov and Hutter, 2018) optimizer and learning rate of 0.001. We also use a learning rate scheduler to reduce the learning rate by 10% if there is no improvement in F1_{micro} on the validation set for 5 epochs. We apply a drop-out probability of 0.3. We specify the LSTM hidden size $u = 256$ and projection size $d_a = 256$ for the Top-50 version and $u = 512$, $d_a = 512$ for the Full version as these are the optimal hyper-parameters reported in Vu et al. (2020). The text input results in Table 8 verify that our re-implementation of the LAAT model reproduces comparable performance on the same dataset as reported in Vu et al. (2020) following the same pre-processing steps and hyper-parameters.

A.2.2 KGE

We obtain KGE for CUI entities following training steps described in Chang et al. (2020) using DGL-KE (Zheng et al., 2020) implementation of TransE (Bordes et al., 2013). The *case4* train, dev,

[10] https://github.com/pokarats/CoDER

Version	Model	Precision		Recall		F1		AUC		P@5
		macro	micro	macro	micro	macro	micro	macro	micro	
Top-50	$GCN_{0.7}$	**62.12**	**67.81**	38.22	45.02	47.33	**54.11**	**84.54**	**87.40**	**56.00**
	$GCN_{0.83}$	56.44	63.22	**41.61**	**47.05**	**47.98**	53.95	83.75	86.22	54.23

Table 7: Results from GCN_{BASE} experiments on the MIMIC-III Top-50 with CUI input type, comparing entity linking threshold of 0.7 and 0.83. All GCN models use KGE as node embeddings to represent each CUI node in a graph.

Encoder	Implementation	F1		AUC		P@5
		macro	micro	macro	micro	
LAAT (50)	Vu et al. (2020)	66.60	71.50	92.50	94.60	67.50
	Ours	66.55	71.01	92.79	94.60	67.28
LAAT (Full)	Vu et al. (2020)	8.70	58.10	92.60	98.80	81.80
	Ours	9.87	57.20	89.84	98.56	80.91

Table 8: Text input results on the MIMIC-III Top-50 and Full test sets from our implementation of the LAAT model in comparison to the results reported in Vu et al. (2020).

and test files are downloaded from SNOMED CT Knowledge Graph Embeddings Git repository[11]. We use the following key configuration parameters for training:

```
model_name: TransE_l2, max_step: 60000,
batch_size: 1024, batch_size_eval: 1000,
neg_sample_size: 64,
neg_sample_size_eval: 90000,
hidden_dim: 100, lr: 0.1,
gamma: 10.0,
adversarial_temperature: 1.0,
regularization_coef: 1e-07,
pairwise: false, loss_genre: Logsigmoid
```

A.2.3 GCN

Our 2-layer GCN classification is implemented using DGL (Wang et al., 2020) with PyTorch (Paszke et al., 2017) backend. We control the hyper-parameters to be as similar to the LAAT specifications as possible. For the GCN layers, we specify the hidden size $u = 256$ and projection size $d_a = 256$ for the Top-50 and $u = 512$, $d_a = 512$ for the Full versions analogous to the hyper-parameters for the LSTM encoder in the LAAT experiments. We train for 50 epochs, using a batch size of 8, and learning rate of 0.001, and AdamW (Loshchilov and Hutter, 2018). We also use the same learning rate scheduler and dropout probability.

[11] https://github.com/dchang56/snomed_kge

A.3 From EHR to GCN Graph Construction

To demonstrate the relational characteristics in EHR structural features and clinical reasoning, we manually annotate 5 randomly selected discharge summaries from the Top-50 version of the MIMIC-III (Johnson et al., 2016) training dataset. The annotations in Figure 5 illustrate that extracted concepts representing parts of a note provide sufficient information for a clinical domain expert to relate the assigned ICD codes to relevant parts of the document. Despite having only clinical domain knowledge without ICD coding training, we are able to identify relevant spans of text and CUIs that relate to the assigned ICD codes.

A.4 CUI Extraction Entity Linking Threshold Comparison

We notice many CUIs in the randomly selected samples do not seem relevant to the clinical presentation described in the note nor assigned ICD codes. We verify if a more selective (higher) threshold has an impact on performance by experimenting with the Top-50 GCN_{BASE} and setting the EL threshold to 0.83. Results in Table 7 show performance scores of the GCN_{BASE} model with EL thresholds of 0.7 and 0.83. Evaluation scores are higher in most metrics with the 0.7 threshold. $Recall_{macro,micro}$ and $F1_{macro}$ are the only metrics where the 0.83 threshold shows higher performance. Considering the evaluation scores between the two EL thresholds are within a few % points of each other, it does not seem computationally worthwhile to repeat all experiments with the 0.83 threshold.

A.5 Runtime Comparison

LAAT experiments are run on NVIDIA GeForce RTX 3090 and GCN on NVIDIA RTX A6000. Table 9 illustrates training runtime and mean input document lengths in number of text or CUI tokens for the LAAT model. CUI input models (W2V and KGE) show training runtimes that are multitudes less than the text input model. The shorter

Note id: 16525_134157

Text Input Type

name known lastname known firstname unit no numeric identifier admission date discharge date date of birth sex f service addendum this is an addendum to the previous **MICU** green discharge summary l hospital course the patient was admitted on after **being found down** l at that time she was noted to have **multiple intracranial bleeds**l the patient then had a **seizure** and was **intubated on** [SIC - TEMPORAL] l **ct scan** on showed **increasing midline shift** and at that time the decision was made to make the patient comfort measures only per the family l the patient was then **extubated on** [SIC - TEMPORAL] and **transferred to the medicine floor for comfort measures** where she was under the care of dr first name4 namepattern1 last name namepattern1lin conjunction with the palliative care consult service the patient was made comfortable utilizing **morphine drip** and discharge status the patient expired l final diagnosis l **subarachnoid hemorrhage with intraparenchymal hemorrhage**lname6 md name8 md m dlmd number dictated by last name l namepattern1 medquist36 d t job job number

MIMIC-III ICD Codes (Top-50 Version)

276.1: Hyposmolality and/or hyponatremia (C0020645)

401.9: Essential Hypertension (C0085580)

285.9: Anemia (C0002871)

96.04: Intubation, Intratracheal (C0021932)

96.72: Continuous invasive mechanical ventilation for 96 consecutive hours or more (C2349745)

38.93: Venous catheterization, not elsewhere classified (C0162203)

38.91: Arterial catheterization (C0007431)

CUI Input Type

C0151699 **C0019080** **C0036572** **C0751495** **C0270844** **C0021925** **C0040405** C3472245 C1699633 C0034606 **C0442808** **C0221106** C0178415 C0020459 C0455769 C0423908 C1879489 C0242485 **C0553891** **C0348011** **C0030704** **C0150521** C0184573 C0150192 C0009763 C0030231 **C2698261** C1260880 C0586514 C0011900 C0002928 **C0019080** C0029163 C0426747 **C0475072** **C3698285** **C2937358** C0026850, C0242271 C0242271

Highlighted Input Document CUI - Text Definition (SG:TUI)

C0151699 - Intracranial hemorrhage **(DISO:T046)**
C0019080 - Hemorrhage **(DISO:T046)**
C0036572 - Seizure **(DISO:T184)**
C0751495 - Partial Seizure **(DISO:T047)**
C0270844 - Tonic seizure **(DISO:T047)**
C0021925 - Intubation **(PROC:T061)**
C0040405 - Computerized axial tomography **(PROC:T060)**
C0442808 - Increasing **(CONC:T169)**
C0221106 - Alkalemia **(DISO:T047)**
C0553891 - Extubation of trachea **(PROC:T061)**
C0348011 - Surgical transfer action **(CONC:T169),**
C0030704 - Patient transfer **(PROC:T058)**
C0150521 - Comfort measures **(PROC:T058),**
C2698261 - Measurement of morphine **(PROC:T059)**
C0475072 - Cerebral hemorrhage following injury **(DISO:T037),**
C3698285 - Nontraumatic intraparenchymal cerebral hemorrhage **(DISO:T046)**
C2937358 - Cerebral hemorrhage **(DISO:T046)**

Annotation Color References:

ICD Codes CUI - Semantic Group (SG) - Semantic Type (TUI)

C0020645 **DISO - T033**
C0085580 **DISO - T047**
C0002871 **DISO - T047**
C0154298 **DISO - T047**
C0021932 **PROC - T061**
C2349745 **PROC - T061**
C0162203 **PROC - T058**
C0007431 **PROC - T061**

SG - TUI - Description

DISO T033 - Finding
DISO T037 - Injury or Poisoning
DISO T046 - Pathologic Function
DISO T047 - Disease or Syndrome
DISO T184 - Sign or Symptom
CONC T169 - Functional Concept
PROC T060 - Diagnostic Procedure
PROC T061 - Therapeutic or Preventive Procedure
PROC T058 - Health Care Activity
PROC T059 - Laboratory Procedure

Figure 5: Spans of text and extracted CUIs in the input document are highlighted with colors that correspond to the assigned ICD codes. Red-highlighting designates codes that we cannot definitively infer from the input document. Additional information provided by the UMLS such as Semantic Type Information (TUI), Semantic Group (SG), and corresponding CUI to each ICD code demonstrate correspondence between the input document and output label space. The additional highlight colors in the annotation references group CUIs by their SG: DISO, CONC, and PROC.

Version	Model	Training Runtime (hh:mm:ss)	Mean Training Input Length (tokens)
Top-50	LAAT$_{text}$	04:53:10	1783
	LAAT$_{W2V}$	00:41:13	396
	LAAT$_{KGE}$	00:40:31	396
Full	LAAT$_{text}$	21:35:34	1731
	LAAT$_{W2V}$	05:40:58	385
	LAAT$_{KGE}$	05:36:59	385

Table 9: Training runtime comparison between text and CUI input types for the Top-50 and Full versions of the MIMIC-III dataset. Mean number of tokens for the training partition is provided. Runtime is only for training the model and is exclusive of time required for concepts extraction and pre-processing.

Version	Model	Training Runtime (hh:mm:ss)	Mean # Nodes	Mean # Edges	Mean # Sub-graphs
Top-50	GCN$_{BASE}$	00:17:48	246	419	167
	GCN$_{EHR}$	00:09:44	246	513	145
	GCN$_{COMBO}$	00:14:11	246	684	90
Full	GCN$_{BASE}$	01:04:40	241	408	165
	GCN$_{EHR}$	00:49:25	241	1024	50
	GCN$_{COMBO}$	01:08:22	241	1189	28

Table 10: Training runtime comparison between GCN graph construction approaches for the Top-50 and Full versions of the MIMIC-III dataset. Mean node, edge, and sub-graphs (connected components) numbers for the training partition are provided. Runtime is for training the GCN model and is exclusive of time spent on pre-processing or building graph datasets.

runtime appears to correlate with shorter average input lengths. Table 10 compares training runtime across GCN graph construction approaches. Contrary to LAAT models, there does not seem to be a notable relationship between runtime and graph nodes, edges, or sub-graphs numbers. As noticeable in the table, graph construction heuristic affects the number of sub-graphs on average; more edges result in fewer sub-graphs. Due to the multiple steps involved in our proposed pipeline, from concepts extraction to graph construction heuristics, application to other datasets requires additional data preparation and pre-processing time.

The LAAT model suffers from time and memory complexity issues associated with the LSTM encoder and attention mechanism. The GCN models are also limited by the memory requirement to store a completed adjacency matrix; additional sampling algorithms and alternative models are required for scalability to larger datasets (Ma et al., 2022).