Modeling Diagnostic Label Correlation for Automatic ICD Coding

Shang-Chi Tsai[*] **Chao-Wei Huang**[*] **Yun-Nung Chen**
National Taiwan University, Taipei, Taiwan
{d08922014,f07922069}@csie.ntu.edu.tw y.v.chen@ieee.org

Abstract

Given the clinical notes written in electronic health records (EHRs), it is challenging to predict the diagnostic codes which is formulated as a multi-label classification task. The large set of labels, the hierarchical dependency, and the imbalanced data make this prediction task extremely hard. Most existing work built a binary prediction for each label independently, ignoring the dependencies between labels. To address this problem, we propose a two-stage framework to improve automatic ICD coding by capturing the label correlation. Specifically, we train a label set distribution estimator to rescore the probability of each label set candidate generated by a base predictor. This paper is the first attempt at learning the label set distribution as a reranking module for medical code prediction. In the experiments, our proposed framework is able to improve upon best-performing predictors on the benchmark MIMIC datasets. [1]

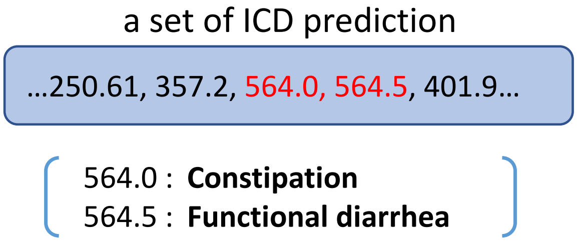

Figure 1: An example of conflicting predictions. While these two codes share the same root in the hierarchical ICD structure and are semantically similar, they are unlikely to appear together.

1 Introduction

Clinical notes from electronic health records (EHRs) are free-from text generated by clinicians during patient visits. The associated diagnostic codes from the International Classification of Diseases (ICD) represent diagnostic and procedural information of the visit. The ICD codes provide an standardized and systematic way to encode information and has several potential use cases (Choi et al., 2016).

Considering that manual ICD coding has been shown to be labor-intensive (O'malley et al., 2005), several approaches for automatic ICD coding has been proposed and investigated by the research community (Perotte et al., 2013; Kavuluru et al., 2015). With recent introduction of deep neural networks, the performance of automatic ICD coding has been improved significantly (Choi et al., 2016; Shi et al., 2017; Mullenbach et al., 2018; Baumel et al., 2018; Xie and Xing, 2018; Li and Yu, 2020; Vu et al., 2020; Cao et al., 2020). Prior work on neural models mostly treated the task of automatic ICD coding as a multi-label classification problem. These models mostly employ a shared text encoder, and build one binary classifier for each label on top of the encoder. This architecture along side with binary cross-entropy loss make the prediction of each label independent of each other, which might lead to incomplete or conflicting predictions. An example of such error is shown in Figure 1. This issue is especially problematic in ICD code prediction, since the ICD codes share a hierarchical structure. That is, the low-level codes are more specific, and the high-level ones are more general. In some cases, the low-level codes under the same high-level category are more likely to be jointly diagnosed. Rare codes also have more opportunity to be considered from the frequent codes in the same high-level class. Prior work considered the hierarchical dependencies between ICD codes by using hierarchical SVM (Perotte et al., 2013) or by introducing new loss terms to leverage the ICD structure (Tsai et al., 2019). However, they borrowed the dependency from domain experts and did not consider

[*] Equal contribution.

[1] The source code of this project is available at https://github.com/MiuLab/ICD-Correlation.

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 4043–4052
June 6–11, 2021. ©2021 Association for Computational Linguistics

the label correlation in the data.

Inspired by the success of reranking techniques on automatic speech recognition (Ostendorf et al., 1991) and dependency parsing (Zhu et al., 2015; Sangati et al., 2009), we propose a two-stage reranking framework for ICD code prediction, which captures the label correlation without any expert knowledge. In the first stage, we use a base predictor to generate possible label set candidates. In the second stage, a label set reranker is employed to rerank the candidates. We design two rerankers to help to capture the correlation between labels. The experimental results show that our proposed framework consistently improves the results of different base predictors on the benchmark MIMIC datasets (Saeed et al., 2011; Johnson et al., 2016). The results also show that the proposed framework is model agnostic, i.e., we can use any base predictor in the first stage.

Data privacy is a major difficulty for medical NLP research. The personal health information (PHI) which explains a patient's ailments, treatments and outcomes is highly sensitive, making it hard to distribute due to privacy concerns. In addition, EHRs across multiple hospitals or languages may contain different writing style, typos and abbreviations. It is labor-demanding to train separate models for each hospital with their in-house data only. The advantage of our proposed two-stage framework is that we can train base predictors with in-house data, while enjoying the universality of ICD codes to train a reranker on ICD labels from different sources. This reranker is able to generally work with various base predictor trained on health records from any specific hospital.

The contributions of this paper are 3-fold:

- This paper is the first attempt to improve multi-label classification with a reranking method for automatic ICD coding.
- The experiments show that the proposed approaches are capable of improving best-performing base predictors on the benchmark datasets MIMIC-2 and MIMIC-3, demonstrating its great generalizability.
- The proposed framework has the great potential of benefiting from extra ICD labels, reducing the demand of paired training data towards scalability in the medical NLP field.

2 Related Work

This paper focuses on multi-label medical code prediction; hence, We briefly describe the related background about medical code prediction and multi-label classification.

2.1 Medical Code Prediction

ICD code prediction is a challenging task in the medical domain. It has been studied since 1998 (de Lima et al., 1998) and several recent work attempted to approach this task with neural models. Choi et al. (2016) and Baumel et al. (2018) used recurrent neural networks (RNN) to encode the EHR data for predicting diagnostic results. Li and Yu (2020) recently utilized a multi-filter convolutional layer and a residual layer to improve the performance of ICD prediction. On the other hand, several work tried to integrate external medical knowledge into this task. In order to leverage the information of definition of each ICD code, RNN and CNN were adopted to encode the diagnostic descriptions of ICD codes for better prediction via attention mechanism (Shi et al., 2017; Mullenbach et al., 2018). Moreover, the prior work tried to consider the hierarchical structure of ICD codes (Xie and Xing, 2018), which proposed a tree-of-sequences LSTM to simultaneously capture the hierarchical relationship among codes and the semantics of each code. Also, Tsai et al. (2019) introduced various ways of leveraging the hierarchical knowledge of ICD by adding refined loss functions. Recently, Cao et al. (2020) proposed to train ICD code embeddings in hyperbolic space to model the hierarchical structure. Additionally, they used graph neural network to capture the code co-occurrences.

2.2 Multi-Label Classification

Multi-label classification problems are of broad interest to the machine learning community. The goal is to predict a subset of labels associated with a given object. One simple solution to a multi-label classification problem is to transform the problem into n binary classification problems, where n denotes the number of labels.

This approach makes an assumption that the predictions of each label are independent. However, in practice, the labels are usually dependent, making these predictors produce undesired predictions.

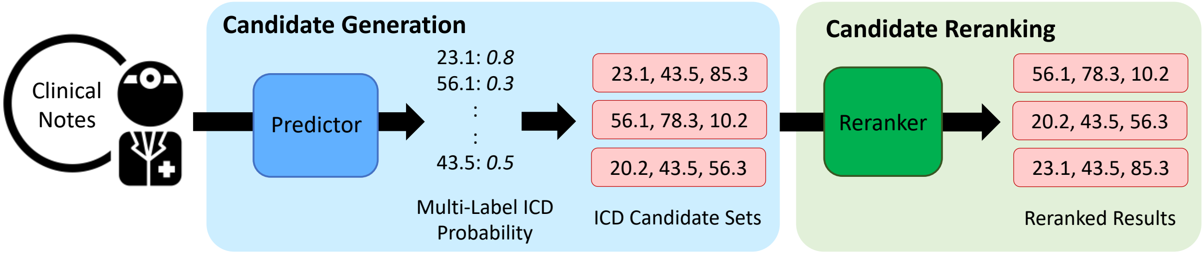

Figure 2: Illustration of our proposed framework.

There are numerous methods developed to alleviate this issue. Read et al. (2009) proposed classifier chains (CC), which introduce sequential dependency between predictions by adding the decision of one classifier to the input of the next classifier. Cheng et al. (2010) generalized CC to probabilistic classifier chains (PCC), where the proposed approach estimates the joint probability of labels and provides a proper interpretation of CC. Nevertheless, the recurrence between classifiers makes these methods less efficient and not applicable to tasks with large amount of labels.

Another line of research has leveraged the label dependencies that are known beforehand. Deng et al. (2014) used label relation graphs for object classification. Tsai et al. (2019) utilized the hierarchical structure of ICD codes to improve the ICD code prediction. These methods relied on known structures of the labels, which may not be easily accessible and less general.

Some prior work tried to learn label correlation and dependencies directly from the dataset. Zhang et al. (2018) introduced residual blocks to capture label correlation. This method requires paired training data, while our framework can learn from ICD codes only.

The concept of retrieve-and-rerank has been widely used in automatic speech recognition (Ostendorf et al., 1991), natural language processing (Collins and Koo, 2005) and machine translation (Shen et al., 2004). Li et al. (2019) proposed to rerank the possible predictions generated by a base predictor with a calibrator. This method is conceptually similar to our framework, where we both follow the retrieve-and-rerank procedure. The main difference between is that they leveraged an extra dataset for training the calibrator, while we train a distribution estimator on the same dataset as our base predictor.

3 Proposed Framework

The task of ICD code prediction is usually framed as a multi-label classification problem (Kavuluru et al., 2015; Mullenbach et al., 2018). Given a clinical record $\mathbf{x}$ in EHR, the goal is to predict a set of ICD codes $\mathbf{y} \subseteq \mathcal{Y}$, where $\mathcal{Y}$ denotes the set of all possible codes. This subset is typically represented as a binary vector $\mathbf{y} \in \{0,1\}^{|\mathcal{Y}|}$, where each bit y_i indicates the presence or absence of the corresponding label.

The proposed framework is illustrated in Figure 2 and consists of two stages:

1. **Label set candidate generation** provides multiple ICD set candidates through a base binary predictor, which is detailed in Section 3.1.
2. **Label set candidate reranking** estimates the probability by leveraging the label correlation for reranking the candidates, which is detailed in Section 3.2.

3.1 Candidate Generation

In the first stage of the framework, we employ a base predictor to perform probabilistic prediction for all labels, and we use the predicted probabilities to generate top-k most probable label sets. More formally, given a clinical note $\mathbf{x}$, we perform a base predictor and obtain the prediction for all labels:

$$P_{\text{base}}(y_i = 1 \mid \mathbf{x}, \theta_{\text{base}}), \quad i = 1, 2, \cdots, |\mathcal{Y}|,$$

where θ_{base} denotes the parameters of the base predictor. The predicted results are used to generate top-k probable sets, i.e., $\hat{\mathbf{y}} \subseteq \mathcal{Y}$ with top-k highest probability prediction:

$$P_{\text{base}}(\hat{\mathbf{y}} \mid \mathbf{x}, \theta_{\text{base}}) = \prod_{i=1}^{|\mathcal{Y}|} P_{\text{base}}(y_i = \hat{\mathbf{y}}_i \mid \mathbf{x}, \theta_{\text{base}}).$$

Although there are $2^{|\mathcal{Y}|}$ possible subsets, the top-k sets can be efficiently generated with dynamic

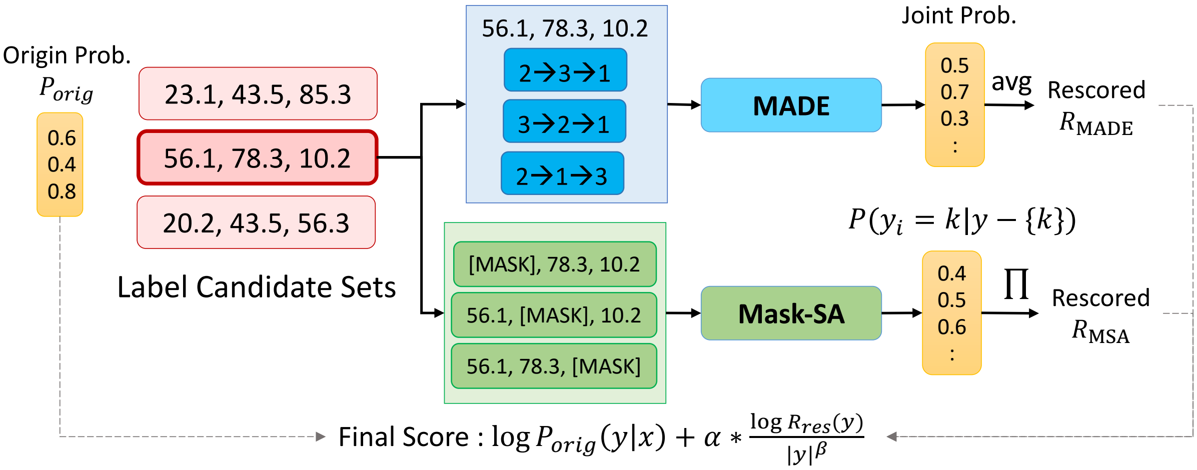

Figure 3: Illustration of the proposed reranking process.

programming as described in the prior work (Li et al., 2016).

3.2 Candidate Reranking

One drawback of the base predictor is the assumption about independent labels. To address this issue, in the second stage of the framework, we introduce a label set reranker to rerank the label set candidates generated in the previous stage. The reranker is designed to capture correlation and co-occurrence between labels. Given a label set candidate $\hat{\mathbf{y}}$, a reranker should be able to provide a reranking score $R(\hat{\mathbf{y}})$, where higher score indicates that the label set is more probable to appear. Similar to prior work (Zhu et al., 2015), We rerank the candidates according to their new scores defined as

$$\log P_{\text{base}}(\hat{\mathbf{y}} \mid \mathbf{x}, \theta_{base}) + \alpha \cdot R(\hat{\mathbf{y}}),$$

where α is a hyperparameter. We use the label set with the highest score after reranking as the final prediction. Note that that reranking is done on label sets, not individual labels.

We employ two rerankers and describe them in the following subsections. Note that we do not restrict the design of rerankers to those we proposed; one can design their own reranker and plug it into the proposed framework. Our reranking framework is illustrated in Figure 3.

3.2.1 MADE Reranker

One intuitive way to assign scores to $\hat{\mathbf{y}}$ is using the joint probability $P(\hat{\mathbf{y}})$. Higher joint probability indicates that the label set is more probable to appear, which aligns with our requirement to rerankers. However, the joint probability $P(\hat{\mathbf{y}})$ is often intractable. Therefore, we can only make an estimation with a density estimator.

Here we employ a masked autoencoder (MADE) (Germain et al., 2015) as the density estimator. MADE estimates the joint probability of a binary vector $P(\hat{\mathbf{y}})$ by decomposing it in an autoregressive fashion with a random ordering

$$P_{\text{MADE}}(\hat{\mathbf{y}}) = \prod_{i=1}^{|\mathcal{Y}|} P_{\text{MADE}}(y_i = \hat{\mathbf{y}}_i \mid \hat{\mathbf{y}}_{o<i}, \theta_{\text{MADE}}),$$

where o denotes a random permutation of $\{1, 2, \cdots, |\mathcal{Y}|\}$, $o(i)$ denotes the new ordering of i, $\hat{\mathbf{y}}_{o<i} = \{\hat{\mathbf{y}}_j \mid o(j) < o(i)\}$ denotes the set of all elements precede $\hat{\mathbf{y}}_i$ in the new ordering, and θ_{MADE} denotes parameters of the MADE model. MADE introduces a sequential dependency between labels. It enforces this dependency by masking certain connections in the multi-layer perceptron, making the output of y_i only depends on $\hat{\mathbf{y}}_{o<i}$.

The MADE model is trained with the labels from the training set of the ICD code prediction task. We use stochastic gradient descent to optimize the parameters θ_{MADE}, and the training objective is to minimize the binary cross-entropy loss $\mathcal{L}(\hat{\mathbf{y}})$:

$$-\frac{1}{|\hat{\mathbf{y}}|}\sum_{i=1}^{|\hat{\mathbf{y}}|}\Big(\hat{\mathbf{y}}_i \log P_{\text{MADE}}(y_i = 1 \mid \hat{\mathbf{y}}_{o<i}, \theta_{\text{MADE}})$$
$$+(1-\hat{\mathbf{y}}_i)\log P_{\text{MADE}}(y_i = 0 \mid \hat{\mathbf{y}}_{o<i}, \theta_{\text{MADE}})\Big).$$

Because we do not know which ordering performs the best, we can sample n different orderings and use the ensemble of these orderings to improve estimation:

$$P_{\text{MADE}}(\hat{\mathbf{y}}) = \frac{1}{n}\sum_{j=i}^{n}\prod_{i=1}^{|\mathcal{Y}|} P_{\text{MADE}}(y_i = \hat{\mathbf{y}}_i \mid \hat{\mathbf{y}}_{o_j<i}, \theta_{\text{MADE}}).$$

The illustration can be found in the blue box of Figure 3.

Given a label set candidate $\hat{\mathbf{y}}$, we define the score $R_{\text{MADE}}(\hat{\mathbf{y}})$ as

$$R_{\text{MADE}}(\hat{\mathbf{y}}) = \frac{\log P_{\text{MADE}}(\hat{\mathbf{y}})}{|\hat{\mathbf{y}}|^{\beta}},$$

where $|\hat{\mathbf{y}}|$ denotes the size of the subset $\hat{\mathbf{y}}$, and β is a hyperparameter. $|\hat{\mathbf{y}}|^{\beta}$ serves as a length penalty similar to the one used in sequence generation (Wu et al., 2016). We find that this length penalty is crucial to the reranker, and without it the score would favor subsets with smaller size.

3.2.2 Masked Self-Attention Reranker (Mask-SA)

As described in the previous subsection, MADE uses a sequential factorization to estimate the joint probability of a label set. This formulation forces the prediction of y_i to only condition on a subset of inputs $\hat{\mathbf{y}}_{o<i}$. With this restriction, the MADE model may fail to capture some crucial dependencies.

Inspired by the masked language modeling objective (Devlin et al., 2019), we propose a masked self-attention reranker (Mask-SA). Mask-SA takes as input a set of predicted labels $\hat{\mathbf{y}} \subseteq \mathcal{Y}$, which is the set representation of the predicted labels. It employs a cloze-style prediction method, where we mask one input at a time and ask the model to predict the masked input. The advantage of this prediction method is that the output is conditioned on all inputs except for itself, which solves the restriction of the MADE model. This procedure is very similar to a denoising autoencoder (Vincent et al., 2008). The illustration can be found in the green box of Figure 3.

The Transformer architecture (Vaswani et al., 2017) has been shown to be efficient and effective in language modeling (Dai et al., 2019). We use it as the architecture of the Mask-SA model, with a slight modification where we remove the positional encodings due to the fact that the predicted ICD codes have no sequential order.

More formally, Mask-SA estimates a distribution over the label vocabulary for the masked input given all other elements in the set $P_{\text{MSA}}(\hat{\mathbf{y}}_i \mid \hat{\mathbf{y}} - \{\hat{\mathbf{y}}_i\}, \theta_{MSA})$, where θ_{MSA} denotes the parameters of the Mask-SA model. θ_{MSA} can be optimized with stochastic gradient descent to minimize the cross-entropy loss function. Given a label set candidate $\hat{\mathbf{y}}$, we compute the score $R_{\text{MSA}}(\hat{\mathbf{y}})$ as

$$R_{\text{MSA}}(\hat{\mathbf{y}}) = \frac{\log \prod_{i=1}^{|\hat{\mathbf{y}}|} P_{\text{MSA}}(\hat{\mathbf{y}}_i \mid \hat{\mathbf{y}} - \{\hat{\mathbf{y}}_i\}, \theta_{\text{MSA}})}{|\hat{\mathbf{y}}|^{\beta}},$$

where β is a hyperparameter. Note that in this formulation, the product of the conditional probabilities is not an exact estimation of the joint probability of $\hat{\mathbf{y}}$, but an analogy to the factorization made in the MADE model.

4 Experiments

In order to evaluate the effectiveness of our proposed framework, we conduct experiments on two benchmark datasets. We employ three different base predictors to validate the generalizability of the proposed framework.

4.1 Setup

We evaluate our model on two benchmark datasets for ICD code prediction.

- **MIMIC-2** Following the prior work (Mullenbach et al., 2018; Li and Yu, 2020), we evaluate our method on the MIMIC-2 dataset. We follow their setting, where 20,533 summaries are used for training, and 2,282 summaries are used for testing. There are 5,031 labels in the dataset.
- **MIMIC-3** The Medical Information Mart for Intensive Care III (MIMIC-3) (Johnson et al., 2016) dataset is a benchmark dataset which contains text and structured records from a hospital ICU. We use the same setting as the prior work (Mullenbach et al., 2018), where there are 47,724 discharge summaries for training, with 1,632 summaries and 3,372 summaries for validation and testing, respectively. There are 8,922 labels in the dataset. We also follow the setting in (Shi et al., 2017) where only the top-50 most frequent codes are considered. This setting has 8,067 summaries for training, 1,574 summaries for validation, and 1,730 summaries for testing.

We follow the preprocessing steps described in Mullenbach et al. (2018) with the provided

Model	Top-50 Dev		Top-50 Test		All Dev		All Test	
	MacroF	MicroF	MacroF	MicroF	MacroF	MicroF	MacroF	MicroF
CAML (2018)	54.03	61.76	53.46	61.41	7.70	54.29	8.84	53.87
+ MADE	**56.91**†	**62.31**†	**56.64**†	**62.40**†	7.79†	**54.70**†	9.11†	**54.29**†
+ Mask-SA	56.57†	62.14†	56.33†	62.26†	**8.06**†	54.53†	**9.27**†	54.09†
MultiResCNN (2020)	60.77	66.98	60.84	66.78	7.38	56.05	8.50	55.31
+ MADE	**62.10**†	67.13†	62.00†	67.13†	7.75†	57.08†	8.81†	56.21†
+ Mask-SA	61.52†	**67.15**†	**62.06**†	**67.22**†	**7.97**†	**57.12**†	**9.28**†	**56.49**†
LAAT (2020)	65.53	**70.38**	65.05	70.01	7.48	57.18	8.74	56.56
+ MADE	65.92†	70.34	65.29†	70.13†	7.92†	**57.80**†	9.16†	**57.26**†
+ Mask-SA	**66.10**†	70.30	**65.44**†	**70.15**†	**8.08**†	57.76†	**9.41**†	57.23†

Table 1: Results on the MIMIC-3 (%). † indicates the improvement achieved by the proposed rescoring framework. The best scores for each base predictor are marked in bold.

Model	Macro F1	Micro F1
CAML	4.90	44.79
+ MADE	5.31†	**46.11**†
+ Mask-SA	**5.55**†	46.08†
MultiResCNN	5.06	45.89
+ MADE	5.50†	47.49†
+ Mask-SA	**5.88**†	**47.55**†
LAAT	6.41	47.54
+ MADE	7.23†	**49.15**†
+ Mask-SA	**7.42**†	49.05†

Table 2: Results on the MIMIC-2 test set using all codes (%). † indicates the improvement achieved by the proposed framework. The best scores are marked in bold.

scripts [2]. All discharge summaries are truncated to a maximum length of 2,500 tokens.

4.2 Base Predictors

In order to validate the generalizability of our proposed framework, we employ three different base predictors that are proposed in prior work:

- **CAML** Convolutional attention for multi-label classification (CAML) is a method proposed in (Mullenbach et al., 2018). CAML aimed at improving ICD code prediction by exploiting the textual description of codes with attention mechanism (Bahdanau et al., 2014).
- **MultiResCNN** The multi-filter residual convolutional neural network (MultiResCNN) improved the design of CAML with multiple convolution filters and residual connections (Li and Yu, 2020).
- **LAAT** Vu et al. (2020) proposed a label attention model which augments the label attention mechanism in CAML with additional transformations. It achieved state-of-the-art performance on MIMIC-2 and MIMIC-3 datasets.

4.3 Training and Evaluation Details

We train our rerankers with the label sets in the training set for 30 epochs. Adam is chosen as the optimizer with a learning rate of $2e-5$. The batch-size is set to 64. The MADE reranker has one hidden layer with 500 neurons, and we find that using $n = 10$ different orderings provides good estimation without using too much computation power. The Mask-SA reranker employs the transformer architecture with 6 self-attention layers, each with 8 attention heads. The hidden size is set to 256. For each pair of the base predictor and the reranker, we apply a grid search over possible values of α and β on the validation set to find the best-performing hyperparameters, and we use them to perform evaluation on the test set. During reranking, we generate top-50 label set candidates to rerank. Note that our approach is to rerank the label set candidates instead of modifying the predicted probabilities from the base predictors. Therefore, common metrics considering the predicted probabilities of each label, such as Precision@K and AUC, are not suitable for our evaluation. Instead, we evaluate our methods with two metrics, macro F1 and micro F1.

[2] https://github.com/jamesmullenbach/caml-mimic

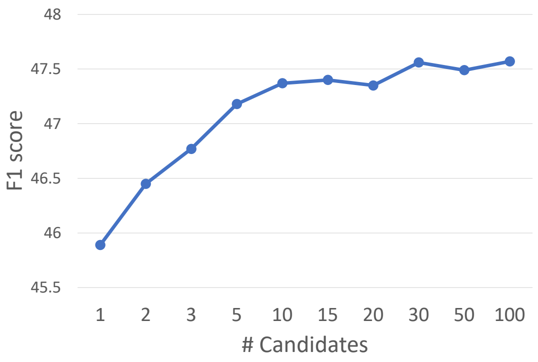

Figure 4: F1 scores (%) with different number of candidates.

4.4 Results

The results on the MIMIC-3 and MIMIC-2 datasets are shown in Table 1 and Table 2 respecitively. All results are obtained by averaging the scores of 5 different training runs. We list the results before reranking in the first row of each base predictor.

In all scenarios, our proposed reranking framework achieves consistent improvement over the base predictors except for LAAT, where the macro F-score on MIMIC-3 top-50 dev set slightly decreased. The relative improvement in macro F-score for all-code settings are more significant, ranging from 1% to 16%. Considering that the all-code setting is much more challenging and macro F-score is difficult to improve due to the data imbalance issue, the achieved improvement demonstrates the great potential of the proposed framework for better practicality. The MADE and Mask-SA reranker are both effective for the purpose of reranking. Their gains are similar across different settings and datasets. We believe that this trend is reasonable given that their formulations are similar, i.e., they both calculate score as product of conditional probabilities. We also observe that in the settings using all ICD codes, Mask-SA reranker provides larger improvement to the macro F-score consistently.

Our proposed framework improves upon the best-performing methods on all settings. Note that the proposed framework is complementary to the base predictor. The results show that our reranking method can improve upon any predictor that is designed with the independent assumption, demonstrating the great flexibility and generalizability of our method.

Model	Avg. Rank
LAAT	24.58
+ MADE	19.73
+ Mask-SA	19.50

Table 3: Average rank of the best-performing label set among the top-50 candidates.

4.5 Effect of Candidate Numbers

The reranking results reported in Table 1 and Table 2 are generated with top-50 candidates. In order to investigate the effect of number of candidates to the final performance, we plot the performance with regard to different number of candidates in Figure 4. As shown in the figure, the reranked score increases consistently when the number of candidates is less than 10. No significant improvement is observed when the number of candidates is larger than 10.

We hypothesize that this phenomenon is due to our formulation of the final score. When calculating the final scores, we combine the original score from the base predictor and the score from the reranker. For the candidates that originally ranked after 10 by the base predictor, the original score may be too low; hence it is almost impossible to be selected after reranking.

4.6 Effectiveness of Reranking

The ultimate goal of our reranker is to bring the best-performing label set to the highest rank. In order to further examine the effectiveness of our rerankers, we calculate the average ranking of the best-performing label set, i.e. the set with the highest micro F-score with respect to the ground truth, before and after reranking. The results are shown in Table 3, implying that the proposed model can bring the best candidate from the 24-th place to the 19-th place for better practicality in terms of the systems with doctors' interactions. Our rerankers improve the average ranking by more than 20% relative, demonstrating that the reranking process is effective.

4.7 Effect on Infrequent Labels

The task of ICD code prediction is extremely hard due to the large set of labels and the imbalanced data: the top-50 most frequent codes take up more than a third of all the outputs. To investigate the effect of the proposed framework on the infrequent labels, we bucket the labels according to their fre-

	Baseline + Rescoring
Sample 1	427.1 427.41 427.5 693.0 99.6 995.0
(+ Mask-SA)	427.1 427.41 427.5 693.0 99.6 995.0 99.62 96.04 96.71
Sample 2	571.5 733.00 733.09 96.04 96.72 V66.7
(+ MADE)	571.5 733.00 ~~733.09~~ 96.04 96.72 V66.7 305.1 431 96.6

Table 4: Sample results before and after reranking from MIMIC-3 data.

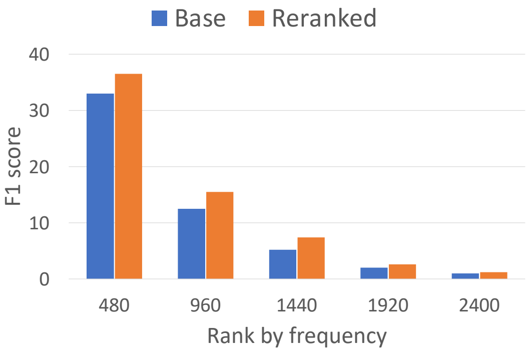

Figure 5: F1 scores (%) with regard to the frequencies of labels. We bucket labels by their frequencies into 6 buckets.

quencies, and calculate the performance for each bucket. We plot the performance of MultiResCNN on MIMIC-3 full set with regard to label frequencies in Figure 5. The figure demonstrates that with reranking, the performance of the infrequent labels also increases. This result indicates that the reranking method is helpful for the extreme multi-label classification problem.

5 Qualitative Analysis

After comparing the results produced from two rerankers, we find that both methods have similar tendency of prediction. In other words, the original candidate sets would be improved by adding or deleting similar ICD codes after rescoring from both MADE and Mask-SA. To further analyze prediction change, Table 4 shows the original and reranked results for two data samples.

5.1 Addition and Deletion of Predictions

For the first sample, we find that the reranking module tends to add missing ICD codes to the predicted set. Specifically, the first sample has no 96 category in the original prediction, and the rescoring process adds 96.04 and 96.71 (highlighted in blue) in the candidate set for better performance. By checking their meanings, we could know that 96.04 is about insertion of endotracheal tube and 96.71 is about invasive mechanical ventilation, and both treatments are important for patients in ICU maintaining their respiratory function. Due to their strong dependency, we find that these codes frequently co-occur in the training data. Apparently, the reranker learn the correlation and is capable of improving the prediction in terms of both diversity and accuracy.

In the second sample, it can be found that our module can also help remove the unreasonable codes. Specifically, the code 733.09 (highlighted in red) is not proper to be the selected code due to the appearance of 733.00, which is the correct disease from the record. Therefore, the reranker can help not only provide additional accurate codes but also delete unreasonable ones for better performance.

5.2 Reranking Analysis

We further analyze our methods from the top-10 ranking candidates sets in the second sample to confirm if the sets with more accurate ICD codes would be at the top of the reranked sets. In this sample, the unreasonable ICD code 733.09 appears in every top-10 predictions before reranking. With reranking, our reranker is able to bring the set without 733.09 to the top-1. This example demonstrates that the reranker's ability to identify conflicting predictions and that we are able to correct them with the proposed framework.

6 Conclusions

This paper proposes a novel framework to improve multi-label classification for automatic ICD coding, which includes candidate generation and candidate reranking modules. In the first stage, a base predictor is performed to generate top-k probable label set candidates. In the second stage, we propose a reranker to capture the correlation between ICD labels without any external knowledge. Two types of the reranker, MADE and Mask-SA, are employed to rerank the candidate sets. Our experiments show that both rerankers can consistently improve the performance of all predictors in MIMIC-2 and MIMIC-3 datasets, demonstrating the generalizability of our framework and the great potential of the flexible usage.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.

Tal Baumel, Jumana Nassour-Kassis, Raphael Cohen, Michael Elhadad, and Noemie Elhadad. 2018. Multi-label classification of patient notes: case study on icd code assignment. In Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence.

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.

Weiwei Cheng, Eyke Hüllermeier, and Krzysztof J Dembczynski. 2010. Bayes optimal multilabel classification via probabilistic classifier chains. In ICML, pages 279–286.

Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F Stewart, and Jimeng Sun. 2016. Doctor ai: Predicting clinical events via recurrent neural networks. In Machine Learning for Healthcare Conference, pages 301–318.

Michael Collins and Terry Koo. 2005. Discriminative reranking for natural language parsing. Computational Linguistics, 31(1):25–70.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive language models beyond a fixed-length context. In ACL, pages 2978–2988, Florence, Italy. Association for Computational Linguistics.

Luciano RS de Lima, Alberto HF Laender, and Berthier A Ribeiro-Neto. 1998. A hierarchical approach to the automatic categorization of medical documents. In CIKM, pages 132–139. ACM.

Jia Deng, Nan Ding, Yangqing Jia, Andrea Frome, Kevin Murphy, Samy Bengio, Yuan Li, Hartmut Neven, and Hartwig Adam. 2014. Large-scale object classification using label relation graphs. In ECCV, pages 48–64. Springer.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In NAACL, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Mathieu Germain, Karol Gregor, Iain Murray, and Hugo Larochelle. 2015. Made: Masked autoencoder for distribution estimation. In ICML, pages 881–889.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, 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.

Ramakanth Kavuluru, Anthony Rios, and Yuan Lu. 2015. An empirical evaluation of supervised learning approaches in assigning diagnosis codes to electronic medical records. Artificial intelligence in medicine, 65(2):155–166.

Cheng Li, Virgil Pavlu, Javed Aslam, Bingyu Wang, and Kechen Qin. 2019. Learning to calibrate and rerank multi-label predictions. In ECML-PKDD.

Cheng Li, Bingyu Wang, Virgil Pavlu, and Javed Aslam. 2016. Conditional bernoulli mixtures for multi-label classification. In ICML, pages 2482–2491.

Fei Li and Hong Yu. 2020. Icd coding from clinical text using multi-filter residual convolutional neural network. In AAAI.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable prediction of medical codes from clinical text. In NAACL, pages 1101–1111, New Orleans, Louisiana. Association for Computational Linguistics.

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(5p2):1620–1639.

M. Ostendorf, A. Kannan, S. Austin, O. Kimball, R. Schwartz, and J.R. Rohlicek. 1991. Integration of diverse recognition methodologies through reevaluation of N-best sentence hypotheses. In Speech and Natural Language: Proceedings of a Workshop Held at Pacific Grove, California, February 19-22, 1991.

Adler Perotte, Rimma Pivovarov, Karthik Natarajan, Nicole Weiskopf, Frank Wood, and Noémie Elhadad. 2013. Diagnosis code assignment: models and evaluation metrics. Journal of the American Medical Informatics Association, 21(2):231–237.

Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. 2009. Classifier chains for multi-label classification. In ECML-KDD, pages 254–269. Springer.

Mohammed Saeed, Mauricio Villarroel, Andrew T Reisner, Gari Clifford, Li-Wei Lehman, George Moody, Thomas Heldt, Tin H Kyaw, Benjamin Moody, and Roger G Mark. 2011. Multiparameter intelligent monitoring in intensive care ii (mimic-ii): a public-access intensive care unit database. Critical care medicine, 39(5):952.

Federico Sangati, Willem Zuidema, and Rens Bod. 2009. A generative re-ranking model for dependency parsing. In Proceedings of the 11th International Conference on Parsing Technologies (IWPT'09), pages 238–241, Paris, France. Association for Computational Linguistics.

Libin Shen, Anoop Sarkar, and Franz Josef Och. 2004. Discriminative reranking for machine translation. In HLT-NAACL, pages 177–184.

Haoran Shi, Pengtao Xie, Zhiting Hu, Ming Zhang, and Eric P Xing. 2017. Towards automated icd coding using deep learning. arXiv preprint arXiv:1711.04075.

Shang-Chi Tsai, Ting-Yun Chang, and Yun-Nung Chen. 2019. Leveraging hierarchical category knowledge for data-imbalanced multi-label diagnostic text understanding. In LOUHI, pages 39–43, Hong Kong. Association for Computational Linguistics.

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 NIPS, pages 5998–6008.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. 2008. Extracting and composing robust features with denoising autoencoders. In ICML, pages 1096–1103.

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. International Joint Conferences on Artificial Intelligence Organization.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144.

Pengtao Xie and Eric Xing. 2018. A neural architecture for automated ICD coding. In ACL, pages 1066–1076, Melbourne, Australia. Association for Computational Linguistics.

Yinyuan Zhang, Ricardo Henao, Zhe Gan, Yitong Li, and Lawrence Carin. 2018. Multi-label learning from medical plain text with convolutional residual models. In Machine Learning for Healthcare Conference, pages 280–294.

Chenxi Zhu, Xipeng Qiu, Xinchi Chen, and Xuanjing Huang. 2015. A re-ranking model for dependency parser with recursive convolutional neural network. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 1159–1168, Beijing, China. Association for Computational Linguistics.