Auxiliary Knowledge-Induced Learning for Automatic Multi-Label Medical Document Classification

Xindi Wang[1,2], **Robert E. Mercer**[1], **Frank Rudzicz**[2,3,4]
[1]Department of Computer Science, University of Western Ontario, Canada
[2] Vector Institute for Artificial Intelligence, Canada
[3] Faculty of Computer Science, Dalhousie University, Canada
[4] Department of Computer Science, University of Toronto, Canada
xwang842@uwo.ca, mercer@csd.uwo.ca, frank@dal.ca

Abstract

The International Classification of Diseases (ICD) is an authoritative medical classification system of different diseases and conditions for clinical and management purposes. ICD indexing assigns a subset of ICD codes to a medical record. Since human coding is labour-intensive and error-prone, many studies employ machine learning to automate the coding process. ICD coding is a challenging task, as it needs to assign multiple codes to each medical document from an extremely large hierarchically organized collection. In this paper, we propose a novel approach for ICD indexing that adopts three ideas: (1) we use a multi-level deep dilated residual convolution encoder to aggregate the information from the clinical notes and learn document representations across different lengths of the texts; (2) we formalize the task of ICD classification with auxiliary knowledge of the medical records, which incorporates not only the clinical texts but also different clinical code terminologies and drug prescriptions for better inferring the ICD codes; and (3) we introduce a graph convolutional network to leverage the co-occurrence patterns among ICD codes, aiming to enhance the quality of label representations. Experimental results show the proposed method achieves state-of-the-art performance on a number of measures.

Keywords: extreme multi-label text classification, knowledge-enhanced text classification, graph convolutional network, ICD classification

1. Introduction

Electronic health records (EHRs)[1] contain all of the key administrative clinical data relevant to a person's care under a particular provider, including demographics, past history notes, progress notes, laboratory reports, diagnoses, and medications. EHRs have been increasingly used in a variety of settings which provide opportunities to enhance patient care and facilitate clinical research. The International Classification of Diseases (ICD)[2] is often used as a surrogate for clinical outcomes of interest, as it is designed to provide diagnostic assistance and classify health disorders. ICD is a medical classification taxonomy maintained by the World Health Organization (WHO)[3], which serves a broad range of uses in diagnostic processes, epidemiology, health management, and other clinical activities. There are two types of codes in the ICD coding system, namely procedure codes[4] (that are used to identify specific surgical, medical, or diagnostic interventions) and diagnosis codes[5] (that are used to identify diseases, disorders and symptoms). In the 10$^{\text{th}}$ edition, there are over 70,000 procedure codes and over 69,000 diagnosis codes[6].

The task of ICD indexing aims to associate ICD codes with EHR documents. Currently, ICD indexing is carried out manually by human annotators, which is labour-intensive and error-prone (O'Malley et al., 2005). Therefore, automatic annotation has gained interest in the research community. Automatic ICD indexing can be regarded as an extreme multi-label text classification (XMTC) problem, where each EHR document can be labeled with multiple ICD codes. Compared with standard multi-label classification tasks, XMTC finds relevant labels from an extremely large set of labels. Large-scale ICD indexing is severely challenged by several problems. First, the distribution of ICD codes is extremely long-tailed: while some ICD codes occur frequently, many others seldom appear, if at all, because of the rarity of the diseases. For instance, among the 942 unique 3-digit ICD codes in the MIMIC-III (Johnson et al., 2016) dataset (the largest publicly available medical dataset), the ten most common codes account for 26% of all code occurrences and the least common 437 codes account for only 1% (Bai and Vucetic, 2019). Second, unstructured clinical texts are long (containing an average of 1596 words in the MIMIC-III dataset) and noisy (including irrelevant information, misspellings, and non-standard abbreviations). These difficulties

[1]https://www.cms.gov/Medicare/E-Health/EHealthRecords
[2]https://www.who.int/standards/classifications/classification-of-diseases
[3]https://www.who.int
[4]https://en.wikipedia.org/wiki/Procedure_code
[5]https://en.wikipedia.org/wiki/Diagnosis_code
[6]https://www.cdc.gov/nchs/icd/icd10cm_pcs.htm

LREC-COLING 2024, pages 2006–2016

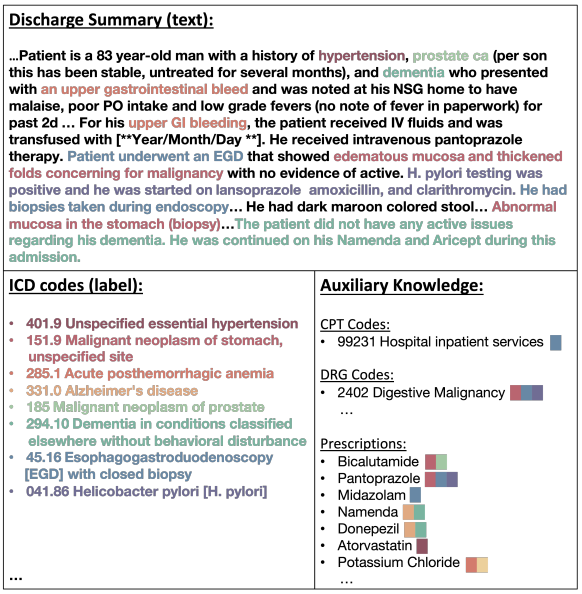

Figure 1: An example of a patient record from the MIMIC-III dataset which includes the discharge summary, assigned ICD codes and auxiliary knowledge. We colour each code and its corresponding mentions in the discharge summary and auxiliary knowledge. We use the auxiliary knowledge of the notes to predict relevant codes of summary.

make extracting relevant information from clinical texts, for all ICD codes, very challenging.

We propose a novel auxiliary knowledge-induced medical code labelling architecture to address these issues. To lessen the problems caused by the long-tailed distribution of ICD codes, we leverage code co-occurrence and join auxiliary knowledge with the clinical texts to improve coding accuracy.

Code Co-occurrence The co-occurrence of codes in clinical texts provides valuable insights into the relationships between different diseases or conditions. For instance Figure 1 shows that the code for "Dementia in conditions classified elsewhere without behavioral disturbance" (294.10) can be easily captured from the text (i.e., the highlighted words in desaturated cyan). However, inferring the code for "Alzheimer's disease" (331.0) is more challenging as the clues are less explicit. Fortunately, there is a strong association between these two diseases, with "Alzheimer's disease" being one of the most common causes of "dementia". This association can be captured by leveraging the fact that the codes for these two diseases often co-occur in clinical texts. By leveraging code co-occurrence patterns, we can capture the dependencies and correlations among codes. This allows us to better understand the context in which specific codes occur and make more accurate predictions based on these relationships beyond using only the clinical texts themselves.

Auxiliary Knowledge EHR auxiliary knowledge is widely available, but is often overlooked in previous studies. In addition to clinical texts, an EHR document is also associated with various auxiliary knowledge such as code systems (other than ICD codes) and drug prescriptions. Specifically, we are interested in two code terminologies (diagnosis-related group (DRG)[7] codes and current procedural terminology (CPT)[8] codes), as well as the medications prescribed to patients, which could be strong indicators of ICD predictions. For instance, Figure 1 shows "Namenda" in drug prescriptions, which would strongly suggest the patient is most likely to have Alzheimer's disease. By incorporating auxiliary knowledge, we augment the information available for coding tasks. This external knowledge provides additional context and insights that can aid in accurately assigning appropriate ICD codes.

To alleviate the long text issue, we introduce a multi-level dilated residual convolutional network to ensure the extracted representations focus on the long clinical notes. With multi-level dilation rates, convolutions can capture broader contexts while preserving spatial resolution, enabling the network to have a larger receptive field without increasing the number of parameters.

The contributions of this paper are as follows:

- We propose a framework that is capable of simultaneously dealing with both long-tail and long-text issues in the ICD prediction task.
- To alleviate the long-tail issue, we propose a graph convolutional network to leverage code co-occurrence which captures the connections among codes with different frequencies. We integrate external knowledge using an auxiliary knowledge mask which constrains the large space of possible ICD codes.
- To handle long texts, we use a multi-level dilated residual convolutional network, enabling the model to capture long-range dependencies and local context with different dilation rates.
- We evaluate on a widely used automatic ICD coding dataset, MIMIC-III, and the results show that our proposed model outperforms previous methods on a number of measures.

2. Related Work

2.1. Automatic ICD Indexing

Automatic ICD indexing is a long-standing task in the healthcare domain. To the best of our knowl-

[7] https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/AcuteInpatientPPS/MS-DRG-Classifications-and-Software

[8] https://www.ama-assn.org/amaone/cpt-current-procedural-terminology

edge, the earliest work was proposed by Larkey and Croft (1996), who combined three classifiers (K-nearest-neighbour, relevance feedback, and Bayesian independence classifiers) to automatically assign ICD codes to dictated inpatient discharge summaries. de Lima et al. (1998) proposed a hierarchical model that used the topology of the code structure, and then calculated the cosine similarity of TF-IDF representations between clinical texts and ICD codes. A variety of rule-based methods (Crammer et al., 2007; Farkas and Szarvas, 2008) and statistical machine learning algorithms, such as support vector machines (Lita et al., 2008), were later applied to the ICD coding task.

With deep neural networks, many previous works have proven the effectiveness of convolutional neural networks (CNNs), recurrent neural networks (RNNs), and their variants for ICD coding. Mullenbach et al. (2018) combined a CNN with an attention mechanism to capture relevant information in the clinical texts for each ICD code. Xie et al. (2019) further improved the CNN attention model by incorporating multi-scale feature attention. Many other CNN variants were proposed to deal with lengthy and noisy clinical texts, such as MultiResCNN (Li and Yu, 2020), DCAN (Ji et al., 2020), and EffectiveCAN (Liu et al., 2021). MultiResCNN introduced a multi-filter residual CNN to capture text patterns of different lengths and used a residual convolutional layer to enlarge the receptive field. DCAN stands for 'dilated convolutional attention network', which used a single filter and the dilation operation to control the receptive field. EffectiveCAN used a CNN based encoder with squeeze-and-excitation networks together with residual networks to aggregate the information across clinical texts. RNN-based models, which have also been widely used in the ICD coding task, are able to capture contextual information across input texts. Shi et al. (2017) proposed a character-aware long short-term memory (LSTM) recurrent network to learn the representations of the clinical texts. Xie and Xing (2018) used a tree-of-sequences LSTM architecture and adversarial learning to capture hierarchical relationships among ICD codes. Baumel et al. (2018) presented a hierarchical attention-bidirectional gated recurrent unit (HA-GRU) to label a document by identifying the sentences relevant for each ICD code. LAAT (Vu et al., 2020) used a bidirectional Long-Short Term Memory (BiLSTM) encoder and customized label-wise attention mechanism to learn label-specific vectors across clinical text fragments.

To tackle the hierarchical relationships among ICD codes, graph convolutional neural networks (GCNNs) (Kipf and Welling, 2017) can be employed. For instance, Rios and Kavuluru (2018) and Xie et al. (2019) leveraged GCNN to capture both the hierarchical relationships among ICD codes and the semantics of each code. HyperCore (Cao et al., 2020) considered both code hierarchy and code co-occurrence to learn code representations in the co-graph by exploiting the GCNN. Our work does not consider the ICD hierarchy because the parent-child hierarchy is shallow, i.e., the ICD hierarchy has only three levels. Also, the possibility of a parent code and a child code both being assigned to the same discharge summary is essentially zero in the MIMIC-III dataset, since the child code is a more specific description of the parent code.

Besides employing ICD code information, some other external knowledge has also been considered. For instance, Bai and Vucetic (2019) proposed a knowledge source integration (KSI) model that incorporates external knowledge from Wikipedia to calculate matching scores between a clinical note and disease-related Wikipedia documents, in order to obtain useful information for ICD predictions. Yuan et al. (2022) proposed a multiple synonym matching network (MSMN) to leverage synonyms of the ICD codes for better code representation learning. Yang et al. (2022) further incorporated a pre-trained language model with three domain-specific knowledge sources: code hierarchy, synonyms, and abbreviations to help the code classification.

2.2. Extreme Multi-label Text Classification

Extreme Multi-label Text Classification (XMTC) is designed to assign relevant labels to objects from an extremely large set of potential labels. Deep learning methods have been employed for XMTC tasks to learn semantic representations of text. For instance, XML-CNN (Liu et al., 2017) used a 1-dimensional convolutional network with different vertical filters and dynamic pooling to learn the text representations. The model utilizes convolutional filters of varying sizes (vertical filters), which operate across different n-gram sizes. Dynamic pooling is then added to allow the model to handle input texts of varying lengths by aggregating the feature maps produced by the convolutional layers into fixed-size representations. Furthermore, AttentionXML (You et al., 2019) employs a BiLSTM layer followed by an attention mechanism, a strategy designed to capture the most relevant text features for each label. Wang et al. (2022) introduced a knowledge-enhanced mask attention module in the KenMeSH framework, designed to refine the candidate label set by reducing its size. This innovative module leverages external knowledge to guide the attention mechanism, focusing it on the most relevant labels for a given text. By filtering out less pertinent labels, the model can concentrate on a more manageable subset of candidates, effectively improving the accuracy of the predictions.

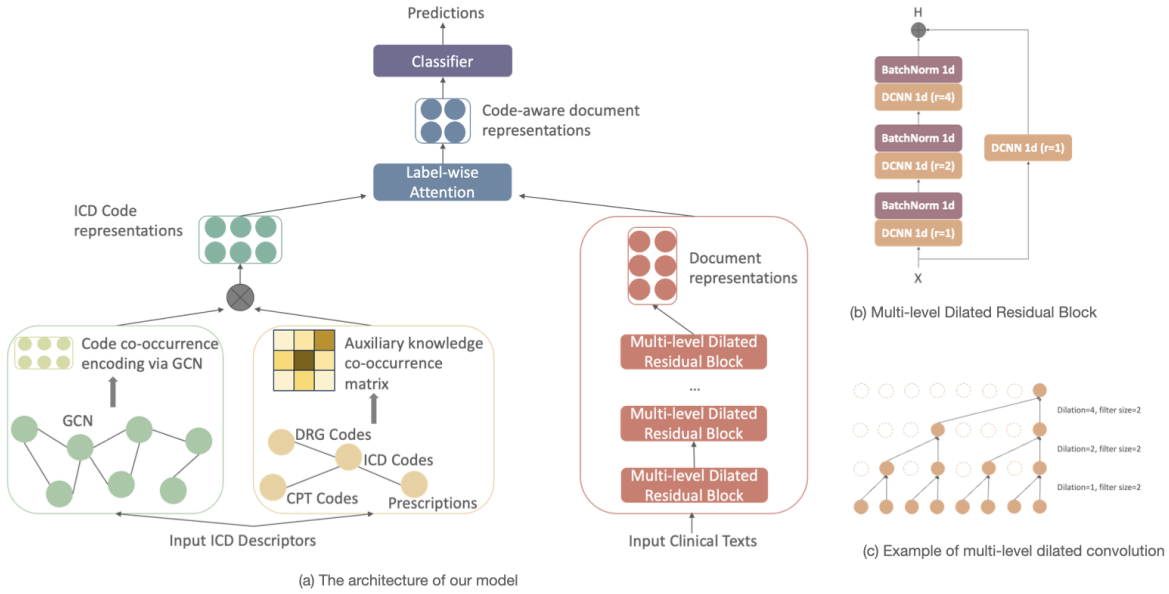

Figure 2: The architecture of our model. There are four main components in our method: a document encoder that contains multiple multi-level dilated residual blocks, a label encoder that includes label co-occurrence representation learned by GCN and an auxiliary knowledge mask, a label-wise attention layer and a classifier.

3. Method

We treat ICD indexing as an extreme multi-label text classification problem in which a set of medical records $\mathcal{X} = \{x_1, x_2, ..., x_N\}$ and a set of ICD codes $\mathcal{Y} = \{y_1, y_2, ..., y_L\}$ is given. The objective of multi-label classification is to learn L binary classifiers in which each classifier is to determine $y_j \in \{0, 1\}$ using the training set $\mathcal{D} = \{(x_i, Y_i)\}$, $Y_i \subset \mathcal{Y}$, $i = 1, ..., N$, where j is the j-th label in $\mathcal{Y}$, and N is the number of records in the set.

In this section, we present a neural architecture for ICD indexing shown in Figure 2(a). Our model is composed of a clinical text encoder that extracts the long-term dependencies and generates higher-level semantic representations for each clinical text, a label encoder that utilizes label co-occurrence relations and auxiliary knowledge to generate dynamic code presentations for each clinical note, a label-wise attention layer that produces the code-aware document representation, and a classifier that produces the final predictions of the ICD codes.

3.1. Clinical Text Encoder

3.1.1. Input Layer

Our model leverages a clinical record C as the input that consists of a sequence of words $\{w_1, w_2, ..., w_n\}$, where n is the sequence length. The embedding matrix $\tilde{E} \in \mathbb{R}^{d_e}$ is pre-trained using word2vec (Mikolov et al., 2013a) from the raw texts of the dataset, where d_e is the dimension of the word vectors. A word w_i in the medical record corresponds to an embedding vector e_i by looking up $\tilde{E}$. Therefore, the word embedding matrix for the input medical record is $E = \{e_1, e_2, ..., e_n\} \in \mathbb{R}^{n \times d_e}$, where n is the sequence length.

3.1.2. Multi-level Dilated Residual Block

To transform the clinical record into informative representations, we apply multiple multi-level dilated residual (Dilated-Res) convolutional blocks to generate representations of semantic units with different lengths. Each Dilated-Res block, as shown in Figure 2(b), is composed of two parallel modules that are referred to as the multi-level dilated convolutional module and the residual module.

We introduce a dilated convolutional layer (DCNN) to learn the high-level semantic representations of the input texts. The concept of dilated convolution has become popular in semantic segmentation in computer vision (Wang et al., 2018) and audio signal modeling (van den Oord et al., 2016), and it has been applied to natural language processing tasks such as neural machine translation (Kalchbrenner et al., 2016) and text classification (Lin et al., 2018). The main concept of DCNN expands the kernel by inserting holes between its consecutive elements in the filters, which aggregates multi-scale contextual information, such as words, phrases, and sentences. Inspired by Lin et al. (2018), we use a three-level DCNN with different dilation rates to generate high-level semantic representations of the input texts. We input the

word embeddings E of the medical records into the 1-dimensional convolution operator with kernel size K and dilation rates $[r_1, r_2, ..., r_m]$, where m is the number of layers of 1-dimensional convolutions. The dilated convolutional procedure can be formalized:

$$H^i = (E *_r f)(s) = \sum_{j=0}^{K-1} f(i) \cdot E_{s-r \cdot j} \quad (1)$$

where H^i is the output channel for layer $i = 1, ..., m$, $*$ denotes the convolution operation, r is the dilation rate, s is the element of the input sequence, K is the kernel size, and $s - r \cdot j$ refers to past time steps. We force the length of the output of the m-layer 1-dimensional DCNN to be the same as the input E, to keep the sequence length unchanged after the convolution, that is, $H^m \in \mathbb{R}^{n \times d_e}$. To achieve this, we set a padding size $p = \frac{r(K-1)}{2}$ with a stride of 1.

In addition to the multi-level dilated convolutional module, we also simultaneously transform input embedding E and add it to H^m as in the residual network (He et al., 2016), which reduces the gradient vanishing issue in the deep encoder structure. We use a 1-dimensional convolutional layer with dilation rate r_1 to transform the input embedding E into $\tilde{H}$. Then we add $\tilde{H}$ with H^m, the output from the multi-level dilated convolutional module, to form:

$$D = \sigma(H^m + \tilde{H}), \quad (2)$$

where $D \in \mathbb{R}^{n \times d_e}$ represents the final clinical text, and $\sigma(\cdot)$ denotes an activation function.

3.2. Label Encoder

3.2.1. Label Co-occurrence Encoding

The co-occurrence of disease codes in clinical texts is often observed when certain diseases are concurrent or have a causal relationship with each other. This means that the codes representing these related diseases tend to appear together in clinical text data. In order to capture the co-occurrence between disease codes in clinical texts, we construct a code co-occurrence graph. This is built using the code co-occurrence matrix, which serves as the adjacency matrix for the graph. To generate the code co-occurrence adjacency matrix, we model the correlation dependency between labels in terms of conditional probabilities. Specifically, we calculate the probability $P(L_j \mid L_i)$, which denotes the probability of occurrence of label L_j when label L_i appears. To facilitate graph construction, we binarize the correlation probability P. This entails converting the probability values into binary values, indicating whether a correlation exists between two labels. The operation can be written as:

$$A_{ij} = \begin{cases} 0, \text{if } P < \lambda \\ 1, \text{if } P \geq \lambda, \end{cases} \quad (3)$$

where A is the binary correlation matrix, and λ is the hyper-parameter threshold to filter the noise edges. In our experiment, $\lambda = 1$, which means that the two labels in each pair will always appear together. With $\lambda = 1$, we obtain 39,166 ICD pairs (which are also the number of edges in the graph). Decreasing the probability rate causes an exponential growth of the number of edges, which greatly increases the complexity of the graph. We hypothesized that the performance of the model would not benefit from a denser graph, since a larger number of edges would result in overfitting. This hypothesis can be explored in the future.

We employ a two-layer Graph Convolutional Network (GCN) to incorporate the co-occurrence relationships among labels. Specifically, we use the ICD full descriptors to generate a feature vector for each code. To calculate the feature vector for a specific code, we start by obtaining the word embeddings for each word in its descriptors, and then average the word embeddings to obtain a consolidated representation for the code:

$$v_i = \frac{1}{Z} \sum_{j=1}^{Z} w_j, i = 1, 2, ..., L, \quad (4)$$

where $v_i \in \mathbb{R}^{d_e}$, Z is the number of words in its descriptor, and L, the number of codes. The code vector set can be represented as $V = \{v_1, v_2, ..., v_L\}$. In our graph structure, each node represents an ICD code, and the edges between nodes represent code co-occurrence relationships. In each layer of the GCN, the node features are aggregated based on these edge types to generate new label features for the subsequent layer:

$$h^{l+1} = \sigma(A \cdot h^l \cdot W^l), \quad (5)$$

where h^l and $h^{l+1} \in \mathbb{R}^{L \times d_e}$ indicate the node representation of the l^{th} and $(l+1)^{th}$ layers, $h^0 = V$, A is the adjacency matrix of the label co-occurrence, W is the layer-specific weight matrix and $\sigma(\cdot)$ denotes an activation function. We denote the last layer representation as $H_{label} \in \mathbb{R}^{L \times d_e}$, which captures the code co-occurrence correlations.

3.2.2. Dynamic Auxiliary Knowledge Mask

ICD codes have a wide range of occurrence frequencies. Consequently, each ICD code has significantly more negative examples than positive ones. Inspired by Wang et al. (2022), to improve the classifier's performance, we dynamically generate a unique mask for each clinical note by integrating auxiliary knowledge in the EHR system.

The dynamic mask that is selected for each summary helps down-sample the negative examples and focus the classifier on candidate labels.

To generate the auxiliary knowledge masks, we consider three external knowledge sources: diagnosis-related group (DRG) codes, current procedural terminology (CPT) codes, and medications prescribed to patients. DRG codes are used to facilitate inpatient billing and reimbursement, and they categorize patients by their ICD codes and the cost associated with treatments. DRG codes are divided into medical DRGs (which don't reflect operating room procedures) and surgical DRGs. CPT codes are used to describe clinical procedures and services in healthcare. They provide a standardized way of documenting and billing for medical services. Such code terminologies play a crucial role in improving ICD code predictions. Prescribed drugs also appear to be highly informative in predicting ICD codes, since they are often the final step of the episode of care. As patients near the end of their treatment or care, the prescribed medications play a crucial role in managing their conditions. Consequently, these medications serve as strong indicators or signals of the underlying health conditions or diagnoses, making them valuable in predicting the corresponding ICD codes. We build an auxiliary knowledge-label co-occurrence matrix using conditional probabilities, i.e., $P(L_i \mid M_j)$, which denote the probabilities of occurrence of label L_i when auxiliary knowledge M_j appears.

$$P(L_i \mid M_j) = \frac{C_{L_i \cap M_j}}{C_{M_j}}, \tag{6}$$

where $C_{L_i \cap M_j}$ denotes the number of co-occurrences of L_i and M_j, and C_{M_j} is the number of occurrences of M_j in the training set. To avoid the noise of rare co-occurrences, a threshold τ filters noisy correlations. $\tilde{M}_j$ denotes the selected ICD set for auxiliary knowledge j.

$$\tilde{M}_j = \{L_k | P(L_k|M_j) > \tau,\ k = 1, ..., L\}. \tag{7}$$

We then join the ICD codes generated from the auxiliary knowledge co-occurrences for the DRG codes, CPT codes and prescribed drugs to form the final ICD mask set T:

$$T = \tilde{M}_{DRG} \cup \tilde{M}_{CPT} \cup \tilde{M}_{drug}. \tag{8}$$

Then we assign a value to each label in $\mathcal{Y}$ to form $T_{vec} \in [0, 1]^{\mathcal{Y}}$. We assign 1 if the label appears in T, and 0 otherwise. The label order of T_{vec} is the same as H_{label}. We then apply the mask to the label co-occurrence representation H_{label} to form:

$$H_{masked} = H_{label} \odot T_{vec}, \tag{9}$$

where $H_{masked} \in \mathbb{R}^{L \times d_e}$ indicates the masked code representation.

3.3. Label-wise Attention Layer

After encoding the clinical notes and their associated ICD codes, we obtain a clinical text representation denoted as D and a masked code representation denoted as H_{masked}. As we aim to assign multiple codes to each clinical note and recognize that different codes may be relevant to different sections of the document, we employ a code-wise attention mechanism. This mechanism allows the model to learn the relevant document representations specific to each code. To generate the code-wise attention vector, we use a matrix-vector product:

$$\alpha = \mathrm{softmax}(D \cdot H_{masked}) \tag{10}$$

Finally, we leverage the document representation D and corresponding code-wise attention vector α to generate the code-aware document representation:

$$C = \alpha \cdot D, \tag{11}$$

where $C \in \mathbb{R}^{L \times d_e}$.

3.4. Classifier

To perform classification, we compute the probability for each code by using a fully connected layer followed by a *sigmoid* transformation:

$$\hat{y} = \mathrm{sigmoid}(W \cdot C), \tag{12}$$

where $W \in \mathbb{R}^{d_e \times 1}$ indicates the weight matrix. Our model is trained using the multi-label binary cross-entropy loss:

$$L = \sum_{i=1}^{L} [-y_i \cdot \log(\hat{y}_i) - (1 - y_i) \cdot \log(1 - \hat{y}_i)], \tag{13}$$

where y_i is the ground truth of code i.

4. Experiments

4.1. Dataset and pre-processing

We use the MIMIC-III dataset (Johnson et al., 2016), which is the largest publicly available clinical dataset for text, and comprises hospital records associated with over 40,000 patients. We focus on the discharge summaries that are human expert-labeled with a set of ICD-9 codes. We follow the experimental setting of Mullenbach et al. (2018) to form MIMIC-III-full and MIMIC-III-top 50. To pre-process the clinical notes, we first remove all de-identified information and replace punctuation and atypical alphanumerical character combinations (e.g., 3a, 4kg) with white space. We then transform every token into its lowercase. The maximum length of a token sequence is 4,000 and any that exceed this length is truncated.

Hyper-parameters	Values
embedding size	**100**, 200
filter size	3, 5, **9**
prediction threshold	0.0005
dropout	**0.2**, 0.5
dilation rate	**[1, 2, 4]**, [2, 5, 9]
learning rate	**0.0001**, 0.0003, 0.0005
batch size	8, 16, **32**

Table 1: Hyper-parameter settings. Bold: the optimal values.

4.2. Evaluation Metrics and Implementation Details

Following previous work (Mullenbach et al., 2018), we evaluate our method using both macro and micro F1 and AUC metrics, as well as precision at K ($P@K$) that indicates the proportion of the correctly predicted labels in the top-K predictions.

We implement our model in PyTorch (Paszke et al., 2019) on a single NVIDIA A100 40G GPU. The training process completes in 5 hours, while inference requires only 30 minutes (about 0.5s per note). This performance underscores the computational efficiency of our method, indicating it is not only effective in handling complex tasks but also practical in terms of computational resources and time. We use word2vec (Mikolov et al., 2013b) to pre-train the word embeddings of dimension 100 on the pre-processed MIMIC-III texts. We use a three-level dilated convolution with dilation rate [1, 2, 4], and the filter size of the convolution is 9. We use the Adam optimizer and early stopping strategies. The learning rate is initialized to 0.0001, and the decay rate is 0.9 in every epoch. The gradient clip is applied to the maximum norm of 5. The batch size is 32. Table 1 shows our detailed hyper-parameter settings. We evaluate with 5 different random seeds for the model and report the average test results. Our code is available at https://github.com/xdwang0726/MIMIC-ICD-Classification.

5. Results and Discussions

To evaluate the effectiveness of our proposed model, we compare with existing state-of-the-art methods, which are given in Table 2. Each row represents the evaluation metrics for a specific method. The best score for each metric is highlighted. According to the reported results, our model demonstrates superior performance across the majority of evaluation metrics, with the exceptions of Macro-AUC and Macro-F1 on the MIMIC-III-full dataset. Under the Top-50 codes setting, our model performs better than the KEPTLongformer on all metrics and achieves state-of-the-art scores. These results confirm the effectiveness of leveraging auxiliary knowledge and label co-occurrence relations.

5.1. Ablation Studies

We aim to investigate the influence of various modules of our model, and we seek to understand how these modules contribute to the performance of the model in terms of both effectiveness and robustness. In order to conduct a fair comparison and isolate the effects of specific modules, we systematically remove certain modules from our model. Specifically, we conduct controlled experiments with three different settings: (a) examining the influence of different embedding methods by replacing built from scratch embeddings with pre-trained contextual embeddings, i.e., Clinical-Longformer (Li et al., 2023) and pre-trained biomedical context-free embeddings, i.e., BioWordVec (Zhang et al., 2019); (b) replacing the co-occurrence graph learning with a fully connected layer; (c) removing the auxiliary knowledge mask from our model. The experimental results are shown in Table 3.

Effectiveness of Embedding Methods As shown in Table 3, using pre-trained context-free word embeddings (BioWordVec) and pre-trained contextual embeddings (Clinical-Longformer) have negative impacts on the performance. This observation shows that although the use of pre-trained word embeddings has shown impressive performance across a wide range of natural language processing tasks, their performance on clinical datasets can sometimes be suboptimal. Understanding why this can occur will require further study.

Effectiveness of Learning Label Representations Table 3 shows the positive contribution of label representations learned by GCN. By using GCN, our model gains the ability to capture and leverage the relationships and dependencies between labels, leading to improvements in performance. This indicates that the incorporation of label co-occurrence information in a GCN enables the model to learn from the collective behaviour of labels, facilitating a more comprehensive understanding of the underlying label relationships.

Effectiveness of Involving Auxiliary Knowledge Mask We have three types of auxiliary knowledge involved to build the mask: DRG codes, CPT codes, and drug prescriptions. As reported in Table 3, performance drops when removing the auxiliary knowledge mask, suggesting that the auxiliary knowledge mask plays a crucial role in guiding the model's attention towards relevant information and aiding in the classification process. This result provides further evidence supporting the premise that the auxiliary knowledge mask effectively leverages external knowledge to mitigate the challenges posed by an extensive pool of potential ICD codes.

Models	MIMIC-III-full						MIMIC-III-top 50				
	AUC		F1		P@K		AUC		F1		P@5
	Macro	Micro	Macro	Micro	P@8	P@15	Macro	Micro	Macro	Micro	
CAML (Mullenbach et al., 2018)	0.895	0.986	0.088	0.539	0.709	0.561	0.875	0.909	0.532	0.614	0.609
DR-CAML (Mullenbach et al., 2018)	0.897	0.985	0.086	0.529	0.690	0.548	0.884	0.916	0.576	0.633	0.618
MultiResCNN (Li and Yu, 2020)	0.910	0.986	0.085	0.552	0.734	0.584	0.899	0.928	0.606	0.670	0.641
LAAT (Vu et al., 2020)	0.919	0.988	0.099	0.575	0.738	0.591	0.925	0.946	0.666	0.715	0.675
Joint-LAAT (Vu et al., 2020)	0.921	0.988	0.107	0.575	0.735	0.590	0.925	0.946	0.661	0.716	0.671
EffectiveCAN (Liu et al., 2021)	0.915	0.988	0.106	0.589	0.758	0.606	0.915	0.938	0.644	0.702	0.656
MSMN (Yuan et al., 2022)	**0.950**	0.992	0.103	0.584	0.752	0.599	0.928	0.947	0.683	0.725	0.680
KEPTLongformer (Yang et al., 2022)	-	-	**0.118**	0.599	0.771	0.615	0.926	0.947	0.689	0.728	0.672
Ours	0.948 ± 0.022	**0.994** ± 0.013	0.112 ± 0.027	**0.605** ± 0.021	**0.784** ± 0.022	**0.637** ± 0.011	**0.928** ± 0.014	**0.950** ± 0.018	**0.692** ± 0.016	**0.734** ± 0.012	**0.683** ± 0.023

Table 2: Comparison to previous methods across three main evaluation metrics MIMIC-III dataset. We report the mean ± standard deviation of each result. Bold: best scores in each column.

Methods	AUC		P@K	
	Macro	Micro	P@8	P@15
Full Model	**0.948**	**0.994**	**0.784**	**0.637**
embedded w/ Longformer	0.918	0.987	0.751	0.592
embedded w/ BioWordVec	0.923	0.989	0.765	0.609
w/o label feature	0.904	0.986	0.736	0.583
w/o masked attention	0.912	0.986	0.756	0.592

Table 3: Ablation experiment results. Bold: the optimal values.

By incorporating external knowledge through the auxiliary knowledge mask, the model gains the ability to narrow down and focus on relevant labels, thereby enhancing its efficiency and accuracy in the final prediction. To select the proper mask for each clinical note, one hyper-parameter is used: threshold τ of auxiliary knowledge-label co-occurrence. With $\tau = 0.005$, 99.22% of the gold-standard ICD codes are guaranteed to be in the mask, and the average number of codes in the mask is 1460 which is about $\frac{1}{6}$ of the complete set of codes.

5.2. Case Studies

We conduct case studies to qualitatively understand the effects of incorporating the label co-occurrence (as shown in Figure 3) and the auxiliary knowledge (as shown in Figure 4). For each patient, we show the discharge summary, ground truth ICD codes, label co-occurrence information / auxiliary knowledge information as well as the top-8 predicted ICD codes of the full model and ablated models. In Case 1, the ground truth ICD codes include "46.85 Dilation of intestine", a diagnosis not explicitly mentioned in the discharge summary. The observed label co-occurrence between "560.2 Volvulus" and "46.85 Dilation of intestine" serves as a robust indicator, effectively suggesting the presence of the "46.85 Dilation of intestine" diagnosis in the patient. Without the label co-occurrence signals, the ablated model makes a wrong prediction "789.07 Abdominal pain, generalized" that ignores the latent label information. In Case 2, the patient has been diagnosed with "331.0 Alzheimer's disease" with less explicit information in the discharge summary. Notably, the presence of "Donepezil" in the drug prescription, an element of the auxiliary knowledge, indicates that the patient is most likely to have Alzheimer's disease. The ablated model, lacking the auxiliary knowledge, mistakenly predicts "285.8 Other specified anemias". Case 1 and Case 2 exemplify the advantages of incorporating label co-occurrence and auxiliary knowledge, respectively.

Case 1: Effectiveness of Incorporating Label Co-occurrence	
Discharge Summary	History of Present Illness: [**Age over 90 **] y/o highly functional man with GERD presented with 4-5 days of abdominal pain and distention. He saw his PCP after having [**Name Initial (PRE) **] few episodes of loose stool. At that time his symptoms were attributed to resolving gastrointestinal infection, however hisabdominal pain progressed and he developed progressive distention. He then presented to [**Hospital1 **] [Location (un) 620] where a CT abdomen was completed which showed impressive large bowel obstruction without volvulus. He was then transferred to [Hospital1 18] for surgical evaluation.Brief Hospital Course:[**Age over 90 **] yo M w history of GERD, highly functional baseline, presents with colonic obstruction and was found to have recurrentvolvulus refractory to decompression. ACTIVE ISSUESColonic obstruction: The pt was found to have a sigmoid volvulus on CT. The pt was transfered to [**Hospital1 **] for further evaluation, andthe pt was offered operative treatment. The pt did not want a surgery, but agreed to attempt to have colonoscopic decompression. ...
Ground Truth ICD Codes	V66.7 Encounter for palliative care; V49.86 Do not resuscitate status; 560.2 Volvulus; 530.81 Esophageal reflux; 46.85 Dilation of intestine
Examples of Label Co-occurrence Information	560.2 Volvulus relates to 46.85 Dilation of intestine
Top-8 Predictions of Full Model	560.2 Volvulus; 46.85 Dilation of intestine; 789.07 Abdominal pain, generalized; 530.81 Esophageal reflux; 530.7 Gastroesophageal laceration-hemorrhage syndrome; V66.7 Encounter for palliative care; 560.9 Unspecified intestinal obstruction; V12.72 Personal history of colonic polyps
Top-8 Predictions of No Label Co-occurrence	560.2 Volvulus; 789.07 Abdominal pain, generalized; 530.81 Esophageal reflux; 530.7 Gastroesophageal laceration-hemorrhage syndrome; V66.7 Encounter for palliative care; 560.9 Unspecified intestinal obstruction; V45.72 Acquired absence of intestine (large) (small); V12.72 Personal history of colonic polyps

Figure 3: Case study on the effectiveness of incorporating label co-occurrence. Correctly predicted labels are marked in green and the incorrect ones are marked in red.

6. Conclusion

In this paper, we propose a novel auxiliary knowledge-induced medical code labelling frame-

Case 2: Effectiveness of Incorporating Auxiliary Knowledge	
Discharge Summary	Brief Hospital Course: [**Age over 90 **] year old male with history of prostate cancer, stage II CKD,DM2, CAD s/p CABG in [2112] and "quintuple bypass" in [2113] and history of frequent falls presented after unwitnessed fall with altered mental status/delirium from his Rehab center. # Acute on chronic metabolic encephalpathy:Patient experienced an unwitnessed fall at [**Hospital **] Rehab center.In the ED, patient was acutely agitated and was intubated in order to obtain CT head/spine, which revealed vasogenic edema concerning for malignancy. The CT-spine was negative and therewas no bleed seen on either scan. MRI head was performed without evidence of metastasis but with chronic ischemic changes and old area of infarcts. Patient was admitted initially to the MICU, and extubated one day later and was found to be agitated and confused. Due to persistent delirium, he was changed to seroquelwith improved response. The etiology for his delirium was likely acute on chronic delirium given multiple falls, head trauma, recurrent hospitalizations, and polypharmacy. Patient'soutpatient medications of lorazepam, zolpidem, ritalin and citalopram [Last Name (un) 8966] discontinued. Infectious workup was negative;though CXR demonstrated RLL opacity concerning for pneumonia vs. pneumonitis, patient remained afebrile without leukocytosis and antibiotics were not initiated....
Ground Truth ICD Codes	348.31 Metabolic encephalopathy; 348.5 Cerebral edema; 293.0 Delirium due to conditions classified elsewhere; 585.4 Chronic kidney disease, Stage IV (severe); 311 Depressive disorder, not elsewhere classified; 274.9 Gout, unspecified; 285.9 Anemia, unspecified; 530.81 Esophageal reflux; 272.0 Pure hypercholesterolemia; 250.92 Diabetes with unspecified complication, type II or unspecified type, uncontrolled; 403.10 Hypertensive chronic kidney disease, benign, with chronic kidney disease stage I through stage IV, or unspecified; 414.01 Coronary atherosclerosis of native coronary artery; V45.81 Aortocoronary bypass status; V10.46 Personal history of malignant neoplasm of prostate; V49.86; Do not resuscitate status; E888.9; Unspecified fall; E849.7; Accidents occurring in residential institution; 707.03; Pressure ulcer, lower back; 707.21; Pressure ulcer, stage I; 331.0; Alzheimer's disease; 294.10; Dementia in conditions classified elsewhere without behavioral disturbance; 96.04; Insertion of endotracheal tube
Examples of Auxiliary Knowledge information	1. Donepezil relates to 331.0 Alzheimer's disease 2. Quetiapine fumarate relates to 293.0 Delirium due to conditions classified elsewhere
Top-8 Predictions of Full Model	293.0 Delirium due to conditions classified elsewhere; 274.9 Gout, unspecified; 585.4 Chronic kidney disease, Stage IV (severe); 331.0; Alzheimer's disease; 294.10; Dementia in conditions classified elsewhere without behavioral disturbance; 401.9 Unspecified essential hypertension; V10.46 Personal history of malignant neoplasm of prostate; 285.8 Other specified anemias
Top-8 Predictions of No Auxiliary Knowledge	293.0 Delirium due to conditions classified elsewhere; 274.9 Gout, unspecified; 585.4 Chronic kidney disease, Stage IV (severe); 294.10 Dementia in conditions classified elsewhere without behavioral disturbance; 401.9 Unspecified essential hypertension; V10.46 Personal history of malignant neoplasm of prostate; 285.8 Other specified anemias; 414.0 Coronary atherosclerosis

Figure 4: Case study on the effectiveness of incorporating auxiliary knowledge. Correctly predicted labels are marked in green and the incorrect ones are marked in red.

work which uses multiple multi-level dilated residual blocks and jointly exploits label co-occurrence and auxiliary knowledge. Specifically, incorporating label co-occurrence relations and external knowledge through the auxiliary knowledge mask serves as a valuable mechanism for addressing the inherent complexity and size of the label space, ultimately leading to improved performance and more effective utilization of the model's resources. Moreover, to deal with the length of the clinical texts, the multi-level dilated residual block helps capture and understand long dependencies. Experimental results demonstrate that our proposed model outperforms the baseline models. We are interested in integrating more external knowledge in the future, such as the Unified Medical Language System (UMLS), to seek further improvements.

7. Limitations

Our work is limited to evaluate the MIMIC-III-full and MIMIC-III-top 50, which are mostly focused on common diseases (i.e., the most frequent ICD codes). It is not possible to define rare diseases simply from the distribution of ICD codes in the dataset since rare ICD codes do not necessarily indicate the presence of rare diseases exclusively. This limits evaluation on diseases that are rare *a priori*. A list of rare diseases proposed by domain experts for more specific medical tasks would be helpful to explore more focused use cases.

Our auxiliary knowledge masks are limited by external knowledge including DRG codes, CPT codes, and drug prescriptions. Other knowledge sources, including disease-symptom, disease-lab relations, for example, could potentially be useful for the auto ICD coding task.

Acknowledgements

We would like to thank all reviewers for their comments, which helped improve this paper considerably. Computational resources used in preparing this research were provided, in part, by Compute Ontario[9], Digital Research Alliance of Canada[10], the Province of Ontario, the Government of Canada through CIFAR, and companies sponsoring the Vector Institute[11]. This research is partially funded by The Natural Sciences and Engineering Research Council of Canada (NSERC) through a Discovery Grant to R. E. Mercer. F. Rudzicz is supported by a CIFAR Chair in AI.

8. Bibliographical References

Tian Bai and Slobodan Vucetic. 2019. Improving medical code prediction from clinical text via incorporating online knowledge sources. *The World Wide Web Conference*, pages 72–82.

Tal Baumel, Jumana Nassour-Kassis, Michael Elhadad, and Noémie Elhadad. 2018. Multi-label classification of patient notes a case study on ICD code assignment. In *The Workshops of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 409–416.

[9] https://www.computeontario.ca
[10] https://ccdb.alliancecan.ca
[11] https://www.vectorinstitute.ai/partners

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 (Volume 1: Long Papers)*, pages 3105–3114.

Koby Crammer, Mark Dredze, Kuzman Ganchev, Partha Pratim Talukdar, and Steven Carroll. 2007. Automatic code assignment to medical text. In *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 129–136.

Luciano R. S. de Lima, Alberto H. F. Laender, and Berthier A. Ribeiro-Neto. 1998. A hierarchical approach to the automatic categorization of medical documents. In *Proceedings of the Seventh International Conference on Information and Knowledge Management*, CIKM '98, pages 132–139.

Richárd Farkas and György Szarvas. 2008. Automatic construction of rule-based ICD-9-CM coding systems. *BMC Bioinformatics*, 9:S10 – S10.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Shaoxiong Ji, Erik Cambria, and Pekka Marttinen. 2020. Dilated convolutional attention network for medical code assignment from clinical text. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 73–78.

Nal Kalchbrenner, Lasse Espeholt, Karen Simonyan, Aaron van den Oord, Alex Graves, and Koray Kavukcuoglu. 2016. Neural machine translation in linear time. *arXiv*, 1610.10099v2.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *Proceedings of the 5th International Conference on Learning Representations*.

Leah S. Larkey and W. Bruce Croft. 1996. Combining classifiers in text categorization. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 289–297.

Fei Li and Hong Yu. 2020. ICD coding from clinical text using multi-filter residual convolutional neural network. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*, pages 8180–8187.

Yikuan Li, Ramsey M Wehbe, Faraz S Ahmad, Hanyin Wang, and Yuan Luo. 2023. A comparative study of pretrained language models for long clinical text. *Journal of the American Medical Informatics Association*, 30(2):340–347.

Junyang Lin, Qi Su, Pengcheng Yang, Shuming Ma, and Xu Sun. 2018. Semantic-unit-based dilated convolution for multi-label text classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4554–4564.

Lucian Vlad Lita, Shipeng Yu, Stefan Niculescu, and Jinbo Bi. 2008. Large scale diagnostic code classification for medical patient records. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-II*, pages 877–882.

Jingzhou Liu, Wei-Cheng Chang, Yuexin Wu, and Yiming Yang. 2017. Deep learning for extreme multi-label text classification. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 115–124.

Yang Liu, Hua Cheng, Russell Klopfer, Matthew R. Gormley, and Thomas Schaaf. 2021. Effective convolutional attention network for multi-label clinical document classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5941–5953.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pages 3111–3119.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26.

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)*, pages 1101–1111.

Kimberly J. O'Malley, Karon F. Cook, Matt D. Price, Kimberly Raiford Wildes, John F. Hurdle, and Carol M. Ashton. 2005. Measuring diagnoses:

ICD code accuracy. *Health Services Research*, 40(5 II):1620–1639.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32, pages 8024–8035.

Anthony Rios and Ramakanth Kavuluru. 2018. Few-shot and zero-shot multi-label learning for structured label spaces. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3132–3142.

Haoran Shi, Pengtao Xie, Zhiting Hu, Ming Zhang, and Eric P. Xing. 2017. Towards automated ICD coding using deep learning. *ArXiv*, abs/1711.04075.

Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. WaveNet: A generative model for raw audio. In *Proceedings of the 9th ISCA Workshop on Speech Synthesis Workshop (SSW 9)*, page 125.

Thanh Vu, Dat Quoc Nguyen, and Anthony Nguyen. 2020. A label attention model for ICD coding from clinical text. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3335–3341. Main track.

Panqu Wang, Pengfei Chen, Ye Yuan, Ding Liu, Zehua Huang, Xiaodi Hou, and G. Cottrell. 2018. Understanding convolution for semantic segmentation. *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1451–1460.

Xindi Wang, Robert Mercer, and Frank Rudzicz. 2022. KenMeSH: Knowledge-enhanced end-to-end biomedical text labelling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2941–2951.

Pengtao Xie and Eric Xing. 2018. A neural architecture for automated ICD coding. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1066–1076.

Xiancheng Xie, Yun Xiong, Philip S. Yu, and Yangyong Zhu. 2019. EHR coding with multi-scale feature attention and structured knowledge graph propagation. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, CIKM '19, pages 649–658.

Zhichao Yang, Shufan Wang, Bhanu Pratap Singh Rawat, Avijit Mitra, and Hong Yu. 2022. Knowledge injected prompt based fine-tuning for multi-label few-shot ICD coding. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1767–1781.

Ronghui You, Zihan Zhang, Ziye Wang, Suyang Dai, Hiroshi Mamitsuka, and Shanfeng Zhu. 2019. AttentionXML: Label tree-based attention-aware deep model for high-performance extreme multi-label text classification. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 5820–5830.

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.

Yijia Zhang, Qingyu Chen, Zhihao Yang, Hongfei Lin, and Zhiyong Lu. 2019. BioWordVec, improving biomedical word embeddings with subword information and MeSH. *Scientific Data*, 6.

9. Language Resource References

Johnson, Alistair EW and Pollard, Tom J and Shen, Lu and Lehman, Li-wei H and Feng, Mengling and Ghassemi, Mohammad and Moody, Benjamin and Szolovits, Peter and Anthony Celi, Leo and Mark, Roger G. 2016. *MIMIC-III, a freely accessible critical care database*. Nature Publishing Group.