RL-based sample selection improves transfer learning in low-resource and imbalanced clinical settings

Anonymous ACL submission

Abstract

A common strategy in transfer learning is few shot fine-tuning, but its success is highly dependent on the quality of samples selected as training examples. Although active learning methods like uncertainty sampling and diversity sampling can pick useful samples, they underperform in low-resource and class-imbalanced conditions. We introduce a more robust sample selection strategy using reinforcement learning (RL) to identify the most informative samples. Combined with back-translation data augmentation, this approach greatly improved model adaptability in low-resource and class-imbalanced settings. Experimental evaluations on two clinical datasets related to invasive fungal infection (IFI) show our RL-based sample selection strategy enhances model transferability and still maintains robust performance under extreme class imbalance compared to traditional methods. An ablation study on data augmentation reveals that this approach can greatly enhance performance when only a few samples are available, but as sample size grows, the quality of back-translation is also crucial for the model's performance.

1 Introduction

Unlike general texts, clinical reports are typically challenging to analyze using Natural Language Processing (NLP) because they contain specialized symbols, abbreviations, and medical jargon. The effectiveness of NLP techniques in healthcare heavily relies on the quality of annotated datasets (Touvron et al., 2023; Liu et al., 2024a). However, due to data restrictions and the rarity of many disease conditions, acquiring large amounts of gold standard data in healthcare can be difficult. The high cost of annotation further restricts the availability of labeled data. Therefore, maximizing the utility of limited data becomes a crucial research focus.

Transfer Learning (TL) (Tan et al., 2018) is an approach in which knowledge learned from a task is reused to boost performance on a related task. It has shown effectiveness across various machine learning applications (Weiss et al., 2016), and opens new avenues for addressing low-resource scenarios. Previous works have attempted to leverage pretrained embeddings (Maimaiti et al., 2021) and few shot examples (Alyafeai et al., 2020) to facilitate transfer learning in NLP. However, when the target task offers very few labeled instances, these approaches may generate unreliable outputs. This is an especially acute problem in healthcare, where reliability is paramount.

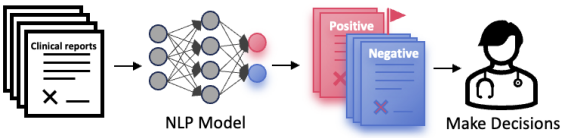

Figure 1: Overview of the NLP-based IFIs detection surveillance framework.

Class-imbalance (Johnson and Khoshgoftaar, 2019) is another challenge often present in low-resource settings. In clinical datasets, there is often a scarcity of positive cases due to the low prevalence of many conditions, making such instances both valuable and limited in number. Differences in data-collection protocols can lead some cohorts to contain very few positive samples, while others may be overwhelmingly positive. These extreme disparities in class distribution further hinder the transferability of NLP models across heterogeneous clinical datasets.

Invasive fungal infections (IFIs) pose significant risks to patients with weakened immune systems, requiring timely detection (Neofytos et al., 2009). Previous works have shown the efficacy of NLP techniques in detecting IFIs from clinical reports (Rozova et al., 2023; Martinez et al., 2015). However, available IFI related datasets are limited and also face the class-imbalance prob-

lem. These datasets come from diverse sources, such as reports from CT and PET scans, or cytology and histopathology findings, where the content structures, terminology, and linguistic expressions differ. In different document types, IFI detection cues overlap to some degree, but existing IFI detection models still fall short of human performance in transferring knowledge between them. As the preparation of gold standard annotated datasets is time-consuming, effective learned knowledge transfer from existing datasets to new but similar tasks becomes valuable. This not only improves annotation efficiency but also enhances the models' adaptability in dealing with similar tasks.

In this work, we propose a more robust strategy for knowledge transfer between similar but different sources, particularly for low-resource and class-imbalanced environments. Firstly, we employ a reinforcement learning (RL) based strategy to identify the most informative samples within new, unlabeled datasets. The sampling policy relies on learner feedback derived from existing knowledge. It considers both content representations and model confidence. After selection, medical experts annotate the selected examples for use in fine-tuning. Finally, we apply back-translation to address the shortage of medical content in low resource conditions. Experimental results show that our approach improves both the adaptability and performance of IFI detection between different sources. In the context of transfer learning, this approach offers a promising way to both reduce annotation effort and enhance model robustness in low-resource and class-imbalanced settings.

Our contributions are summarized as follows:

- This work is the first to address the challenges posed by low-resource and class-imbalance scenarios in IFI detection from medical reports.
- We propose a more robust RL-based sample selection strategy tailored to scenarios with both data scarcity and class imbalance.
- We demonstrate that back-translation based data augmentation further mitigates the challenges posed by low-resource conditions.
- Extensive experiments on two IFI-related datasets confirm that our transfer learning approach is more effective between similar but different sources, even under low-resource and class-imbalanced conditions.

2 Related Work

With high-quality annotated datasets, NLP methods have shown promising results in infection detection. Based on the concept features relevant to IFIs, dictionary-based detection approaches have shown effective performance (Rozova et al., 2023; Martinez et al., 2015). Bag-of-words models have also been utilized, often combined with machine learning techniques to further enhance accuracy and scalability in infection detection (Cury et al., 2021; López-Úbeda et al., 2020). Recently, large language models (LLMs), such as BioBERT (Lee et al., 2020) and ClinicalBERT (Huang et al., 2019), pre-trained on large biomedical corpora, have further improved contextual understanding in clinical texts (Consoli et al., 2024; Boligarla et al., 2023).

Low resource settings remain challenging for NLP tasks. Researchers have explored various transfer strategies to improve model performance by learning from limited data or external knowledge. Few-shot fine-tuning (Mann et al., 2020; Gu et al., 2021; Liu et al., 2022a), where large pre-trained models are adapted using only a small number of labeled examples, has shown promising results. Selecting effective few-shot samples is critical, and active learning strategies such as uncertainty sampling (Nguyen et al., 2022) and diversity sampling (Yang et al., 2015) are often employed. However, these active learning approaches typically focus on a single metric and may not perform consistently across diverse scenarios. Reinforcement Learning (RL) (Fang et al., 2017; Liu et al., 2024b) offers a potential solution by optimizing more flexible and adaptive sample selection policies, thereby improving robustness in different contexts. Data augmentation (Li et al., 2022; Shorten et al., 2021; Bayer et al., 2022) is a strategy to address limited training data by generating more training examples from existing ones. Common methods in NLP include paraphrasing, synonym replacement, word perturbation, back-translation, and synthetic text generation. These techniques expand the dataset and help models to improve performance in low-resource settings. Although these strategies are widely studied, their effectiveness in clinical NLP applications, with highly specialized terminology and noisy texts (Liu et al., 2022b), still requires further exploration and validation.

Class imbalance is especially crucial in low-resource clinical NLP tasks (Ghosh et al., 2024). To address this, previous studies have explored various

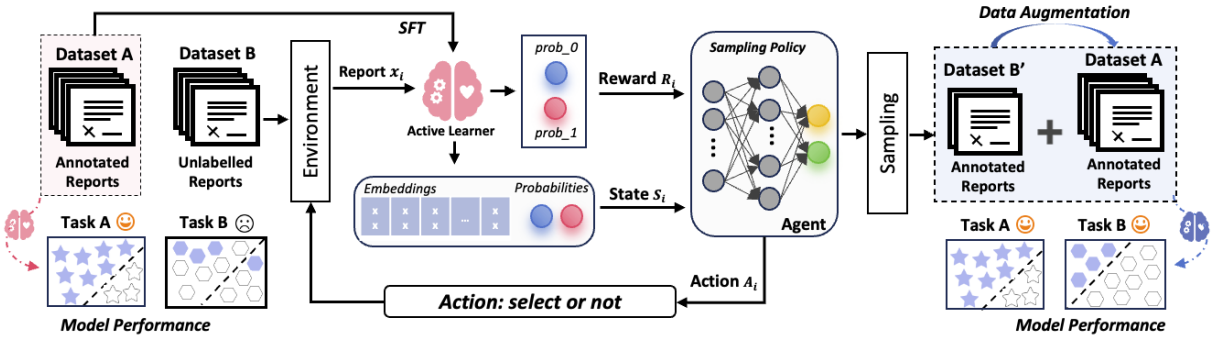

Figure 2: An overview of our framework for transfer learning from A to B. Task A and Task B share some similar knowledge but still have differences. Without transfer learning, the model trained only on Dataset A performs well on Task A but not on a Task B. Our sample selection strategy uses RL to identify key samples from Dataset B and finally applies back-translation. The resulting model achieves good performance on both Task A and Task B. Task A = test set of Dataset A, Task B = test set of Dataset B. Dataset A contains train/dev reports and has labels. Dataset B contains unlabeled train/dev reports. Dataset B' contains the selected samples from Dataset B and have been annotated. SFT = Supervised Fine-tuning.

strategies. Data-level approaches, such as oversampling minority classes (Hairani et al., 2024) and undersampling majority classes (Yang et al., 2024), are typically used to balance class distributions. Algorithm-level methods, such as cost-sensitive learning (Araf et al., 2024) and focal loss adjustments (Aljohani et al., 2023), aim to direct model attention toward underrepresented classes, thereby improving model performance. Despite these advances, effectively managing class imbalance in low-resource clinical scenarios continues to be an active area of investigation.

3 Methodology

3.1 Overview

The overall framework of our proposed method is shown in Figure 2. Initially, we fine-tune an active learner with the source dataset. The active learner then provides feedback in the form of document embeddings and predicted probabilities for a new, unlabeled target dataset. This new dataset contains unlearned knowledge and shares partial similarity with the source domain. These outputs from feedback than serve as the input state representation for a Sampling Policy Network (SPN) which is trained via RL to select the most informative samples. The selected informative samples are subsequently annotated and combined with the original annotated dataset. Finally, the expanded dataset is enhanced through back-translation. With supervised fine-tuning on the final dataset, the knowledge can be effectively transferred and the model performance across heterogeneous datasets also improved.

3.2 Sample Selection Strategy

Our sample selection strategy uses RL to identify the most informative samples between low-resource and class-imbalanced datasets. The aim is to improve the model's transferability across different medical datasets with the selected samples. The pseudocode for this part is provided in Appendix A. This strategy consists of the following key components:

Active Learner. The active learner is a lightweight classifier fine-tuned on a fully annotated dataset. Its primary role is to serve as an analytic tool. It uses learned representations from labeled data to evaluate and provide diagnostic feedback on unlabeled datasets.

State Representation. We construct the state representation for each unlabeled report based on the outputs of the active learner. Specifically, each state vector combines the contextual embedding derived for the unlabeled report and its predicted probabilities for each class. We also integrate two informative metrics: confidence and margin. Confidence is defined as the highest predicted class probability, and the margin is calculated as the absolute difference between class probabilities. Finally, the state vector is defined as:

$$s_i = [h_(x_i); p_0(x_i); p_1(x_i); c(x_i); m(x_i)] \quad (1)$$

where $h_(x_i)$ are the contextual embeddings of the unlabeled report, $p_0(x_i)$ and $p_1(x_i)$ are predicted class probabilities, $c(x_i)$ is the confidence, and $m(x_i)$ represents the probability margin. This comprehensive representation equips the RL agent with

nuanced diagnostic signals to effectively guide the sample selection process. The overall dimension of the state vector is set to 772.

Sampling Policy Network (SPN). We formulate the sample selection process by training a SPN. This is a deep Q-learning network (Hester et al., 2018) trained to decide whether each unlabeled report should be selected (action $a_i = 1$) or discarded (action $a_i = 0$). To be specific, we optimize the network parameters θ by minimizing the mean squared error (MSE) between the predicted Q-values and the target Q-values:

$$\mathcal{L}(\theta) = \mathbb{E}_{(s,a,r,s')\sim\mathcal{B}}\Big[\big(Q(s,a;\theta) - (r + \gamma \max_{a'} Q(s',a';\theta^{-}))\big)^2\Big] \quad (2)$$

where $\mathcal{B}$ is the experience replay buffer, (s, a, r, s') are past experiences, γ is the discount factor, and θ^{-} represents the parameters of the periodically updated target network. Through this training process, the SPN learns an effective sampling policy that prioritizes selecting the most informative samples for improving model transferability across different datasets.

Reward. In our RL framework, we use the classification margin $m(x_i)$ as the reward r_i. This margin quantifies the active learner's confidence in its prediction. A larger margin indicates higher certainty, whereas a smaller margin signals greater uncertainty. After the agent taking action $a_i \in \{0, 1\}$ (select vs. discard), the agent receives r_i and updates its network via the Bellman target:

$$\hat{Q}_i = r_i + \gamma \max_{a'} Q_{\text{target}}(s_{i+1}, a'), \quad (3)$$

Over episodes, this reward structure drives the policy toward an optimal sampling strategy under a limited annotation budget.

Sample Selection. After the SPN is trained, we apply the learned policy to the unlabeled dataset. For each sample, we compute the action with the highest expected reward according to the SPN:

$$a_i^* = \arg\max_{a\in\{0,1\}} Q(s_i, a; \theta) \quad (4)$$

Samples for which $a_i^* = 1$ are selected for annotation. By iteratively applying this selection process, the resulting annotated subset helps improve the knowledge transfer performance in low-resource and class-imbalanced datasets.

3.3 Data Augmentation

To address the scarcity of textual data, we employ multilingual back-translation as a data augmentation strategy. For each report, we translate the English text into multiple target languages and subsequently back-translate it into English, leveraging pretrained MarianMT models[1] from Hugging Face. The back-translation augmentation process for a given text x is defined as follows:

$$x_{\text{BT}} = f_{\text{Tgt}\to\text{En}}\left(f_{\text{En}\to\text{Tgt}}(x)\right) \quad (5)$$

where $f_{\text{En}\to\text{Tgt}}$ and $f_{\text{Tgt}\to\text{En}}$ are pretrained translation models from English to a target language and from the target language back to English.

We utilize several language pairs to enhance linguistic diversity, including English–Chinese (zh), English–French (fr), English–German (de), and English–Spanish (es). This augmentation is applied to both the selected samples and the annotated dataset which contains previously acquired knowledge. By enriching multilingual variations, this approach mitigates the low-resource issues inherent in NLP tasks.

4 Experimental Setup

4.1 Datasets

We chose two IFI-related clinical datasets as benchmarks in this study: the Cytology and Histopathology IFI Reports corpus (CHIFIR[2]) and the PET-CT Invasive Fungal Infection Reports corpus (PIFIR[3]). These data originated from 2 Australian hospitals, one tertiary referral center and one specialized cancer hospital.

Figure 3: Word clouds for the CHIFIR (right) and PIFIR (left) datasets. Word size corresponds to term frequency.

Although both datasets are related to IFI, the vocabulary used varies across them. The CHIFIR dataset is collected from cytology and histopathology reports. These reports assess tissue or fluid

[1] https://huggingface.co/Helsinki-NLP

[2] Available for credentialed users at https://physionet.org/content/corpus-fungal-infections/1.0.2/

[3] Available for credentialed users at https://physionet.org/content/pifir/1.0.0/

Transfer Learning Strategy		Performance on PIFIR				Performance on CHIFIR			
		Accuracy	F1 score	Precision	Recall	Accuracy	F1-score	Precision	Recall
Zero-shot Transfer	Fine-tuned on CHIFIR	0.33	0.12	**1.00**	0.06	**0.94**	**0.76**	**1.00**	**0.63**
	Fine-tuned on PIFIR	**0.88**	**0.92**	0.88	**0.97**	0.48	0.27	0.17	**0.63**
Full-shot Transfer	Fine-tuned on CHIFIR + PIFIR	0.83	0.89	0.85	0.94	0.90	0.55	**1.00**	0.38

Table 1: Performance comparison of fine-tuned BioBERT over CHIFIR and PIFIR with different strategies. Models perform well when fine-tuned and evaluated on the same dataset. In contrast, evaluated on a similar but still different dataset causes a clear performance drop. Fine-tuned on a single dataset but evaluated on another dataset can be regard as the zero-shot transfer, and we set this as our baseline. Fine-tuned and evaluated on both datasets can be regarded as the full-shot transfer, and we assume this is the best of transfer learning performance.

samples and describe the microscopic visualization of fungal organisms. The PIFIR dataset is collected from PET-CT reports. These reports assess metabolic activity and discuss the anatomical and morphological features of fungal lesions via PET-CT imaging. Figure 3 shows differences in their predominant clinical terms. More detailed concept level analysis of differences can be found in Appendix B. Both CHIFIR and PIFIR exhibit class imbalance. CHIFIR is dominated by negative cases (negative: 86%, positive: 14%), whereas PIFIR is dominated by positive cases (negative: 31%, positive: 69%). This imbalance not only increases the difficulty of effective knowledge transfer but also poses challenges for experimental evaluation by potentially skewing performance metrics.

Split	CHIFIR			PIFIR		
	Total	P	N	Total	P	N
Train	202	28	174	139	94	45
Dev	29	4	25	20	14	6
Test	52	8	44	42	31	11

Table 2: Class distribution for CHIFIR and PIFIR across train, development, and test sets. P = the number of positive reports, N = the number of negative reports.

For the CHIFIR dataset, expert annotators have provided report-level classification labels indicating whether a report is positive or negative for an IFI, as well as span-level annotation of concepts relevant to IFI detection. It contains 283 reports from 201 patients, with an average length of 1,384 characters. The PIFIR dataset also has report- and concept- level annotations. It includes 201 reports from 156 patients, with an average of 1,457 characters. The original datasets do not contain a development set. We therefore split each dataset into training, validation, and test parts. The ratio is around 70:10:20, and the original class balance is kept. Table 2 shows the number of positive and negative samples in each split.

4.2 Evaluation Metrics

We employ a variety of metrics for evaluation, including accuracy, F1 score, precision, and recall. Class imbalance in benchmark datasets makes the F1 score particularly important. Recall is also important, given that it is critical not to miss IFI-positive cases.

4.3 Baselines

We select the fine-tuned BioBERT approach from previous work (Anonymous) as the baseline. When fine-tuned separately on CHIFIR and PIFIR dataset, the model performs well on its own data but poorly on the other dataset. Although both datasets focus on IFIs and they share some similarities, the model still struggles to apply what it learned from one dataset to the other. Table 1 shows the baseline results and reveals the challenges of knowledge transferability between these datasets. Although training on both CHIFIR and PIFIR can improve performance, it requires annotating all reports in the unlabeled dataset, which is labor-intensive. We observed that even when training on both datasets together, the evaluation performance on the CHIFIR dataset remains low, with an F1 score of 0.55. It suggests that transferring knowledge from CHIFIR to PIFIR is relatively simple, while the reverse transfer from PIFIR to CHIFIR is more difficult.

5 Experimental Results

5.1 Cross-Dataset Transfer Learning Performance

We set the number of few shot examples to $k \in \{1, 2, 4, 8, 16, 32\}$; this scale is widely used in earlier few shot studies (Brown et al., 2020). The largest value, 32 samples, is safe for our data because it adds about 16% to the training instances in CHIFIR dataset and about 23% to the instances in PIFIR dataset, so the auxiliary samples never outweigh the target domain.

Knowledge Transfer from CHIFIR to PIFIR						
Sample Selection (in PIFIR)			**Performance on PIFIR**			
Strategy	**Num**	**P:N**	Acc	F1	P	R
Randomly	1	–	0.3143	0.1500	0.4800	0.0900
	2	–	0.4476	0.3866	0.7208	0.2968
	4	–	0.5810	0.5974	**0.9359**	0.4710
	8	–	0.6571	0.7180	0.8871	0.6774
	16	–	0.7381	0.8113	0.8693	0.7871
	32	–	**0.7667**	**0.8508**	0.8087	**0.9032**
Uncertainty	1	1.00 : 0.00	0.4524	0.4103	**1.0000**	0.2581
	2	1.00 : 0.00	0.5000	0.5330	0.8571	0.3871
	4	1.00 : 0.00	0.6667	0.7667	0.7931	0.7419
	8	1.00 : 0.00	0.6667	0.8000	0.7179	0.9032
	16	0.94 : 0.06	0.6667	0.8000	0.7179	0.8710
	32	0.94 : 0.06	**0.7381**	**0.8493**	0.7381	**1.0000**
Diversity	1	1.00 : 0.00	0.3571	0.2286	**1.0000**	0.1290
	2	1.00 : 0.00	0.6190	0.6923	0.8571	0.5806
	4	1.00 : 0.00	0.6667	0.7812	0.7576	0.8065
	8	0.87 : 0.13	**0.7857**	**0.8525**	0.8667	0.8387
	16	0.87 : 0.13	0.7381	0.8493	0.7381	**1.0000**
	32	0.91 : 0.09	0.7381	0.8493	0.7381	**1.0000**
Our Method	1	1.00 : 0.00	0.2857	0.1176	0.6667	0.0645
	2	0.50 : 0.50	0.3571	0.2286	**1.0000**	0.1290
	4	0.75 : 0.25	0.6429	0.6809	**1.0000**	0.5161
	8	0.87 : 0.13	0.7381	0.8493	0.7381	**1.0000**
	16	0.75 : 0.25	**0.7857**	**0.8696**	0.7895	0.9677
	32	0.69 : 0.31	**0.7857**	**0.8696**	0.7895	0.9677

Knowledge Transfer from PIFIR to CHIFIR						
Sample Selection (in CHIFIR)			**Performance on CHIFIR**			
Strategy	**Num**	**P:N**	Acc	F1	P	R
Randomly	1	–	0.5115	**0.2071**	**0.1729**	**0.5000**
	2	–	0.6038	0.1804	0.1548	0.3500
	4	–	0.7346	0.1556	0.1433	0.1750
	8	–	0.7692	0.0307	0.0400	0.0250
	16	–	**0.8307**	0.0364	0.0667	0.0250
	32	–	0.8269	0.0000	0.0000	0.0000
Uncertainty	1	0.00 : 1.00	0.7692	0.1429	0.1667	0.1250
	2	0.50 : 0.50	0.3654	**0.3265**	0.1951	**1.0000**
	4	0.33 : 0.67	0.6154	0.2857	0.2000	0.5000
	8	0.13 : 0.87	0.8077	0.1667	**0.2500**	0.1250
	16	0.13 : 0.87	0.7115	0.0000	0.0000	0.0000
	32	0.13 : 0.87	**0.8462**	0.0000	0.0000	0.0000
Diversity	1	0.00 : 1.00	0.4423	**0.2927**	0.1818	**0.7500**
	2	0.00 : 1.00	0.5962	0.2759	0.1905	0.5000
	4	0.25 : 0.75	0.6154	0.2857	**0.2000**	0.5000
	8	0.13 : 0.87	**0.8462**	0.0000	0.0000	0.0000
	16	0.06 : 0.94	**0.8462**	0.0000	0.0000	0.0000
	32	0.06 : 0.94	**0.8462**	0.0000	0.0000	0.0000
Our Method	1	1.00 : 0.00	0.4423	**0.3256**	0.2000	**0.8750**
	2	0.50 : 0.50	0.6923	0.2727	0.2143	0.3750
	4	0.50 : 0.50	0.5769	0.3889	0.2500	**0.8750**
	8	0.38 : 0.62	0.6346	0.2400	0.1765	0.3750
	16	0.31 : 0.69	0.7500	0.1333	0.1429	0.1250
	32	0.25 : 0.75	**0.8654**	0.2222	**1.0000**	0.1250

Table 3: Transfer learning performance from CHIFIR to PIFIR (left) and from PIFIR to CHIFIR (right) under different sample selection strategies. Num = the number of annotated reports; P:N = the positive-to-negative ratio in these annotations. Metrics reported are Accuracy (Acc), F1 score (F1), Precision (P) and Recall (R).

We compare our strategy with several other active learning approaches to analyze the impact of different sample selection methods on knowledge transfer performance. In this analysis, we do not apply back translation data augmentation, thus isolating the effects of the sample selection itself:

1) **Random Selection**: We randomly select k samples from the unlabeled dataset for annotation. Each experiment is run five times to reduce variance and obtain more reliable results. We report the mean evaluation metrics over these five runs.

2) **Uncertainty-based Selection** (Nguyen et al., 2022): Uncertainty score is the lowest confidence predicted by the active learner. Based on this, we select k samples for annotation and evaluate transfer learning performance on a different dataset.

3) **Diversity-based Selection** (Margatina et al., 2021): Diversity score measures report similarity between datasets. We calculate the cosine distance between each report embedding in the unlabeled dataset and the embeddings in the labeled dataset. Higher scores show greater differences between samples. We select the k most diverse samples for annotation, and evaluate the resulting performance.

4) **Our method**: Our RL-based sampling strategy selects the most informative samples based on contextual embeddings, predicted probabilities, confidence, and margin. We annotate these samples and then evaluate their effectiveness for improving knowledge transfer.

The results of our transfer learning experiments across the CHIFIR and PIFIR datasets are shown in Table 3. We analyze how the number of selected samples affects performance. Although both our datasets are focused on IFI-detection, the experimental results show that unlike in context learning, with 1 samples selected, few shot fine-tuning can not greatly improve the robustness of model performance on the other datasets. As k increases from 1 to 32, all methods show steady improvement. However, after $k = 16$, the performance gain becomes smaller, suggesting that additional annotations bring limited benefits once enough informative samples are selected.

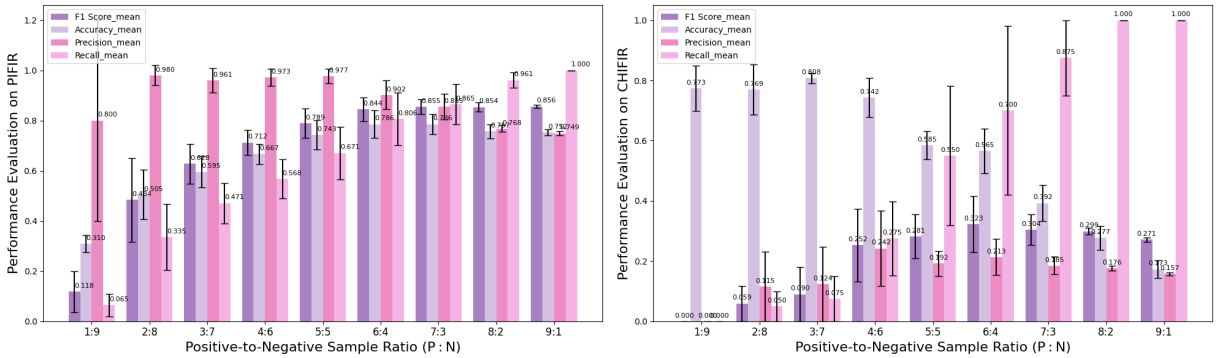

Figure 4: Class imbalance analysis of positive to negative sample ratios for transfer learning from CHIFIR to PIFIR (left) and from PIFIR to CHIFIR (right). For each ratio, we randomly selected 20 samples and repeated the experiment five times to evaluate transfer performance; bars show mean values and black lines indicate variance.

Compared to other sample selection strategies, our approach shows greater robustness. In transfer learning from CHIFIR to PIFIR, with more than 16 samples, uncertainty-based selection and diversity-based selection result in models that predict all cases as positive and no negative samples are correctly identified (False Negative = 0 and True Negative = 0). Our method avoids this problem, maintaining balanced predictions and good performance on both classes. Similarly, in transfer learning from PIFIR to CHIFIR, uncertainty and diversity methods fail to predict positive cases when more than 8 samples are selected. Their F1 score, precision, and recall drop to zero, as models predict all reports as negative. In contrast, our method still has the ability to identify positive samples even under class imbalance, although the performance does not increase with more samples, suggesting that this transfer is more challenging.

In Appendix C, we provide additional evaluations of model performance on the labeled dataset after transfer learning. Interestingly, our method achieves stable results and even slightly outperforms joint fine-tuning on both datasets. This may be because joint fine-tuning risks overfitting when data is limited, while our sample-efficient strategy mitigates this.

5.2 Robustness under Imbalanced Sampling

We evaluate the robustness of our sampling strategy under class imbalance conditions. First, we use the random selection strategy to choose 20 samples from the unlabeled dataset with different positive to negative ratios (from 1:9 to 9:1). Each setting is repeated five times to obtain stable results. Figure 4 shows the outcomes. For knowledge transfer from CHIFIR to PIFIR, the best performance occurs when the positive to negative ratio is around 7:3, which closely matches the actual class distribution in PIFIR. In the reverse transfer, from PIFIR to CHIFIR, increasing the positive ratio raises recall but hurts accuracy. This happens because CHIFIR has many more negative cases, and a high positive ratio makes the model label almost everything as positive. To predict both positive and negative samples well, the ratio still needs to match CHIFIR's true distribution.

Table 3 shows uncertainty and diversity strategies tend to select mostly positive or negative samples. Specifically, when selecting 10 samples, these methods choose only positive cases. This imbalance causes the trained models to classify nearly all reports as positive in the PIFIR dataset. Consequently, these methods perform poorly at identifying negative samples, resulting in low precision and recall for the negative class. In reverse class distribution scenarios, choosing more samples of the minority class from the data already learned helps the model tell positive and negative classes apart. However, in low-resource and class-imbalanced settings, this choice can easily lead the model to overfit on the new task.

Our RL-based sampling method selects samples having class ratios that better match the true distribution of each dataset. By choosing a more suitable ratio of positive and negative examples, our method helps the model better identify both classes. Our method improves knowledge transfer between heterogeneous datasets, outperforming uncertainty and diversity sample selection strategies.

5.3 Effects of Data Augmentation

Since the knowledge transfer performance from PIFIR to CHIFIR is unstable, we concentrate here

on the CHIFIR to PIFIR transfer performance. Figure 5 (upper) shows the knowledge transfer performance after informative sample selection, comparing results with and without back translation for Chinese, French, German, and Spanish together. It shows that back translation has a large impact when the sample size is small, but as the sample size grows, its effect becomes less obvious and may even decrease slightly. Since we applied back translation in four languages, we further evaluate how each language affects model performance. The results are shown in Figure 5 (lower).

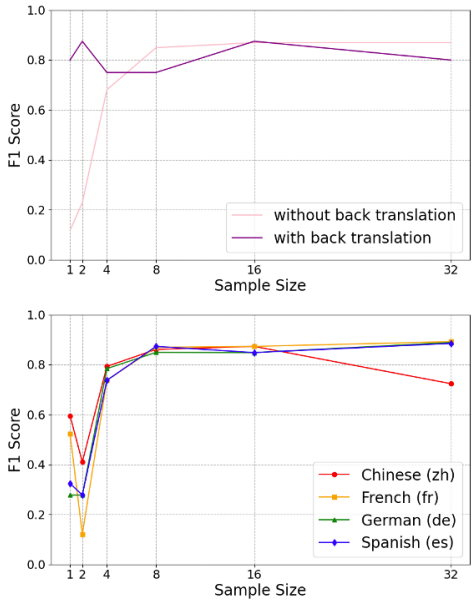

Figure 5: The effects of data augmentation performance on knowledge transfer from CHIFIR to PIFIR after informative sample selection.

To understand the impact of each language, we examined the model's F1 scores at different sample sizes. We found that German and Spanish translations gave stable, high performance even with few samples. Models trained with German translations achieved the highest overall F1 scores and exhibited robustness when the number of samples increased. Spanish translations showed similar trends and reliable results. French translations demonstrated moderate effectiveness. While achieving good results with adequate samples, performance fluctuated significantly when sample sizes were small. Chinese translations, despite adding linguistic diversity, led to unstable and lower overall performance. Models trained on Chinese translations initially improved with limited samples but then dropped noticeably with increased sample sizes. This likely results from significant semantic shifts introduced during translation, causing confusion and negatively affecting learning effectiveness.

To access translation quality quantitatively, we evaluate the BLEU scores (Papineni et al., 2002) between the original texts and their back-translated versions as shown in Figure 6. Higher BLEU scores suggest translations closely resemble the original reports. Overall, BLEU scores were modest across languages, but German and Spanish achieved the highest scores. Chinese translations had the lowest BLEU scores, reflecting substantial differences from the original texts. This explains why model performance dropped slightly when using 32 samples augmented with Chinese back translation.

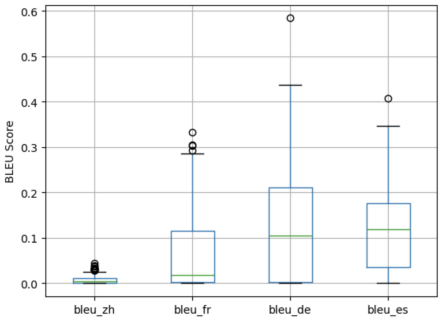

Figure 6: BLEU scores between original reports and their back-translated versions for different languages.

Overall, our ablation study highlights that although multilingual back-translation enhances model adaptability in low resource scenarios, careful selection of target languages is also crucial.

6 Conclusion and Future Work

In this work, we addressed the challenge of transfer learning for IFI detection under low-resource and class-imbalanced conditions. We proposed an RL-based sampling strategy and enhanced the selected data with multilingual back-translation. Our approach improves model performance and adaptability across diverse medical datasets compared to traditional sample selection strategies. The back-translation is effective at small sample sizes and the quality of translation is crucial for knowledge transfer. Future work will explore additional advanced data augmentation techniques to further enhance model robustness. Investigating methods to better align multilingual representations and integrate clinical domain-specific knowledge more effectively will also be valuable. Finally, extending this framework to other NLP tasks and broader clinical contexts could demonstrate its general applicability and effectiveness.

Limitations

Despite demonstrating promising results, our approach has several limitations. First, the datasets used for evaluation originate from specific clinical contexts. Future validations should include more diverse datasets from multiple domains. Second, the effectiveness of our RL-based sample selection depends on the diagnostic feedback provided by the active learner. This places high quality demands on the original gold dataset. Any inaccuracies or biases in active learner predictions could negatively influence the quality of selected samples. Finally, while back-translation proved beneficial, it can introduce semantic shifts, especially in languages significantly different from English. Careful consideration and further linguistic validation are required when applying multilingual augmentation strategies in sensitive clinical scenarios.

Acknowledgments

Supported by anonymous grant and relevant Ethics approvals.

References

Naif Radi Aljohani, Ayman Fayoumi, and Saeed-Ul Hassan. 2023. A novel focal-loss and class-weight-aware convolutional neural network for the classification of in-text citations. *Journal of Information Science*, 49(1):79–92.

Zaid Alyafeai, Maged Saeed AlShaibani, and Irfan Ahmad. 2020. A survey on transfer learning in natural language processing. *arXiv preprint arXiv:2007.04239*.

Imane Araf, Ali Idri, and Ikram Chairi. 2024. Cost-sensitive learning for imbalanced medical data: a review. *Artificial Intelligence Review*, 57(4):80.

Markus Bayer, Marc-André Kaufhold, and Christian Reuter. 2022. A survey on data augmentation for text classification. *ACM Computing Surveys*, 55(7):1–39.

Srikanth Boligarla, Elda Kokoè Elolo Laison, Jiaxin Li, Raja Mahadevan, Austen Ng, Yangming Lin, Mamadou Yamar Thioub, Bruce Huang, Mohamed Hamza Ibrahim, and Bouchra Nasri. 2023. Leveraging machine learning approaches for predicting potential lyme disease cases and incidence rates in the united states using twitter. *BMC Medical Informatics and Decision Making*, 23(1):217.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Sergio Consoli, Peter Markov, Nikolaos I Stilianakis, Lorenzo Bertolini, Antonio Puertas Gallardo, and Mario Ceresa. 2024. Epidemic information extraction for event-based surveillance using large language models. In *International Congress on Information and Communication Technology*, pages 241–252. Springer Nature Singapore Singapore.

Ricardo C Cury, Istvan Megyeri, Tony Lindsey, Robson Macedo, Juan Batlle, Shwan Kim, Brian Baker, Robert Harris, and Reese H Clark. 2021. Natural language processing and machine learning for detection of respiratory illness by chest ct imaging and tracking of covid-19 pandemic in the united states. *Radiology: Cardiothoracic Imaging*, 3(1):e200596.

Meng Fang, Yuan Li, and Trevor Cohn. 2017. Learning how to active learn: A deep reinforcement learning approach. *arXiv preprint arXiv:1708.02383*.

Kushankur Ghosh, Colin Bellinger, Roberto Corizzo, Paula Branco, Bartosz Krawczyk, and Nathalie Japkowicz. 2024. The class imbalance problem in deep learning. *Machine Learning*, 113(7):4845–4901.

Yuxian Gu, Xu Han, Zhiyuan Liu, and Minlie Huang. 2021. Ppt: Pre-trained prompt tuning for few-shot learning. arxiv 2021. *arXiv preprint arXiv:2109.04332*.

Hairani Hairani, Triyanna Widiyaningtyas, and Didik Dwi Prasetya. 2024. Addressing class imbalance of health data: A systematic literature review on modified synthetic minority oversampling technique (smote) strategies. *JOIV: International Journal on Informatics Visualization*, 8(3):1310–1318.

Todd Hester, Matej Vecerik, Olivier Pietquin, Marc Lanctot, Tom Schaul, Bilal Piot, Dan Horgan, John Quan, Andrew Sendonaris, Ian Osband, and 1 others. 2018. Deep q-learning from demonstrations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2019. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*.

Justin M Johnson and Taghi M Khoshgoftaar. 2019. Survey on deep learning with class imbalance. *Journal of big data*, 6(1):1–54.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Bohan Li, Yutai Hou, and Wanxiang Che. 2022. Data augmentation approaches in natural language processing: A survey. *Ai Open*, 3:71–90.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024a. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin A Raffel. 2022a. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. *Advances in Neural Information Processing Systems*, 35:1950–1965.

Jinghui Liu, Daniel Capurro, Anthony Nguyen, and Karin Verspoor. 2022b. "Note Bloat" impacts deep learning-based NLP models for clinical prediction tasks. *Journal of Biomedical Informatics*, 133:104149.

Ying Liu, Haozhu Wang, Huixue Zhou, Mingchen Li, Yu Hou, Sicheng Zhou, Fang Wang, Rama Hoetzlein, and Rui Zhang. 2024b. A review of reinforcement learning for natural language processing and applications in healthcare. *Journal of the American Medical Informatics Association*, 31(10):2379–2393.

Pilar López-Úbeda, Manuel Carlos Díaz-Galiano, Teodoro Martín-Noguerol, Antonio Luna, L Alfonso Ureña-López, and M Teresa Martín-Valdivia. 2020. Covid-19 detection in radiological text reports integrating entity recognition. *Computers in Biology and Medicine*, 127:104066.

Mieradilijiang Maimaiti, Yang Liu, Huanbo Luan, and Maosong Sun. 2021. Enriching the transfer learning with pre-trained lexicon embedding for low-resource neural machine translation. *Tsinghua Science and Technology*, 27(1):150–163.

Ben Mann, N Ryder, M Subbiah, J Kaplan, P Dhariwal, A Neelakantan, P Shyam, G Sastry, A Askell, S Agarwal, and 1 others. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 1:3.

Katerina Margatina, Giorgos Vernikos, Loïc Barrault, and Nikolaos Aletras. 2021. Active learning by acquiring contrastive examples. *arXiv preprint arXiv:2109.03764*.

David Martinez, Michelle R Ananda-Rajah, Hanna Suominen, Monica A Slavin, Karin A Thursky, and Lawrence Cavedon. 2015. Automatic detection of patients with invasive fungal disease from free-text computed tomography (ct) scans. *Journal of biomedical informatics*, 53:251–260.

Dionissios Neofytos, D Horn, E Anaissie, W Steinbach, A Olyaei, J Fishman, M Pfaller, C Chang, K Webster, and K Marr. 2009. Epidemiology and outcome of invasive fungal infection in adult hematopoietic stem cell transplant recipients: analysis of multicenter prospective antifungal therapy (path) alliance registry. *Clinical Infectious Diseases*, 48(3):265–273.

Vu-Linh Nguyen, Mohammad Hossein Shaker, and Eyke Hüllermeier. 2022. How to measure uncertainty in uncertainty sampling for active learning. *Machine Learning*, 111(1):89–122.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Vlada Rozova, Anna Khanina, Jasmine C Teng, Joanne SK Teh, Leon J Worth, Monica A Slavin, Karin A Thursky, and Karin Verspoor. 2023. Detecting evidence of invasive fungal infections in cytology and histopathology reports enriched with concept-level annotations. *Journal of Biomedical Informatics*, 139:104293.

Connor Shorten, Taghi M Khoshgoftaar, and Borko Furht. 2021. Text data augmentation for deep learning. *Journal of big Data*, 8(1):101.

Chuanqi Tan, Fuchun Sun, Tao Kong, Wenchang Zhang, Chao Yang, and Chunfang Liu. 2018. A survey on deep transfer learning. In *Artificial Neural Networks and Machine Learning–ICANN 2018: 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4-7, 2018, Proceedings, Part III 27*, pages 270–279. Springer.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. 2016. A survey of transfer learning. *Journal of Big data*, 3:1–40.

Cynthia Yang, Egill A Fridgeirsson, Jan A Kors, Jenna M Reps, and Peter R Rijnbeek. 2024. Impact of random oversampling and random undersampling on the performance of prediction models developed using observational health data. *Journal of Big Data*, 11(1):7.

Yi Yang, Zhigang Ma, Feiping Nie, Xiaojun Chang, and Alexander G Hauptmann. 2015. Multi-class active learning by uncertainty sampling with diversity maximization. *International Journal of Computer Vision*, 113:113–127.

A Algorithm

We provide the algorithm for the strategy of sample selection in our approach, as depicted in Algorithm 1. The state vector has 772 elements: the 768-dimensional `[CLS]` embedding, two class probabilities, the confidence (max-probability) score and the margin between the two probabilities.

Algorithm 1 RL-based sample selection strategy

Require: pool $\mathcal{U}$, budget B, classifier f_θ, tokenizer T

$env \leftarrow \textsc{ActiveLearningEnv}(\mathcal{U}, f_\theta, T, B)$
initialise Q_ϕ, $\hat{Q}_\phi \leftarrow Q_\phi$, replay buffer $\mathcal{D}$
for episode $= 1$ to N **do**
  $s \leftarrow$ *env.reset()*
  while not done **do**
    $a \leftarrow \epsilon$-greedy(Q_ϕ, s)
    $(s', r, done) \leftarrow$ *env.step*(a)
    store $(s, a, r, s', done)$ in $\mathcal{D}$
    if $|\mathcal{D}| \geq$ batchSize **then** $\textsc{UpdateNets}(Q_\phi, \hat{Q}_\phi, \mathcal{D})$
    end if
    $s \leftarrow s'$
  end while
  if episode mod $K = 0$ **then** $\hat{Q}_\phi \leftarrow Q_\phi$
  end if
end for
Selection
$\mathcal{S} \leftarrow \varnothing$, $s \leftarrow$ *env.reset()*
while not done **do**
  $a \leftarrow \arg\max_{a'} Q_\phi(s, a')$
  $(s', \_, done) \leftarrow$ *env.step*(a)
  if $a = 1$ **then** add current sample to $\mathcal{S}$
  end if
  $s \leftarrow s'$
end while
return $\mathcal{S}$

B Concept-level Differences Analysis in CHIFIR and PIFIR Datasets

Concept annotations in CHIFIR and PIFIR Datasets are listed in Table 4 and Table 5. CHIFIR dataset totally reports 1,155 concepts and PIFIR dataset has 3,194 concepts. The two corpora serve different clinical niches. CHIFIR comes from cytology and histopathology notes and therefore focuses on microbiology terms such as `FungalDescriptor` and `Stain`. PIFIR is built from PET–CT reports and centres on imaging findings and risk factors, for example `Abnormality_CT` and `Risk_factor`.

Concept	Count	Unique	Diversity
`ClinicalQuery`	68	43	0.63
`FungalDescriptor`	294	86	0.29
`Fungus`	106	19	0.18
`Invasiveness`	39	27	0.69
`Stain`	172	16	0.09
`SampleType`	198	64	0.32
`positive`	118	40	0.34
`equivocal`	8	6	0.75
`negative`	152	12	0.08

Table 4: Summary statistics for the IFI-related concepts in the CHIFIR dataset.

Concept	Count	Unique	Diversity
`Infection_or_IFI`	279	174	0.62
`Risk_factor`	429	179	0.42
`Abnormality`	46	24	0.52
`Abnormality_CT`	460	204	0.44
`Abnormality_PET`	470	224	0.48
`Lung`	372	36	0.10
`Sinus`	19	4	0.21
`Other`	189	92	0.49
`Infection_Inflammation`	354	103	0.29
`IFI_Indication`	37	21	0.57
`improvement`	115	51	0.44
`stable`	29	16	0.55
`worsening`	55	34	0.62
`positive`	124	33	0.27
`equivocal`	129	68	0.53
`negative`	87	23	0.26

Table 5: Summary statistics for the IFI-related concepts in the PIFIR dataset.

To quantify overlap we compute the Jaccard Similarity between the concept vocabularies:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{6}$$

where A and B are the sets of surface forms in CHIFIR and PIFIR, respectively. Figure 7 plots the resulting heat-map. Although both datasets include the labels `positive`, `equivocal` and `negative`, their lexical realisations share little common ground, so the Jaccard scores remain low.

C Cross-Dataset Transfer Performance on Original Dataset

Table 6 reports how well each sample selection strategy performs when evaluated on the source dataset after transfer. We include two settings. In the first, we fine-tune on CHIFIR and PIFIR samples then test on CHIFIR. In the second, we fine-tune on PIFIR and selected CHIFIR samples and test on PIFIR. For reference, Table 1 shows our zero-shot and full-shot baselines. A model trained only on CHIFIR does well on CHIFIR but poorly

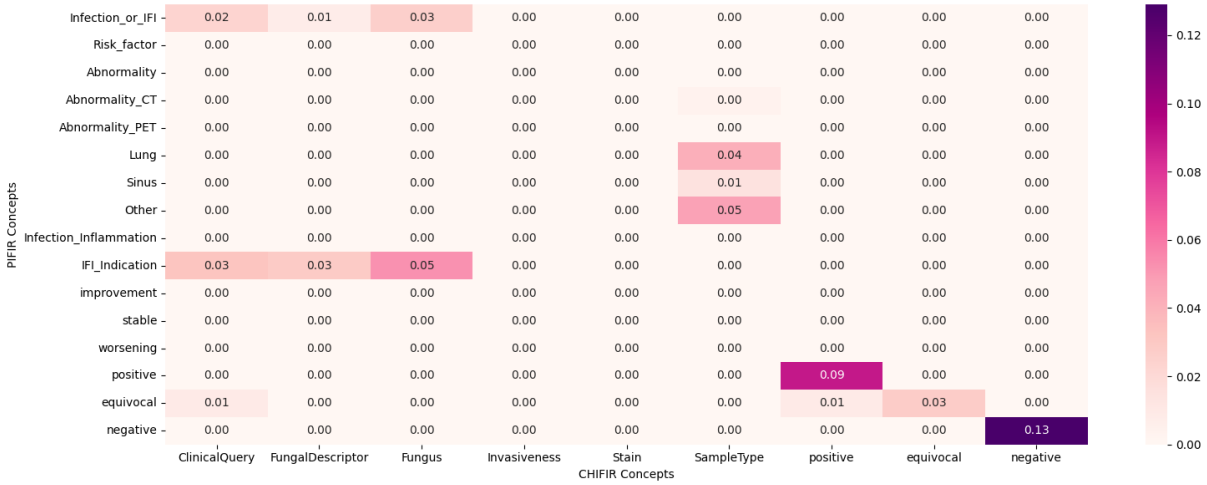

Figure 7: Jaccard similarity heatmap between CHIFIR and PIFIR concepts.

on PIFIR and vice versa. Joint fine-tuning on both datasets yields strong results on both.

In the appendix results, after transfer learning, the model performance on the original datasets remains stable and even improves in some cases compared to before transfer. For example, when transferring from CHIFIR to PIFIR, adding two or sixteen samples chosen by our RL method yields the highest accuracy, F1 score, precision, and recall. In both cases, the model makes only one false positive and one false negative. These results also show that transfer learning not only improves model adaptability but also preserves performance on the source dataset.

Knowledge Transfer from CHIFIR to PIFIR

Sample Selection (in PIFIR)			Performance on CHIFIR			
Strategy	**Num**	**P:N**	Acc	F1	P	R
Randomly	1	–	0.9269	0.7600	0.7714	**0.7500**
	2	–	0.9296	0.7538	**0.8000**	0.7250
	4	–	**0.9308**	**0.7714**	**0.8000**	**0.7500**
	8	–	0.9269	0.7600	0.7714	**0.7500**
	16	–	0.9269	0.7600	0.7714	**0.7500**
	32	–	0.9231	0.7500	0.7500	**0.7500**
Uncertainty	1	1.00 : 0.00	0.9423	0.8000	0.8571	**0.7500**
	2	1.00 : 0.00	0.9432	0.7692	**1.0000**	0.6250
	4	1.00 : 0.00	**0.9615**	**0.8571**	**1.0000**	**0.7500**
	8	1.00 : 0.00	0.9423	0.8000	0.8571	**0.7500**
	16	0.94 : 0.06	0.9423	0.8000	0.8571	**0.7500**
	32	0.94 : 0.06	0.9231	0.7500	0.7500	**0.7500**
Diversity	1	1.00 : 0.00	0.9423	0.8000	0.8571	0.7500
	2	1.00 : 0.00	**0.9615**	**0.8750**	0.8750	**0.8750**
	4	1.00 : 0.00	0.9423	0.8000	0.8571	0.7500
	8	0.87 : 0.13	0.9423	0.8000	0.8571	0.7500
	16	0.87 : 0.13	0.9423	0.8000	0.8571	0.7500
	32	0.91 : 0.09	0.9038	0.5455	**1.0000**	0.3750
Our Method	1	1.00 : 0.00	0.9423	0.8000	0.8571	0.7500
	2	0.50 : 0.50	**0.9615**	**0.8750**	**0.8750**	**0.8750**
	4	0.75 : 0.25	0.9231	0.7500	0.7500	0.7500
	8	0.87 : 0.13	0.9231	0.7500	0.7500	0.7500
	16	0.75 : 0.25	**0.9615**	**0.8750**	**0.8750**	**0.8750**
	32	0.69 : 0.31	0.9423	0.8000	0.8571	0.7500

Knowledge Transfer from PIFIR to CHIFIR

Sample Selection (in CHIFIR)			Performance on PIFIR			
Strategy	**Num**	**P:N**	Acc	F1	P	R
Randomly	1	-	**0.9048**	**0.9375**	**0.9091**	0.9677
	2	-	0.8333	0.8923	0.8529	0.9355
	4	-	0.8333	0.8955	0.8333	0.9677
	8	-	0.8810	0.9254	0.8611	**1.0000**
	16	-	0.8095	0.8824	0.8108	0.9677
	32	-	0.8095	0.8750	0.8485	0.9032
Uncertainty	1	0.00 : 1.00	0.8571	0.9032	0.9032	0.9032
	2	0.50 : 0.50	0.8571	0.9091	0.8571	0.9677
	4	0.33 : 0.67	0.8571	0.9091	0.8571	0.9677
	8	0.13 : 0.87	0.8571	0.9091	0.8571	0.9677
	16	0.13 : 0.87	0.8571	0.9118	0.8378	**1.0000**
	32	0.13 : 0.87	**0.8810**	**0.9254**	**0.8611**	**1.0000**
Diversity	1	0.00 : 1.00	**0.9286**	**0.9538**	0.9118	**1.0000**
	2	0.00 : 1.00	0.8095	0.8710	0.8710	0.8710
	4	0.25 : 0.75	0.9048	0.9394	0.8857	**1.0000**
	8	0.13 : 0.87	0.8571	0.9091	0.8571	0.9677
	16	0.06 : 0.94	**0.9286**	0.9524	**0.9375**	0.9677
	32	0.06 : 0.94	**0.9286**	**0.9538**	0.9118	**1.0000**
Our Method	1	1.00 : 0.00	0.7619	0.8485	0.8000	0.9032
	2	0.50 : 0.50	0.8571	0.9091	0.8571	**0.9677**
	4	0.50 : 0.50	0.7857	0.8615	0.8235	0.9032
	8	0.38 : 0.62	**0.8810**	**0.9231**	**0.8824**	**0.9677**
	16	0.31 : 0.69	0.7857	0.8657	0.8056	0.9355
	32	0.25 : 0.75	**0.8810**	**0.9231**	**0.8824**	**0.9677**

Table 6: Transfer learning performance from CHIFIR to PIFIR (left) and from PIFIR to CHIFIR (right) under different sample selection strategies. Num = the number of annotated reports; P:N = the positive-to-negative ratio in these annotations. Metrics reported are Accuracy (Acc), F1 score (F1), Precision (P) and Recall (R).