Neurocomputing 633 (2025) 129809

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Decoupled contrastive learning for multilingual multimodal medical pre-trained model

Qiyuan Li [a,b,c], Chen Qiu [a,b,c,*], Haijiang Liu [a,b,c], Jinguang Gu [a,b,c], Dan Luo [d]

[a] College of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan, 430065, Hubei, China
[b] Hubei Province Key Laboratory of Intelligent Information Processing and Real-time Industrial System, Wuhan, 430065, Hubei, China
[c] The Key Laboratory of Rich-Media Knowledge Organization and Service of Digital Publishing Content, Institute of Scientific and Technical Information of China, Beijing, 100038, China
[d] Department of Computer Science and Engineering, Lehigh University, Bethlehem, 18015, USA

ARTICLE INFO

Communicated by G. Fenza

Keywords:
Multilingual multimodal learning
Decoupled contrastive learning
Medical pre-training model

ABSTRACT

Multilingual multimodal pre-training aims to facilitate the integration of conceptual representations across diverse languages and modalities within a shared, high-dimensional semantic space. This endeavor in healthcare faces challenges related to language diversity, suboptimal multimodal interactions, and an absence of coherent multilingual multimodal representations. In response to these challenges, we introduce a novel multilingual multimodal medical pre-training model. Initially, we employ a strategic augmentation of the medical corpus by expanding the MIMIC-CXR report dataset to 20 distinct languages using machine translation techniques. Subsequently, we develop a targeted label disambiguation technique to address the labeling noise within decoupled contrastive learning. In particular, it categorizes and refines uncertain phrases within the clinical reports based on disease type, promoting finer-grained semantic similarity and improving inter-modality interactions. Building on these proposals, we present a refined multilingual multimodal medical pre-trained model, significantly enhancing the understanding of medical multimodal data and adapting the model to multilingual medical contexts. Experiments reveal that our model outperforms other baselines in medical image classification and multilingual medical image–text retrieval by up to 13.78% and 12.6%, respectively.

1. Introduction

A multilingual multimodal pre-trained model integrates both multilingual and multimodal data, including but not limited to text, image, and audio. These models necessitate extensive multilingual multimodal corpora for the pre-training, intending to extract and encode shared feature representations across many languages and modalities. Consequently, such models [1,2] enhance their comprehension and generative capabilities concerning heterogeneous data, thereby demonstrating superior performance in tasks such as image–text retrieval and image/text generation. Furthermore, the intricate and varied representational prowess of these models renders them highly promising for a broad spectrum of applications across diverse sectors, including finance [3], healthcare [4], education [5], social media [6], and so on [7].

In healthcare, being aware of the multilingual nature of medical data is critical to delivering comprehensive healthcare services. The 2023 edition of the Healthcare World Language Index [8] from AMN Healthcare indicates a notable diversity in health information dissemination and disease manifestation across cultures and languages. However, the development of medical multilingual pre-trained models faces two principal challenges: first, the scarcity of multilingual medical data, and second, the resulting imbalance in language distribution. In particular, the centralized nature of the available data resources leads to a bias in which Chinese and English data occupy the vast majority. In contrast, data in other languages are relatively scarce, severely limiting the model's ability to cover other languages and its application potential. For example, when a language model trained primarily on English data is used to develop a dialogue system, it may perform well for English speakers but poorly for Mandarin, Arabic, or Swahili speakers. On the other hand, it is worth noting that the scarcity of medical data is one of the main reasons for the unbalanced language distribution. This unbalanced data distribution leads to a bias in the construction and evaluation of medical multilingual pre-trained

* Corresponding author at: College of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan, 430065, Hubei, China.
E-mail addresses: vickyuan@wust.edu.cn (Q. Li), chen@wust.edu.cn (C. Qiu), alecliu@ontoweb.wust.edu.cn (H. Liu), simon@wust.edu.cn (J. Gu), dal417@lehigh.edu (D. Luo).

https://doi.org/10.1016/j.neucom.2025.129809
Received 2 June 2024; Received in revised form 18 February 2025; Accepted 21 February 2025
Available online 1 March 2025

models, affecting the performance and applicability of the models in different linguistic environments. Given this state of affairs, there is a relative paucity of research [9,10] on medical multilingual pre-trained models, and the field is still at an exploratory stage, requiring more attention and investment.

Other than the multilingual characteristic of medical data, it also faces the challenges of multi-modality. The composition of medical data includes not only text, but also images such as X-rays, CTs, and MRIs, which play a critical role in clinical applications [11]. Traditional multimodal pre-trained models in medical image analysis frequently rely on large-scale labeled data, which is costly to acquire. Consequently, self-supervised learning, which utilizes unlabeled data for pre-training through a contrastive learning framework [12], is widely adopted to reduce the dependence on labeled data. For example, ConVIRT [13] employs a joint training mechanism for visual and textual encoders utilizing paired medical images and reports by bidirectionally contrasting targets. GLoRIA [14] advances the modeling of global and local interactions between medical images and their corresponding reports, thereby capturing pathological significance from specific regions within the images. But both of them suffer from false negatives, where the text in a negative sample may describe the same symptoms as the anchor text. Then, MedCLIP [15] introduces a re-training strategy for image–text pairs based on decoupled contrastive learning, which operates on image and text labels. However, this method is susceptible to semantic labeling errors and negative word omission due to the misrecognition of uncertain phrases during the labeling process. Therefore, we need to address the problem of labeling noise, which in turn diminishes the matching accuracy of positive samples.

To address these issues, we propose a novel multilingual multimodal medical pre-trained model called 3M-CLIP. In particular, to mitigate the challenges of scarcity and imbalanced data distribution, 3M-CLIP employs a strategic augmentation of the medical corpus by expanding the linguistic scope of the report texts within the MIMIC-CXR [16] dataset to 20 distinct languages using machine translation techniques. Furthermore, 3M-CLIP develops a targeted label disambiguation technique to address the labeling noise within decoupled contrastive learning. In particular, it categorizes and refines uncertain phrases within the clinical reports based on disease type, promoting finer-grained semantic similarity and improving inter-modality interactions. To this end, we are able to construct a refined multilingual multimodal medical pre-training model that aspires to learn a generalized representation of multilingual multimodality in the medical domain. This model not only extends the applicability of the framework to medical multilingual contexts but also enhances the understanding of medical multimodal data. Experiments demonstrate that the model outperforms other baseline models in medical image classification and multilingual medical image–text retrieval by up to 13.78% and 12.6%, respectively.

Our contributions are summarized as follows:

- We expand the MIMIC-CXR dataset to encompass 20 distinct languages, subsequently training a medical multilingual text encoder with the augmented dataset to bolster the model's capabilities in handling multilingual medical data.
- We introduce a novel targeted label disambiguation method that aims to establish robust semantic similarity, thereby harmonizing the semantics of visual and textual modalities and augmenting inter-modal interaction.
- We present a refined multilingual multimodal medical pre-training model, significantly enhancing the understanding of medical multimodal data and adapting the model to multilingual medical contexts.
- Experiments show that the model outperforms other baselines by up to 13.78% and 12.6% on medical image classification and multilingual medical image–text retrieval tasks, respectively.

2. Related work

This chapter reviews the extant scholarly advancements in multilingual pre-trained models, multimodal pre-trained models, and multilingual multimodal pre-trained models, transitioning from a general domain perspective to a specialized focus on medicine.

2.1. Multilingual pre-trained models

Multilingual pre-trained models learn multilingual generic knowledge representations mainly by pre-training on large-scale textual data in multiple languages. These models can understand and generate text in different languages, thus realizing multilingual tasks with minimal or no language-specific labeled data. Popular models include mBERT [17], XLM-RoBERTa [18], M2M100 [19], mBART-50 [20], CodeGeeX [21], and others. However, there are fewer multilingual pre-trained models in healthcare, as the semantics of medical terminology terms may differ from general knowledge. One such model is the MMedLM 2 [9], recently proposed by Qiu et al. a new multilingual healthcare language model designed to serve a broader multilingual audience. The model outperforms existing open-source models in several benchmarks and is particularly suitable for adaptation to various medical scenarios by fine-tuning with medical instructions, but it only covers six languages. The main challenges encountered in developing medical multilingual pre-trained models are the need for multilingual medical data and the imbalance of medical data distribution, which require further exploration to establish multilingual pre-trained models for the medical field.

In this work, we mitigate data scarcity and imbalance by extending the medical multilingual dataset. Specifically, we extend the dataset's report text to 20 diverse languages through machine translation, then filter and clean it to improve the quality of the translated data, and finally validate the remaining translations to ensure their accuracy and reliability. Moreover, we develop a medical multilingual text encoder trained on the expanded dataset, facilitating the uniform encoding of medical textual data in multiple languages and significantly bolstering the model's capabilities in handling multilingual medical information.

2.2. Multimodal pre-trained models

Multimodal pre-trained models can process various data formats, including text, images, audio, and video. Through pre-training on multimodal data, this model acquires the ability to learn generalized knowledge representations and associations, which allows it to understand and generate content that contains multiple information forms [22–25]. This technology is ideally suited for tasks that involve integrating information from different perceptual modalities, such as image classification [12,26], image recognition [27], cross-modal retrieval [12,28,29], text-to-image generation [30,31], image caption generation [29,30], and visual question answering [30].

Multimodal pre-training tasks are divided into four types, masked language modeling (MLM), masked vision modeling (MVM), vision-language matching (VLM), and vision-language contrastive learning (VCL). Several multimodal pre-training models utilize a staged pre-training approach, performing tasks such as MLM with VLM [28, 32], MLM with MVM [29], or a sequence of MVM, MLM, and VLM tasks [30]. On the other hand, models like CLIP [12] and Filip [33] adopt a one-step pre-training strategy, focusing solely on the VLC task. This approach promotes coherent learning of visual and linguistic information through end-to-end training, enhancing generalization across multimodal tasks. It also reduces training time and optimizes computational resource allocation.

CLIP not only works well in the general domain, but also in the medical domain. Paired image–text data is used for pre-training in some models [13,14,34]. For instance, ConVIRT [13] employs bidirectional contrastive targets for the joint training of visual and textual encoders

using paired medical images and reports. Similarly, the GLoRIA [14] model aims to delineate the global and local interactions between medical images and reports, thereby extracting pathological significance from specific regions within the images. As we have illustrated in the previous chapter, they commonly grapple with the challenge of false negative data. To counteract the issue above, methods [15,35–37] are proposed to mitigate the problem of false negative data. Specifically, MedCLIP [15] seeks to remediate the problem through decoupled contrastive learning, wherein image–text pairings are retrained based on image and text labels. However, it is susceptible to semantic tagging errors and negative word omissions, leading to label noise that diminishes the precision of positive sample matching.

Therefore, we propose a targeted label disambiguation technique specifically tailored to address the labeling noise problem arising from uncertain labels. This technique involves classifying and processing uncertain phrases within the report text according to the relevant disease type. Doing so can effectively address the noise problem and re-pair the image and text. Additionally, we can enhance the intermodal interaction and achieve a cohesive semantic representation.

2.3. Multilingual multimodal pre-trained models

Multilingual multimodal pre-trained models represent a class of sophisticated AI constructs that integrate data from multiple languages and different modalities, including but not limited to text, image, and audio, for pre-training. These models are engineered to extract and synthesize generalized knowledge representations from extensive multilingual multimodal datasets, equipping them with the capacity to excel in diverse downstream tasks. The pre-training process of multilingual multimodal pre-trained models can be categorized into different categories based on the downstream task goals. These categories include comprehension-based tasks like masked language modeling, translated language modeling, and image–text matching, generation-based tasks like text generation and image caption, and comprehension- and generation-based tasks.

There is still a relatively small amount of research on multilingual multimodal pre-trained models in healthcare. These models primarily focus on generative tasks and favor Chinese languages. For instance, the XrayGLM model [38], introduced by Wang et al. facilitates medical image diagnosis and engages in multi-turn interactive dialogues in the Chinese language. Furthermore, XrayPULSE[1] represents an extension of the PULSE[2] framework, aiming to foster biomedical multimodal dialogues in Chinese and English. Additionally, the Med-MLLM model [4], proposed by Liu et al. offers support for English, Chinese, and Spanish, thereby providing a nuanced and robust decision-making framework for COVID-19 in the context of disease reporting, diagnosis, and prognosis. Divergent from the generation-based tasks, our approach is anchored in comprehension-based tasks. We develop a fine-grained multilingual multimodal medical pre-trained model that employs a contrastive learning mechanism. This mechanism is instrumental in learning generalized representations of multilingual multimodality within the medical domain through re-matching between images and multilingual texts. Our model expands its applicability across medical multilingual contexts and significantly bolsters the understanding capacity of medical multimodal data. Moreover, our work unfolds a good start for developing multilingual multimodal pre-trained models based on comprehension tasks in the medical domain.

[1] https://github.com/openmedlab/XrayPULSE.
[2] https://github.com/openmedlab/PULSE.

3. Methodology

We propose a novel multilingual multimodal medical pre-training model to address the challenges of language diversity, suboptimal multimodal interactions, and the lack of coherent multilingual representations in healthcare. Specifically, we implement a strategic augmentation approach that expands the MIMIC-CXR report dataset into 20 distinct languages through machine translation techniques, effectively mitigating the challenges of data scarcity and imbalanced multilingual distribution. Additionally, we introduce a targeted label disambiguation technique to address labeling noise within decoupled contrastive learning, thereby enhancing cross-modal interactions. In the following sections, we will detail our methodology for data augmentation and model construction, as shown in Fig. 1.

3.1. Data augmentation with translation

As discussed above, existing medical data sources severely suffer from data scarcity and language distribution imbalance. To address this problem, we propose data augmentation with translation, which involves translating existing English datasets into multiple languages using a translation model. Given the discrepancy between real and translated data, we will further outline the aspects of data expansion, filtering, and validation.

3.1.1. Data expansion

We choose MIMIC-CXR for our model's training dataset, a commonly used English dataset for multimodal medical pre-training. It is a large chest X-ray database that contains free-text radiology reports from the Beth Israel Deaconess Medical Center in Boston, Massachusetts. Radiology reports typically include sections such as "examination", "indication", "technique", "comparison", "findings", and "impression". The most crucial diagnostic information for X-ray data is usually found in the "findings" or "impression" sections. Therefore, when creating radiology reports, we prioritize extracting the text content from the "findings" section, followed by the "impression" section, as they contain the corresponding X-ray data.

Based on this English dataset, we augment it with data translation. We use the latest NLLB-200 model [39] with 600M pre-training data points to translate 216k English text data extracted from MIMIC-CXR into 20 target languages (see Table A.6 for details). As a result, we can obtain medical reports in 21 different languages.

3.1.2. Filtering and validation

Filtering. We found that the quality of the translation varies between languages, which might impact the performance of our model. Therefore, we filter the translated data to prevent poor-quality data from affecting the model. Our analysis reveals that the most frequent translation errors include multiple translations of single words and the translation of empty text as duplicate long text in multiple languages, as shown in Fig. 2. To address these issues, we introduce two metrics to filter out low-quality translated data.

- **Filtering Continuous and Repeated Translation Data.** We measure the translation quality of a text by determining whether there are more than t continuous and repeated sentences or string segments, which we refer to as the FCR. After data analysis, we set the value of t at 10.
- **Type Token Ratio.** Type Token Ratio (TTR) usually measures the lexical complexity of a text and is calculated as the ratio of $Count_{Types}$ to $Count_{Tokens}$, where $Count_{Types}$ means counting the number of unique words that are not repeated and $Count_{Tokens}$ means calculating the total number of words in a given text.

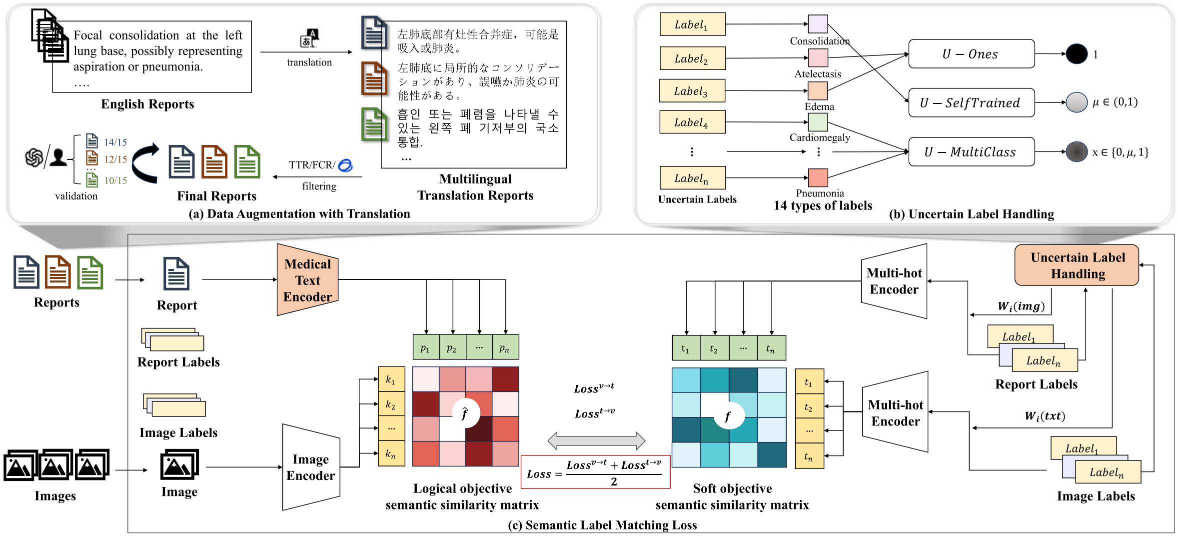


Fig. 1. Overall architecture of our methodology. Our objective is to further our understanding of image and report representations by focusing on the semantic relationships between image and report labels, as well as the raw representation learning of the images and reports themselves. (a) Data Augmentation with Translation. We get the final multilingual reports for input by translating, filtering, and validating. (b) Uncertain Label Handling. We categorize and refine uncertain labels within the clinical reports and images based on disease type. (c) Semantic Label Matching Loss. We combine the semantic similarity losses of logical and soft objectives to obtain the final loss function for the pre-trained model. The semantic similarity of logical objectives is assessed by encoding the input images and reports using visual and text encoders, while soft objectives are calculated by conducting targeted label disambiguation on uncertain labels in the images and reports to determine the weights of these labels and using a multi-hot encoder to combine these weights with deterministic labels.

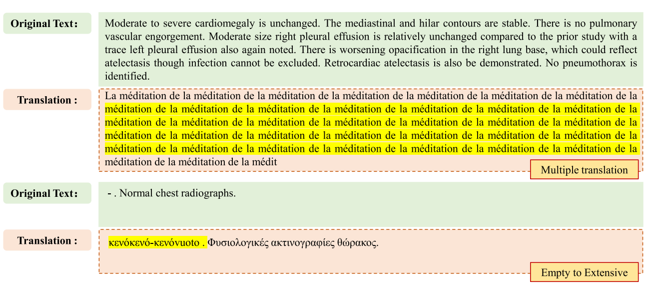


Fig. 2. Examples of frequent translation errors. The above example shows the multiple translations of single words. The following example is translating the empty text as duplicate long text in multiple languages.

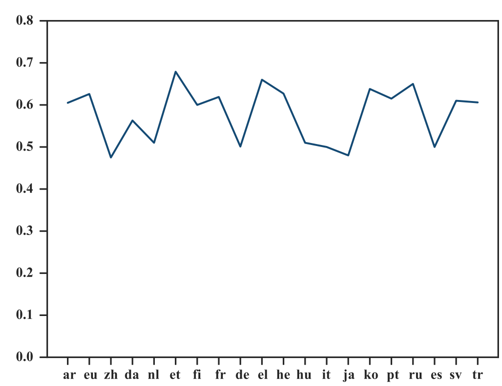


Fig. 3. Filtering thresholds for 20 target languages.

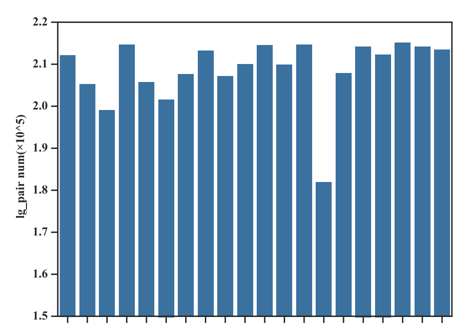


Fig. 4. Distribution of datasets after filtering for 20 target languages.

We evaluate the TTR threshold for each language by manually checking the translation quality of sentences from each of the 20 target languages and then discarding all sentences that scored below the threshold. Considering the varying translation quality across different languages, we apply the same initial TTR threshold (0.5) for all languages. Subsequently, we manually assess the quality of the translated data above and below this threshold, making adjustments to the range to determine the final translation quality threshold. Figs. 3 and 4 show the filtering thresholds and the distribution of the filtered data volume for the 20 target languages, respectively.

Validation. To ensure the quality of the dataset, we first randomly select 1% samples from both the filtered and discarded translation data. Subsequently, we use GLM-4-flash and human scoring to rate their performance on a scale of 1 to 5 based on accuracy, fluency, and contextual consistency (see Table A.7 for specific details of the scoring criteria). Additionally, we also use the XLM-RoBERTa model to generate embeddings, calculate the cosine similarity between translated data, and assess the L1 distance between each language and the mean to comprehensively evaluate the overall quality of the final training dataset [40].

The validation results are shown in Figs. 5 and 6. It is evident from Fig. 5 that the scores of the filtered data are significantly higher than those of the discarded data, regardless of whether they are rated by GLM-4-flash or by humans. Furthermore, Fig. 6 provides a more detailed illustration of the differences in semantic similarity between these two types of data, indicating that the data type with a smaller accumulated error also has a higher quality.

We also examine the impact of the quality of translation data on the results of model training. For this purpose, we randomly sample

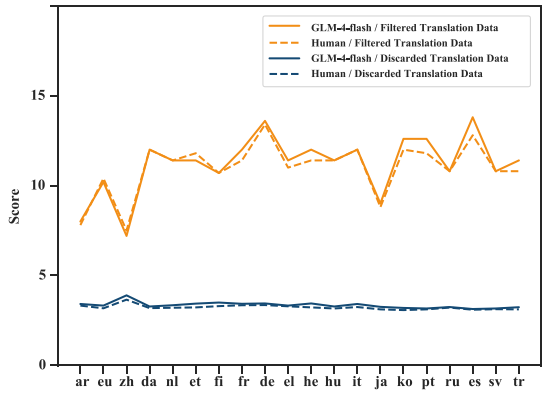


Fig. 5. Results of GLM-4-flash and human scoring on accuracy, fluency, and contextual consistency for both filtered and discarded translation data.

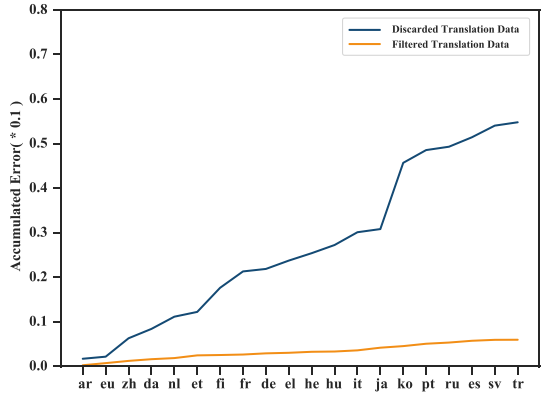


Fig. 6. Accumulated error analysis of L1 distance based on cosine similarity for both filtered and discarded translation data.

10% of the data from the filtered dataset and select the same number of samples from the discarded data to construct a training dataset for model pre-training. Finally, we assess the extent to which translation data quality affects model training performance in medical image classification and image–text retrieval tasks, as shown in Figs. 7 and 8.

Based on the above results, the overall performance of the models trained on filtered data is higher than the models trained on discarded data, in both medical image classification and image–text retrieval tasks. However, this discrepancy is particularly pronounced in the image–text retrieval task. This is because the quality of the multilingual text translation data directly affects the quality of the input text, which in turn impacts the model's ability to learn the association between images and texts. Furthermore, the comparison between the models trained on the 10% filtered data and the full filtered data shows that increasing the volume of training data can indeed improve the model's performance in downstream tasks.

3.2. Model

In this section, we propose a multilingual multimodal medical pre-trained model based on decoupled contrastive learning, as shown in Fig. 1 mentioned above. The 3M-CLIP model consists of three main components: raw vision-text encoder, uncertain vision-text label encoder, and semantic label matching loss.

3.2.1. Raw vision-text encoder

First, we represent the input images and texts as I and T, respectively. For input images, we utilize a vision encoder, denoted as $\mathcal{V}$, to encode images into embeddings $v \in \mathbb{R}^D$. Then the projection head maps the original embedding $v' \in \mathbb{R}^P$, represented as (1a) and (1b).

$$v = \mathcal{V}(I), \tag{1a}$$

$$v' = f_v(v), \tag{1b}$$

where f_v denotes the projection head of the vision encoder.

For input texts, we develop a text encoder to handle textual data in medical reports. To enhance its adaptability to multilingual medical environments, we incorporate translated medical text data into the XLM-RoBERTa model for pre-training. Specifically, we utilize the dual-objective training framework that integrates masked language modeling (MLM) and translation language modeling (TLM) objectives to capture the semantic features of multilingual medical texts. MLM focuses on modeling monolingual text by randomly masking words within individual sentences and having the model predict the masked words. TLM models bilingual parallel texts by concatenating source and target language sentences from parallel pairs and randomly masking words in both sentences, requiring the model to predict the masked words. Our training goal is to minimize the loss function of both, as shown in (2) and (3). After training on these two objectives using multilingual data, we obtained a clinically specialized text encoder that handles multilingual contexts known as Medical-XLM-RoBERTa.

$$L_{mlm}(\theta) = -\frac{1}{N}\sum_{i=1}^{N} \log P(y_i|x,\theta) \tag{2}$$

where x is the input sequence, and θ is a parameter of the model. N represents the total number of masked positions within a batch. y_i denotes the true word at the ith position. $P(y_i|x,\theta)$ indicates the probability that the model predicts the word at the ith position.

$$L_{tlm}(\theta) = -\frac{1}{M}\sum_{j=1}^{M} \log P(z_j|x,y,\theta) \tag{3}$$

where x and y are the input sequences of the two languages, and θ is a parameter of the model. M represents the total number of masked positions within a batch. z_j denotes the true word at the jth position. $P(z_j|x,y,\theta)$ indicates the probability that the model predicts the word at the jth position.

Similar to the vision encoder, we encode the text as an embedding $t \in \mathbb{R}^M$ and subsequently project it onto $t' \in \mathbb{R}^P$, following the Swin Transformer [41] architecture, as represented in (4a) and (4b).

$$t = \mathcal{T}(T), \tag{4a}$$

$$t' = f_t(t), \tag{4b}$$

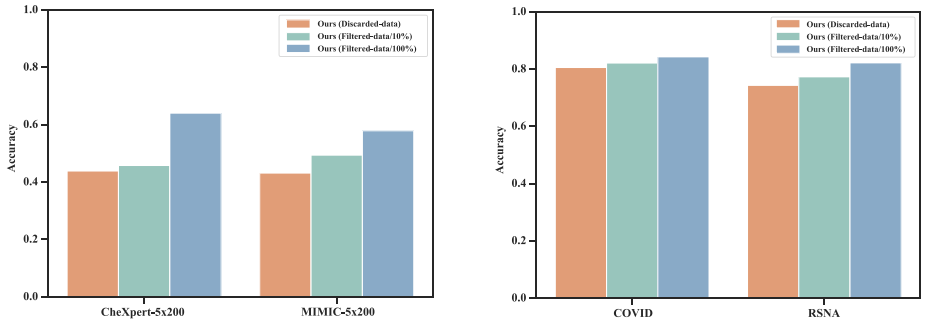
where f_t denotes the projection head and $\mathcal{T}$ denotes the text encoder. Notably, the text encoder has the same dimension P as the visual encoder, ensuring compatibility for contrastive learning.

3.2.2. Uncertain vision-text label encoder

We propose targeted label disambiguation to mitigate label noise ambiguity which is caused by semantic labeling errors or negative word omissions. In particular, we perform targeted disambiguation of uncertainty labels for images and texts based on different disease types.

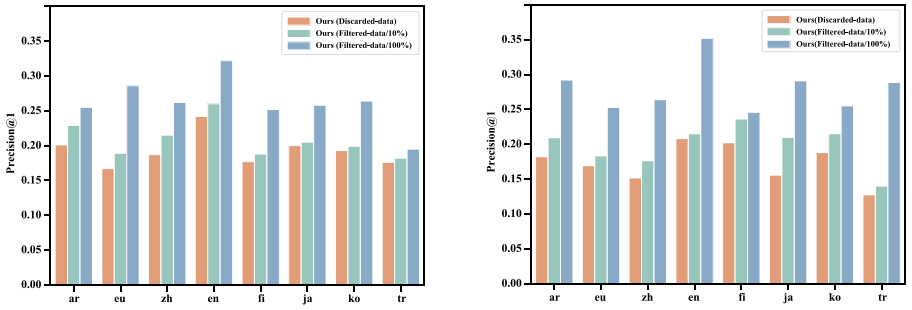
To further enhance the additional supervision, we use external medical knowledge. We extract 14 main entity types (refer to Table A.8 for details) as labels using MetaMap [42] directly from the original sentences. These labels include both deterministic and uncertain labels. For uncertain labels, we define three processing strategies based on the methodology proposed by Irvin et al. [43], combined with five observations from the evaluation of the CheXpert competition task to optimize the management of uncertain labels.

- *U-Ones*: This method maps all uncertain labels to 1.



(a) Routine testing. (b) Zero-shot.

Fig. 7. Accuracy results of our model trained on 10% of the filtered data and the same number of samples from the discarded data in the medical image classification task. (a) shows the routine test results on the CheXpert-5x200 and MIMIC-5x200 datasets. (b) shows the zero-shot results on the COVID and RSNA datasets. We also compare the performance of our model trained on the full filtered dataset.



(a) CheXpert-1000-multi series. (b) MIMIC-1000-multi series.

Fig. 8. Precision@1 results of our model trained on 10% of the filtered data and the same number of samples from the discarded data in the image–text retrieval task, evaluated for one typical language from each of the eight language families in the CheXpert-1000-multi series dataset (a) and the MIMIC-1000-multi series dataset (b). We also compare these results with the performance of our model trained on the full filtered dataset.

- *U-SelfTrained*: This method treats uncertain labels as unlabeled samples and employs semi-supervised learning. First, it trains the model by ignoring uncertain labels, then uses the trained model to predict uncertain labels, and takes the predicted probabilities as new labels for further training.
- *U-MultiClass*: This strategy treats uncertain labels as an independent category. In this method, the model outputs probabilities for three classes (0, 1, and uncertain) for each observation and selects the class with the highest probability as the final prediction.

Finally, we quantitatively present the results of the processing of uncertain labels (u_label) in images and texts in (5).

$$W_{u\_label} = \begin{cases} 1, u\_label \in \{Atelectasis, Edema\} \\ p \in (0,1), u\_label \in Consolidation \\ x \in \{0, p_\mu, 1\}, u\_label \in other\ diseases \end{cases} \tag{5}$$

where 1 represents the result obtained using the U-Ones method, p is the predicted probability value of the uncertain label obtained through the U-SelfTrained method, x indicates the final predicted value of the uncertain label in the U-MultiClass method and p_μ refers to the probability value of the uncertain label category.

Next, we encode the n labels of the images and texts. Considering the characteristics of multi-label types and uncertain labels, we use multi-hot encoding for representation, as shown in (6). To further enhance the model's expressive power, we apply a weighting process to the multi-hot encoding, with the weights determined by the probability values corresponding to each label, as shown in (7). Ultimately, the labels' representation Z can be expressed in (8), which integrates the presence of the labels with their corresponding probabilities, ensuring that the model has greater flexibility and accuracy when handling multi-label and uncertain data.

$$L = [l_1, l_2, \ldots, l_n] \tag{6}$$

where l_k indicates whether the label k exists or not, if the label exists, l_k is 1, otherwise it is 0.

$$W = [w_1, w_2, \ldots, w_n] \tag{7}$$

where weight w_k denotes the probability value of label k.

$$Z = L \odot W = [l_1 \cdot w_1, l_2 \cdot w_2, \ldots, l_n \cdot w_n] \tag{8}$$

3.2.3. Semantic label matching loss

Based on the aforementioned raw vision-text encoder and uncertain vision-text label encoder, we connect images and texts to pre-train the model. In contrast to studies that treat all positive samples equally [44–46], we perform pairwise combinations of images and texts based on the semantic similarity between the image and text labels. Specifically, during each iteration, we independently sample N_{batch} input of images x_{imgs} and texts x_{txts}. To establish the correspondence between images and texts, we construct soft objective by calculating the similarity between the labels, which is defined as (9).

$$Sim = \frac{Z_{img} \cdot Z_{txt}}{\| Z_{img} \| \| Z_{txt} \|} \tag{9}$$

For an image i, we will obtain a set $Sim_{ij} (1 \le j \le N_{batch})$ whose soft objective function is computed by normalizing j by *softmax*.

$$f_{ij}^{v \to t} = \frac{\exp Sim_{ij}}{\sum_{j=1}^{N_{batch}} \exp Sim_{ij}} \tag{10}$$

Similarly, for a text t, we compute its soft objective function as (11).

$$f_{tj}^{t \to v} = \frac{\exp Sim_{tj}}{\sum_{j=1}^{N_{batch}} \exp Sim_{tj}} \tag{11}$$

Table 1
The statistics of the pre-training datasets.

Pretrain	#Images	#Reports	#Classes	#Languages
MIMIC-CXR	377,095	360,654	14	21
CheXpert	223,414	–	14	–

Then, we obtain the logical objective function by calculating the cosine similarity between the image and text embeddings.

$$\hat{Sim}_{ij} = \tilde{v}_i^T \cdot \tilde{t}_j \tag{12}$$

where $\tilde{v}_i$ and $\tilde{t}_j$ are the normalized v_p and t_p, respectively. The softmax function also gives the predicted similarity.

$$\hat{f}_{ij} = \frac{\exp \hat{Sim}_{ij}/\tau}{\sum_{i=1}^{N_{batch}} \exp \hat{Sim}_{ij}/\tau} \tag{13}$$

$$\hat{f}_{tj} = \frac{\exp \hat{Sim}_{tj}/\tau}{\sum_{t=1}^{N_{batch}} \exp \hat{Sim}_{tj}/\tau} \tag{14}$$

where τ is the temperature coefficient, a hyperparameter that controls the sharpness of the probability distribution. Therefore, the final semantic matching loss is the cross entropy between the soft objective function and the logical objective function, which is given by:

$$Loss^{v\to t} = -\frac{1}{N_{batch}} \sum_{i=1}^{N_{batch}} \sum_{j=1}^{N_{batch}} f_{ij} \log \hat{f}_{ij} \tag{15}$$

$$Loss^{t\to v} = -\frac{1}{N_{batch}} \sum_{t=1}^{N_{batch}} \sum_{j=1}^{N_{batch}} f_{tj} \log \hat{f}_{tj} \tag{16}$$

$$Loss = \frac{Loss^{v\to t} + Loss^{t\to v}}{2} \tag{17}$$

Particularly, with multilingual text data encoded by our, we pre-train the model with the final loss function mentioned above to obtain the complete model.

4. Experiments

4.1. Experimental settings

Datasets. In this paper, we use the MIMIC-CXR and CheXpert datasets for pre-training the model. Table 1 shows the statistics of the datasets.

Subsequently, we evaluate the performance of the pre-trained model using five classified datasets, CheXpert-5x200, MIMIC-5x200, COVID, RSNA Pneumonia, and MIMIC&CheXpert-1000-multi series.

- **CheXpert-5x200** is a multi-class categorical dataset extracted from CheXpert, containing 200 positive images from each of the five main classes (Atelectasis, Cardiomegaly, Consolidation, Edema, Pleura Effusion) specifically for the CheXpert competition task.
- **MIMIC-5x200** is similar to the CheXpert-5x200 described above, 200 images from each of the same five major categories are extracted from the MIMIC-CXR.
- **COVID** [47] is a publicly available X-ray dataset with COVID and non-COVID labels. The ratio of positive to negative cases is approximately 1:1. We extracted 3000 fine-tuned data and 3000 test data for evaluation purposes in a 1:1 ratio.
- **RSNA Pneumonia** [48] is a collection of pneumonia cases from the publicly available NIH chest X-ray database. The dataset is dichotomous, consisting of pneumonia and normal cases. We selected 3000 fine-tuned data and 1000 test data at a ratio of COVID plus or minus 1:1 for task assessment.
- **MIMIC&CheXpert-1000-multi series** is a 21-language multilingual translation of texts by a translator for image–text retrieval tasks.

Above all, the first four datasets are utilized for assessing the performance of medical image classification tasks. Specifically, the CheXpert-5x200 and MIMIC-5x200 are designed to evaluate the baseline performance of the models, while the COVID and RSNA Pneumonia focus on zero-shot and fine-tuned classification tasks. Furthermore, the last dataset is dedicated to evaluating the effectiveness of image–text retrieval tasks.

Baselines. Our baseline models are outlined below:

- **CLIP**: CLIP maps the input image and text to the same vector space using separate encoders for images and text. The model is pre-trained using contrastive learning, allowing it to compare and match image and text features directly.
- **ConVIRT**: ConVIRT focuses on vision-text contrastive learning in the medical domain. It optimizes the visual representation by improving the consistency between real image–text pairs and randomly generated text pairs. We use a text encoder called BioClinicalBERT[3] in the reproduction process.
- **GLoRIA**: GLoRIA learns by matching sub-regions of images with keywords in medical reports through a cross-attention mechanism. This captures important features in images and reports more efficiently.
- **MedCLIP**: MedCLIP utilizes image and text labels to re-pair images and text through decoupled contrastive learning. This approach increases the amount of training data while avoiding false negatives, and also learns more fine-grained feature representations.

We conduct medical image classification tasks on the model to assess its ability to disambiguate uncertain labels. Furthermore, we also perform zero-shot and fine-tuned classification on COVID and RSNA datasets. We compare the four baseline models described above and 3M-CLIP. Additionally, to evaluate the models' multilingual capability, we retrieve images and text in multiple languages. The comparison models used are MedCLIP and 3M-CLIP.

Parameter settings. We obtain the text encoder Med-XLM-RoBERTa by retraining the XLM-RoBERTa model. This is accomplished by using translated medical reports in 20 different languages. The pre-training covers the MLM and TLM processes, enabling the encoder to learn comprehensively in monolingual and multilingual environments. The hidden size is 768, with 12 attention heads and hidden layers. The maximum input sequence length is 514, and the model is trained for 200 epochs. The training of the final Med-XLM-RoBERTa model takes approximately 12 h using two A100 GPUs.

Our model employs Med-XLM-RoBERTa as the text encoder and Swin Transformer as the vision encoder for pre-training. We unify the vision and text encoders to facilitate contrastive learning and set the output linear projection head dimension to 514. For the original images, we first apply preprocessing steps such as random horizontal flipping, color jittering, and random affine transformations, then scale them to 257 × 257 and then randomly crop them to 224 × 224. Other hyperparameters in our model are set as follows: the initialization hyperparameter temperature τ is 0.07 [49], the learning rate is 5e−6, the batch size is 100, the weight decay is 1e−4, the number of epochs is 100, and the warmup is 0.1. The pre-training of the final model takes 7 days to complete on two A100 GPUs.

4.2. Results and analysis

4.2.1. Medical image classification

Routine testing. To evaluate the model's performance, we conduct medical image classification tests on the CheXpert-5x200 and MIMIC-5x200 datasets, with results shown in Table 2. Compared to other baseline

[3] https://huggingface.co/emilyalsentzer/Bio_ClinicalBERT.

Table 2
The ACC results of medical image classification task for CheXpert-5 × 200 and MIMIC-5 × 200 datasets.

ACC	CheXpert-5 × 200	MIMIC-5 × 200
	Routine testing	
CLIP	0.2016	0.1918
ConVIRT	0.4188	0.4018
GLoRIA	0.4328	0.3306
MedCLIP	0.5943	0.4856
3M-CLIP	**0.6402**	**0.5788**
	Continue learning	
3M-CLIP	**0.6448**	**0.6102**

Table 3
The ACC(STD) results for COVID and RSNA datasets on zero-shot image classification tasks. The *pmt* version refers to the customized prompts corresponding to each disease category for each method, to further enhance the model's ability to recognize unseen categories.

ACC(STD)	COVID	RSNA
CLIP	0.5069(0.03)	0.4989(0.01)
$CLIP_{pmt}$	0.5090(<0.01)	0.5055(0.01)
ConVIRT	0.5184(0.01)	0.4731(0.05)
$ConVIRT_{pmt}$	0.6647(0.05)	0.4647(0.08)
GLoRIA	0.7090(0.04)	0.5808(0.08)
$GLoRIA_{pmt}$	0.5702(0.06)	0.4752(0.06)
MedCLIP	0.7828(<0.01)	0.6702(<0.01)
$MedCLIP_{pmt}$	0.7758(<0.01)	0.6849(<0.01)
3M-CLIP	**0.8428**(<**0.01**)	0.8218(<0.01)
$3M\text{-}CLIP_{pmt}$	0.8420(<0.01)	**0.8227**(<**0.01**)

models, our method achieves the best performance overall. Additionally, we perform continued learning experiments, which reveal that the model's performance improves on both datasets, thereby partially validating the effectiveness and robustness of our approach. These experimental results demonstrate that our model has a significant advantage in handling medical image classification tasks.

Zero-shot. To assess the model's generalization ability, we conduct a zero-shot classification task, which evaluates the model's ability to recognize new categories (unseen categories) without direct training data. Then, two datasets are used for zero-shot image classification evaluations: COVID, and RSNA Pneumonia. The image–text encoder facilitates zero-shot prediction by matching the encoded image embeddings with the prompts created for each disease category. Each method has a corresponding prompt ensemble version (abbreviated as *pmt*). We report the accuracy (ACC) mean and standard deviation (STD) as evaluation metrics. Each experiment result is obtained by ten repeat runs on different random seeds, owing to the stochastic nature of the prompt generation process. Table 3 presents the results.

As can be seen from Table 3, our method greatly outperforms all other baselines. The prompt ensemble in 3M-CLIP leads to better performance on RSNA dataset. Notably, even though COVID positive images are not included in the pre-training process, all models achieve the best classification results on this dataset. The possible reason is that the model has strong generalization ability or architecture, which is well-suited to the features of the COVID dataset, enabling them to effectively capture the key information in the data even in this scenario.

Fine-tune. To assess the transferability of the model to a downstream supervised task, we fine-tuned the randomly initialized linear classification head using cross-entropy loss on the training data. The experimental results are presented in Table 4. It can be observed that 3M-CLIP outperforms other baselines on all two datasets consistently.

When comparing Tables 3 and 4, it is evident that the performance of 3M-CLIP in zero-shot classification is comparable to its performance in fine-tune classification, and it even surpasses other supervised learning methods. Moreover, the model outperforms zero-shot classification on the RSNA dataset, whereas the opposite is observed in the COVID dataset. This discrepancy partially accounts for the superiority of zero-shot classification in low-resource scenarios.

Table 4
The ACC results for COVID and RSNA datasets on fine-tuned image classification tasks.

ACC	COVID	RSNA
CLIP	0.5866	0.7303
ConVIRT	0.6983	0.7846
GLoRIA	0.7623	0.7981
MedCLIP	0.7702	0.7998
3M-CLIP	**0.8310**	**0.8736**

4.2.2. Image–text retrieval

To evaluate the model's semantic representations in an image–text retrieval task, we conduct the MIMIC&CheXpert-1000-multi dataset. However, since CheXpert lacks publicly available report data, we follow MedCLIP to pair its image data with reports from the MIMIC-CXR dataset. We conduct five rounds of evaluation, with each round retrieving 100 randomly selected sentences from the MIMIC-CXR dataset for each of the five categories in each language and comparing them against 1000 images. The average of the ACC and Precision@K are utilized as metrics to access the performance of the relevance model. ACC measures the accuracy of the retrieval by determining if the report and image belong to the same category, while Precision@K measures the accuracy of the first K reports retrieved to determine if the report matches the query image's category. For Precision@K metrics, we report the results at ranks of 1, 2, 5, and 10.

The results of ACC for MedCLIP and 3M-CLIP on CheXpert and MIMIC-CXR datasets are presented in Fig. 9. Our model performs exceptionally well in all languages, especially in English environments, which indicates that our method effectively provides the necessary semantic information for text retrieval in multilingual environments. On the other hand, the 3M-CLIP models demonstrate different representational capabilities in different linguistic environments, unlike the MedCLIP model which has a smoother ACC across all languages. One possible explanation is that the text encoder used by 3M-CLIP relies more on the specific characteristics of a language. Alternatively, variations in the quality of translation models for different languages can lead to discrepancies in the quality of training data and thus affect performance.

The comparison results between MedCLIP and 3M-CLIP on the CheXpert and MIMIC-CXR datasets are shown in Figs. 10 and 11, respectively. It is evident from these figures that 3M-CLIP outperforms MedCLIP across all values of K and achieves an overall higher score of approximately 0.55. Furthermore, MedCLIP's overall score on both datasets is around 0.78, while the performance of 3M-CLIP, although varying by data and language, is overall better than MedCLIP. Moreover, in most languages, 3M-CLIP performs better on the MIMIC-CXR dataset than the CheXpert dataset. Fig. 12 visually displays the models' results on different datasets with varying Precision@K values, revealing that MedCLIP's Precision@K remains consistent at around 0.2 on both datasets, with Precision@10 being more stable. An interesting finding is that 3M-CLIP performs better than other K values at K = 10 in most languages for precision, but worse in a few languages. Furthermore, the results of 3M-CLIP on the MIMIC-CXR dataset are significantly better than CheXpert, covering a broader range of multilingual environments. Notably, 3M-CLIP's varied performance in different languages on the two datasets compared to MedCLIP's stable performance further supports the same previous results of their performance on ACC metrics.

4.2.3. Ablation study

To analyze the impact of each component of the model on overall performance, we conduct ablation experiments. The experimental setup is as follows:

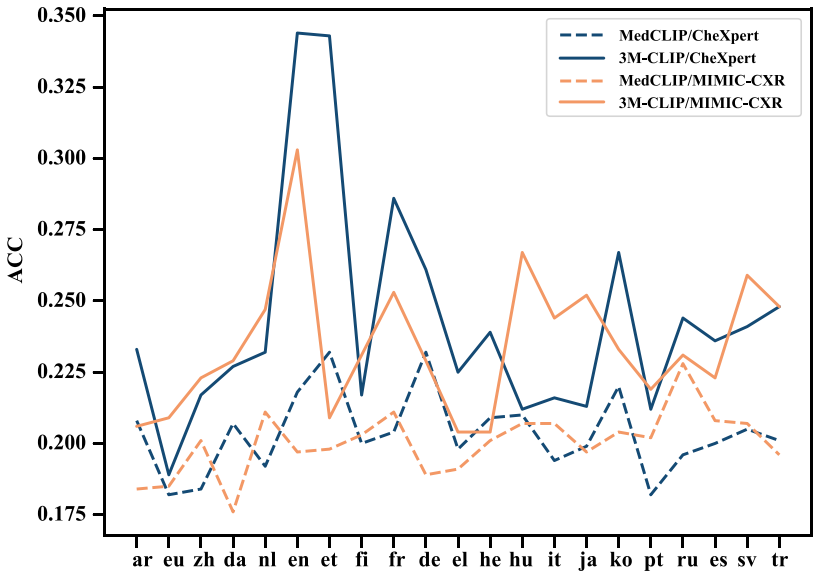


Fig. 9. ACC results for image–text retrieval of MedCLIP and 3M-CLIP on the CheXpert and MIMIC-CXR datasets.

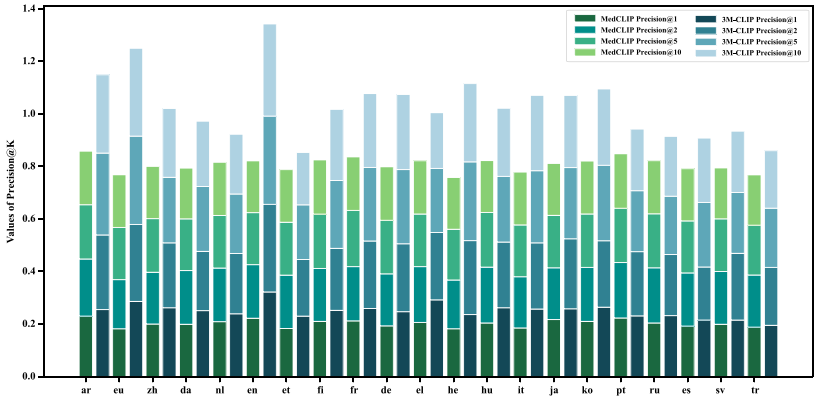


Fig. 10. Precison@K results for image–text retrieval on CheXpert.

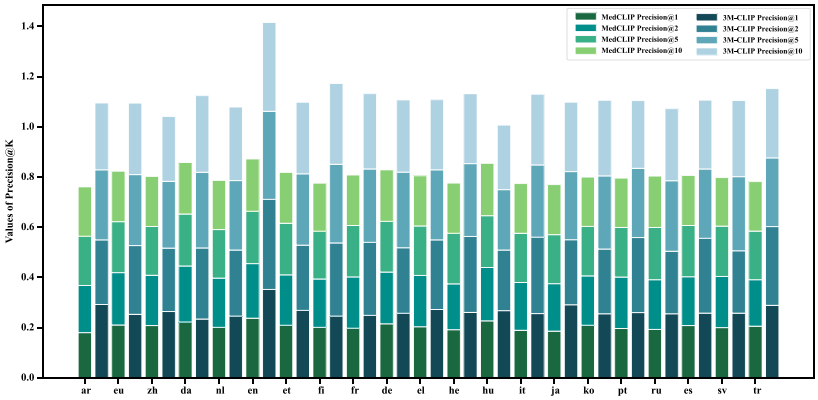
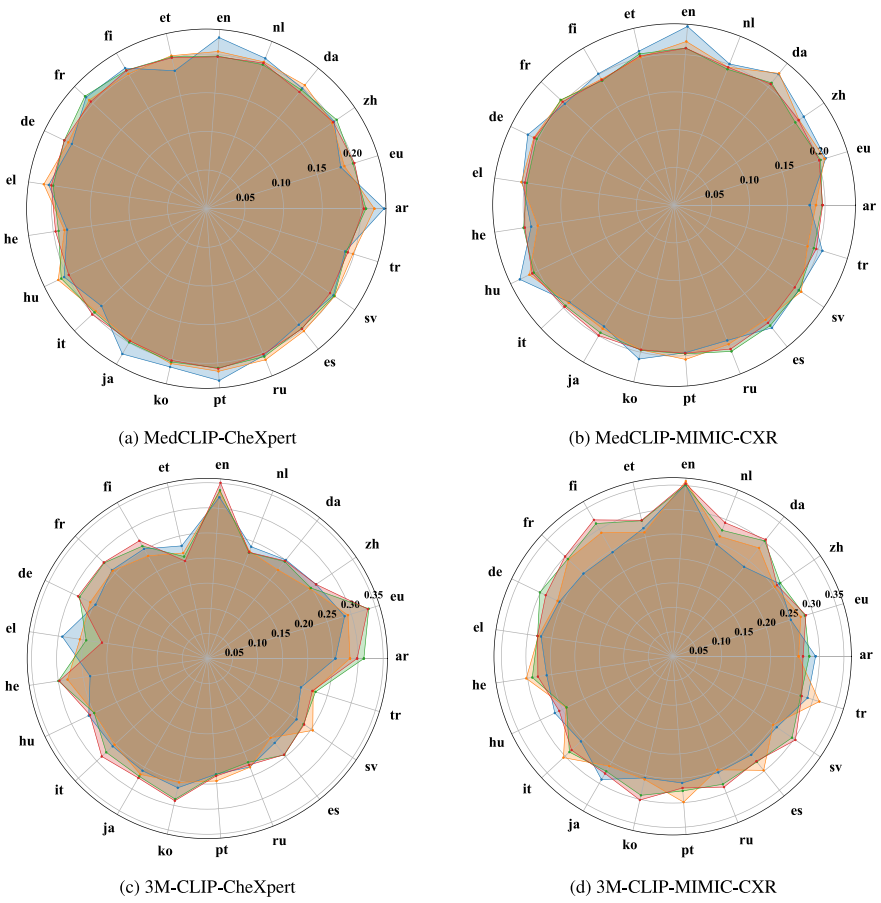


Fig. 11. Precison@K results for image–text retrieval on MIMIC-CXR.

- MedCLIP(bio) refers to using BioClinicalBERT as a monolingual text encoder without changing the labels.
- MedCLIP(xlm) refers to using XLM-RoBERTa as a multilingual text encoder, also without changing the labels.
- MedCLIP(med) refers to using Med-XLM-RoBERTa as a multilingual text encoder without changing the labels.
- MedCLIP(bio+label) indicates using BioClinicalBERT as a monolingual text encoder with label disambiguation applied.
- MedCLIP(xlm+label) denotes using XLM-RoBERTa as a multilingual text encoder with label changes.
- Ours refers to using Med-XLM-RoBERTa as a multilingual text encoder with label changes.



(a) MedCLIP-CheXpert (b) MedCLIP-MIMIC-CXR

(c) 3M-CLIP-CheXpert (d) 3M-CLIP-MIMIC-CXR

Fig. 12. Comparative analysis of model Precison@K on different datasets. (Where the colors represent Blue-Precision@1, Orange-Precison@2, Green-Precision@5, Red-Precision@10).

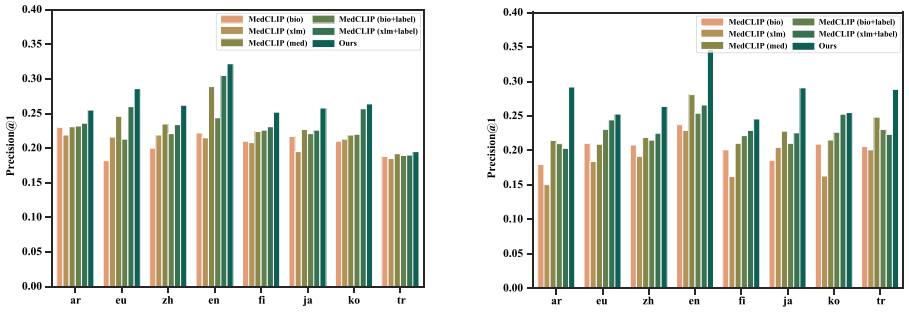
Medical image classification. The experimental results are shown in Table 5. Our model achieves optimal performance overall. Analyzing the performance of each module, we find that the text encoder trained on medical data significantly improves the model's classification accuracy without changing the labels. Specifically, models trained on medical data, such as MedCLIP(bio) and MedCLIP(med), demonstrate significant advantages in classification accuracy compared to those not trained on medical data, like MedCLIP(xlm). For example, MedCLIP(bio) outperforms MedCLIP(xlm) by 15.02% in average classification accuracy over the four datasets, indicating that training on medical data significantly improves model performance even without multilingual support. Even with a smaller parameter size, MedCLIP(bio) still surpasses MedCLIP(xlm), which indicates that training on medical data is more effective in improving performance than simply increasing the model size. Moreover, MedCLIP(med), which combines medical data training and multilingual capability, achieves an average classification accuracy 18.22% higher than MedCLIP(xlm) and 3.2% higher than MedCLIP(bio), further demonstrating the synergistic effect of medical data and multilingual training and validating the crucial role of medical data training in multilingual environments.

Further analysis reveals that the label disambiguation significantly enhances model performance. In all datasets, models with label disambiguation outperform those without this feature in classification accuracy. In particular, our label disambiguation method outperforms the MedCLIP(med) without label disambiguation, achieving an average accuracy improvement of 5.35%. Notably, the effect of label disambiguation even outweighs the improvement of the text encoder enhanced by medical multilingual training, especially when comparing the bio and med series to the no-label and label series. This shows that label disambiguation significantly enhances classification performance, especially in multilingual scenarios. Overall, the results of the ablation experiments demonstrate that each module significantly contributes to the improvement of the model's overall performance, highlighting the importance of model design and the relevance of training data.

Table 5
ACC results of ablation study for medical image classification. The MedCLIP(bio) indicates the use of BioClinicalBERT as a monolingual text encoder without changing the labels. The MedCLIP(xlm) and MedCLIP(med) denote the replacement of the monolingual text encoders with the multilingual text encoders XLM-RoBERTa and Med-XLM-RoBERTa, respectively, while keeping the labels unchanged. The MedCLIP(bio+label) indicates label disambiguation with BioClinicalBERT as a monolingual text encoder. The MedCLIP(xlm+label) and Ours represent encoding with multilingual text encoders XLM-RoBERTa and Med-XLM-RoBERTa while changing the labels.

	CheXpert-5 × 200	MIMIC-5 × 200	COVID	RSNA
MedCLIP(bio)	0.5946	0.4875	0.7758	0.6849
MedCLIP(xlm)	0.4324	0.4424	0.5843	0.4830
MedCLIP(med)	0.5957	0.5038	0.8201	0.7513
MedCLIP(bio+label)	0.6182	0.5230	0.8293	0.7919
MedCLIP(xlm+label)	0.4892	0.5516	0.7827	0.6160
Ours	**0.6402**	**0.5798**	**0.8420**	**0.8227**

Image–text retrieval. We select one typical language from each of the eight language families for comparison, and the experimental results are shown in Fig. 13. In general, our model demonstrates the best performance. Analyzing the performance of each module, we find that in cross-modal interactions, models with integrated medical knowledge, such as MedCLIP(bio) and MedCLIP(med), generally outperform models without integrated medical knowledge, such as MedCLIP(xlm). Furthermore, in all multilingual environments, the MedCLIP(med) model trained on medical data outperforms the multilingual MedCLIP(xlm) model not trained on medical data, with an average accuracy improvement of 3.33% (CheXpert-5x200: 2.41%, MIMIC-5x200: 4.25%).



(a) CheXpert-1000-multi series. (b) MIMIC-1000-multi series.

Fig. 13. Precision@1 results of ablation study for image–text retrieval, evaluated for one typical language from each of the eight language families in the CheXpert-1000-multi series dataset (a) and the MIMIC-1000-multi series dataset (b). The MedCLIP(bio) indicates the use of BioClinicalBERT as a monolingual text encoder without changing the labels. The MedCLIP(xlm) and MedCLIP(med) denote the replacement of the monolingual text encoders with the multilingual text encoders XLM-RoBERTa and Med-XLM-RoBERTa, respectively, while keeping the labels unchanged. The MedCLIP(bio+label) indicates label disambiguation with BioClinicalBERT as a monolingual text encoder. The MedCLIP(xlm+label) and Ours represent encoding with multilingual text encoders XLM-RoBERTa and Med-XLM-RoBERTa while changing the labels.

This suggests that training on medical data significantly influences model performance, and model parameter size may not be the sole determinant of performance.

Further analysis shows that models using label disambiguation methods for preprocessing demonstrate significantly better performance than those without such processing in cross-modal interactions. Particularly in multilingual scenarios, our methods with label disambiguation outperforms the MedCLIP(med) model without it, achieving an average precision@1 improvement of 4.04% (CheXpert-5x200: 2.89%, MIMIC-5x200: 5.20%). These results indicate that label disambiguation significantly enhances performance in medical and multilingual environments. Overall, these results validate the effectiveness of our approach.

Furthermore, to evaluate the robustness and effectiveness of the label disambiguation strategy under different experimental settings and data distributions, we conduct comparative experiments on both the MedCLIP(med) and our proposed model. Specifically, we perform a detailed performance evaluation on five disease classification tasks (Atelectasis, Cardiomegaly, Consolidation, Edema, and Pleural Effusion) using the CheXpert-5x200 and MIMIC-5x200 datasets. By comparing the performance of the two models across different data distributions, we aim to deeply analyze the impact of label disambiguation strategies on model performance.

Fig. 14 compares the accuracy results of two models in the medical image classification task. Our model achieves the highest average accuracy across five disease classification tasks on both the CheXpert-5x200 and MIMIC-5x200 datasets, significantly outperforming the comparison model MedCLIP(med) by 4.19% and 7.51%, respectively. Specifically, in our analysis of individual categories, our model demonstrates significant advantages in four disease categories: Atelectasis(CheXpert-5x200: 0.1691, MIMIC-5x200: 0.2194), Cardiomegaly(CheXpert-5x200: 0.0001, MIMIC-5x200: 0.0943), Consolidation(CheXpert-5x200: 0.2141, MIMIC-5x200: 0.26), and Pleural Effusion(CheXpert-5x200: 0.0121, MIMIC-5x200: 0.0677), except Edema. These results indicate that our proposed label disambiguation strategy can significantly enhance the model's classification performance in most cases.

We also analyze the performance differences of our model across the two datasets. The experimental results show that the overall performance of our model on the CheXpert-5x200 is slightly better than that on the MIMIC-5x200 dataset, which may be due to the higher annotation quality or more balanced data distribution of the CheXpert-5x200. Notably, our model achieves the highest performance on both datasets in the Pleural Effusion classification task(CheXpert-5x200: 0.8516, MIMIC-5x200: 0.8266), further validating its robustness and generalization capability in handling specific disease categories.

In summary, our model demonstrates superior performance in the multi-category medical image classification task, especially with the addition of the label disambiguation strategy, which can effectively improve the model's classification accuracy for complex medical images.

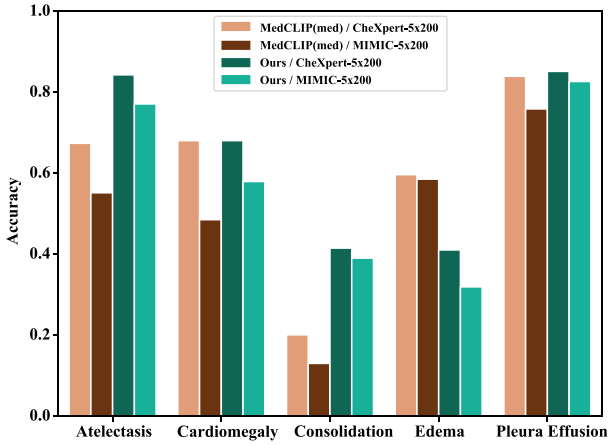


Fig. 14. ACC results of medical image classification for five disease categories by MedCLIP(med) and Ours on Chexpert-5x200 and MIMIC-5x200 datasets.

Fig. 15 shows the Precision@1 results of two models on the image–text retrieval task, covering eight language families in five disease types. Our model achieves an average Precision@1 improvement of 3.82% on the CheXpert-5x200 and 4.55% on the MIMIC-5x200 dataset. Below, we provide a detailed analysis of performance for each disease type:

- Atelectasis. Our model outperforms MedCLIP(med) in Precision@1 across most language families on the CheXpert-5x200 dataset, including Basque, English, Japanese, Korean, and Turkish. The MIMIC-5x200 dataset shows significant superiority in Basque, Chinese, English, Finnish, Japanese, and Turkish, particularly excelling in English(0.7208) and Finnish(0.6195).
- Cardiomegaly. In the CheXpert-5x200 dataset, our model excels in Basque, English, and Finnish. The MIMIC-5x200 dataset shows strong cross-language generalization, outperforming almost all languages except Finnish and Turkish.
- Consolidation. On the CheXpert-5x200 dataset, our model notably surpasses in Finnish. The MIMIC-5x200 outperforms MedCLIP(med) in Arabic, Basque, and Japanese, indicating potential language dependency.
- Edema. In the CheXpert-5x200 dataset, our model outperforms in most languages, especially Arabic and Turkish. The MIMIC-5x200 dataset also surpasses in Japanese, Korean, and Turkish, indicating that performance on Edema may vary by dataset.

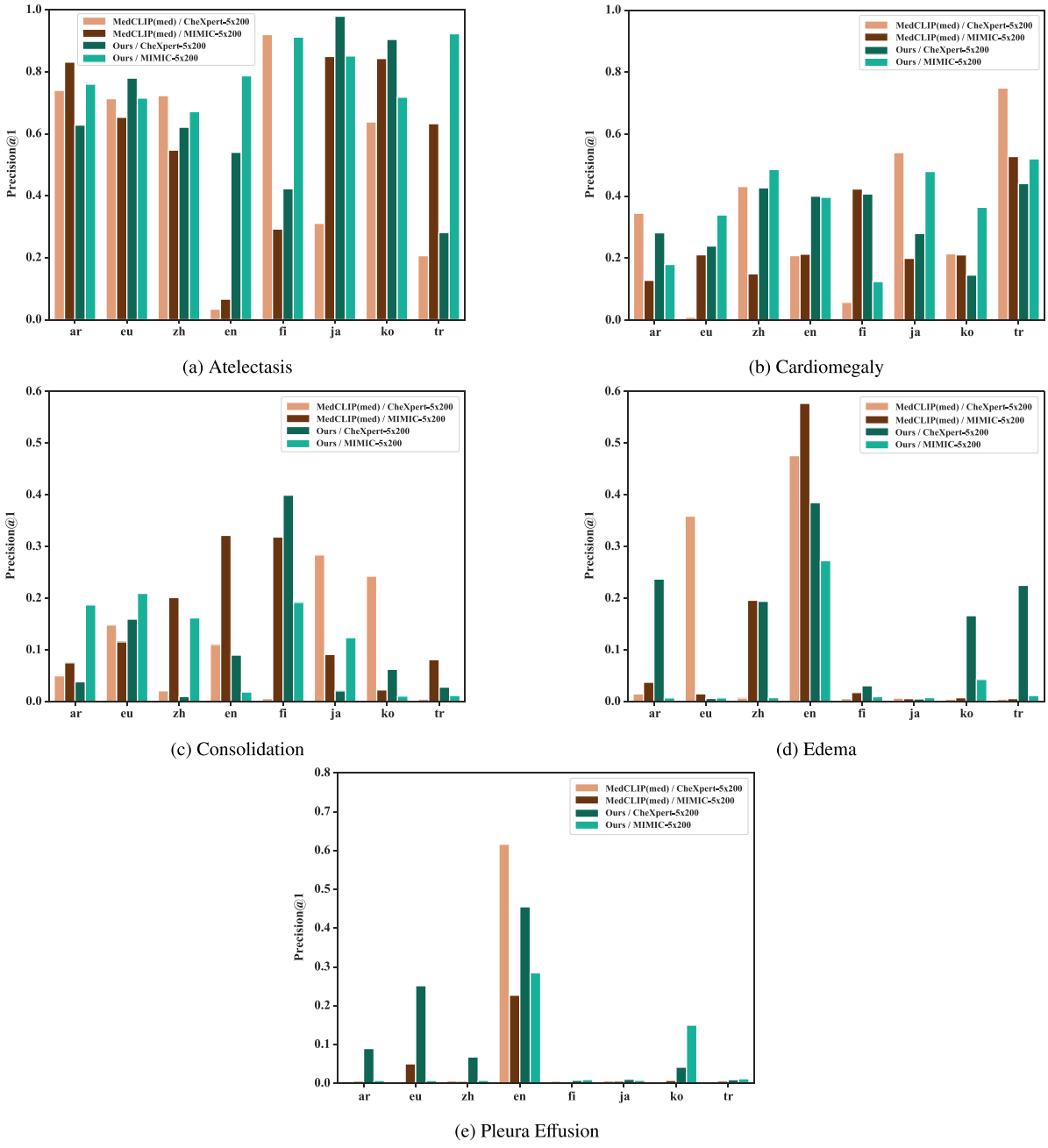


Fig. 15. Precision@1 results from MedCLIP(med) and Ours for image–text retrieval of five disease categories in eight language families on the CheXpert-5x200 and MIMIC-5x200 datasets.

- Pleural Effusion. On the CheXpert-5x200 dataset, our model outperforms in all languages except English, especially excelling in Basque. Similarly, in the MIMIC-5x200 dataset, our model surpasses in all languages except Basque. Notably, the model shows the widest precision range in Pleural Effusion, highlighting the effectiveness of the label disambiguation strategy. While improvements in English are minimal, performance in Basque and Turkish significantly exceeds that of the baseline model.

Our model outperforms MedCLIP(med) across most disease types and languages, particularly excelling in Atelectasis and Pleural Effusion, which highlights the significant advantage of our label disambiguation strategy in multilingual image–text retrieval tasks. Furthermore, the model demonstrates the widest performance range in Pleural Effusion, further validating the effectiveness of the label disambiguation strategy.

4.2.4. *Case study*

To facilitate an intuitive understanding of retrieval performance, we have extracted two instances from the MIMIC-1000-zh test dataset to compare the retrieval efficacy of the MedCLIP and 3M-CLIP models, as shown in Fig. 16. Upon examination of these instances, it is observed that when provided with image inputs, the 3M-CLIP model demonstrated superior performance in retrieving labels that align with the report text corresponding to the input image labels, as compared to the MedCLIP model. Nevertheless, the retrieved labels do not invariably match the original report text associated with the image. The reason for this discrepancy may be that the report texts selected during the testing phase were chosen randomly based on different disease categories, so this can lead to differences in the retrieved report texts, but overall, it does not affect the labeling categories of the report texts.

Notably, the retrieval process has occasionally yielded results that deviate from the expected outcomes, as exemplified in Fig. 17. This

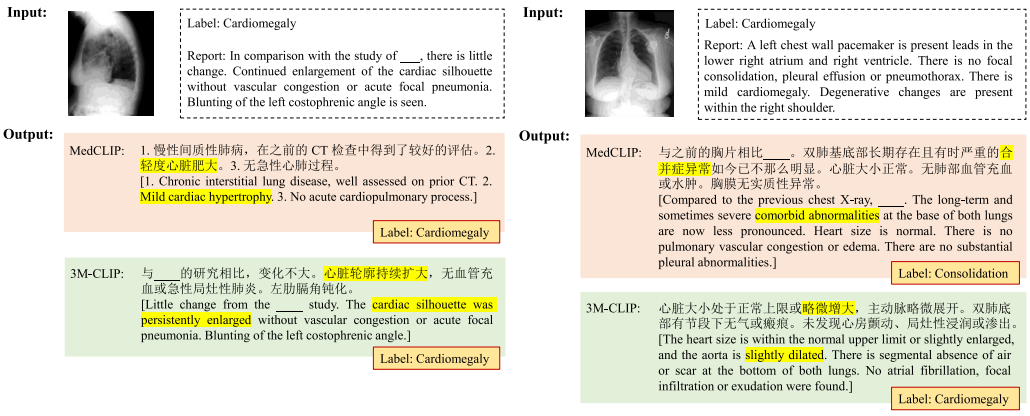


Fig. 16. Examples of MedCLIP and 3M-CLIP for image–text retrieval. The input is an image, and the dotted line box shows the label of the image and the corresponding report text. The output is the report text with the highest relevance to the image, with the label corresponding to the output text in the red box at the bottom right.

Fig. 17. Examples of 3M-CLIP for image–text retrieval. The top is a true case and the bottom is a false case.

divergence may stem from a deficiency in the model's comprehension of the semantic interplay between visual and textual data. Additionally, the model's capacity for feature extraction from visual inputs may be suboptimal, which could result in a diminished ability to amalgamate visual information with its textual counterpart effectively. In subsequent research endeavors, it would be beneficial to focus on enhancing the model's semantic comprehension capabilities and refining its feature extraction mechanisms. Such advancements have the potential to significantly augment the precision of the retrieval process and reduce the occurrence of erroneous results.

5. Discussion

We adopt a strategic augmentation approach to enrich multilingual data in the medical field and utilize the data to train a multilingual medical text encoder, enhancing the model's ability to handle multilingual medical data. We introduce a targeted label disambiguation method to establish robust semantic similarity, thereby aligning the semantics of visual and textual modalities and enhancing cross-modal interactions. We construct a refined multilingual multimodal medical pre-training model, 3M-CLIP, aimed at learning generalized representations of multilingual multimodality in the medical domain. Extensive experiments conducted on four medical datasets demonstrate that 3M-CLIP establishes new state-of-the-art levels in medical image classification and medical image–text retrieval. Our results indicate that 3M-CLIP exhibits strong transferability in medical image classification tasks and validates its effectiveness in multilingual image–text retrieval tasks.

Understanding the factors that influence a model's ability to process multilingual medical data is crucial to our field. Although this is not the primary focus of this study, we still provide some insights that merit further validation. In our observations of limited medical data, we find that high-quality multilingual data works better than low-quality data. This indicates that the quality of training data plays a key role in model performance, particularly in multilingual medical applications. Furthermore, considering the professionalism of medical terms and

Table A.6
List of 21 languages.

Name	Code	Family	Lang
Arabic	ar	Afro-A	arb_Arab
Basque	eu	Vasconic	eus_Latn
Chinese	zh	Sino-T	zho_Hans
Danish	da	Indo-E	dan_Latn
Dutch	nl	Indo-E	nld_Latn
English	en	Indo-E	eng_Latn
Estonian	et	Uralic	est_Latn
Finnish	fi	Uralic	fin_Latn
French	fr	Indo-E	fra_Latn
German	de	Indo-E	deu_Latn
Greek	el	Indo-E	ell_Grek
Hebrew	he	Afro-A	heb_Hebr
Hungarian	hu	Uralic	hun_Latn
Italian	it	Indo-E	ita_Latn
Japanese	ja	Japonic	jpn_Jpan
Korean	ko	Koreanic	kor_Hang
Portuguese	pt	Indo-E	por_Latn
Russian	ru	Indo-E	rus_Cyrl
Spanish	es	Indo-E	spa_Latn
Swedish	sv	Indo-E	swe_Latn
Turkish	tr	Turkic	tur_Latn

the scarcity of multilingual data, enhancing the translation quality of multilingual medical data is particularly important. To validate these results, future research should add accurate medical translations and assess how data quality affects model performance. This will help us understand how the quality of multilingual medical translation affects model performance. We will continue to refine our methods. Future research may focus on improving the quality of multilingual medical data.

6. Conclusion

In this paper, we propose a novel multilingual multimodal medical pre-trained model called 3M-CLIP. Specifically, to mitigate the challenges of scarcity and imbalanced data distribution, 3M-CLIP employs a strategic augmentation method by expanding the MIMIC-CXR report dataset to 20 distinct languages using machine translation techniques. Furthermore, 3M-CLIP develops a targeted label disambiguation technique to address the labeling noise within decoupled contrastive learning. In particular, it categorizes and refines uncertain phrases within the clinical reports based on disease type. To this end, we can construct a refined multilingual multimodal medical pre-training model that aspires to learn a generalized representation of multilingual multimodality in the medical domain. Based on this, we conduct the two sub-tasks. The experiment demonstrates that our model improves on average 10% on zero-shot image classification and has good transferability. Additionally, it performs exceptionally well in the multilingual image–text retrieval task. The natural future is that we can enhance the quality of different languages in the translation process, thereby improving the model's ability to generalize across languages.

CRediT authorship contribution statement

Qiyuan Li: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Data curation, Conceptualization. **Chen Qiu:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Haijiang Liu:** Writing – review & editing, Software. **Jinguang Gu:** Writing – review & editing, Supervision, Funding acquisition. **Dan Luo:** Writing – review & editing.

Table A.7
Specific scoring criteria for GLM-4-flash and human. We assess the quality of the translation data in terms of accuracy, fluency and contextual consistency, then score them according to different scoring criteria, and finally sum them up.

Type	Scoring criteria	Score
Accuracy	The translation is completely faithful to the original text, with no omissions, errors, or misinterpretations of information, and the terminology and factual information is completely correct.	5
	The translation is very close to the original text, with only a very few minor errors or omissions that do not affect overall understanding.	4
	The translation is generally accurate, but there are some errors or omissions that may affect the full understanding of the original text.	3
	The translation is partially accurate with multiple errors or omissions that significantly affect understanding of the original text.	2
	The translation is grossly inaccurate, with numerous errors or omissions that make it difficult for the reader to understand the message of the original text.	1
Fluency	The translation is natural and fluent, fully compliant with the target language, and reads very smoothly without any hard, direct translations.	5
	The translation is relatively fluent, with only a very few expressions that are not natural or are slightly stiff.	4
	The translation is fluent, but several expressions are not natural or hard, affecting the reading experience.	3
	The translation is not fluent enough, with several obvious language errors or unnatural expressions that significantly affect the reading.	2
	The translation is very poor, with many linguistic errors and unnatural expressions, making it difficult to read.	1
Contextual consistency	The translation takes full account of the context of the original text, and the expression of the translated text in the target language fully corresponds to the needs of the context.	5
	The translation takes the context into account well, with only a few areas where the translation fails to fully adapt to the context.	4
	The translation generally fits the context but fails to fully adapt to the context in several places, which may affect comprehension.	3
	The translation fails to take the context into account well, with some places where the expression does not fit the context, significantly affecting understanding.	2
	The translation completely ignores the context and there are a large number of expressions that do not fit the context, making it difficult for the reader to understand the contextual meaning of the original text.	1

Table A.8
List of 14 main types of findings.

Finding types
Atelectasis
Cardiomegaly
Consolidation
Edema
Enlarged cardiomediastinum
Fracture
Lung lesion
Lung opacity
No finding
Pleural effusion
Pleural other
Pneumonia
Pneumothorax
Support devices

Data and code availability

Three of the four datasets are public. The remaining dataset may be available from PhysioNet upon the request of qualified parties. We will release the code at https://github.com/vicky-yuan/3m-clip.

Declaration of competing interest

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

Acknowledgments

This work was supported by the National Key Research and Development Program of China under Grants 2022YFC3300801, the Knowledge Innovation Program of Wuhan-Shuguang Project under Grants 2023010201020409, and the Natural Science Foundation of Hubei Province (CN) under Grants JCZRQN202500208. This study utilized the computational resources provided by Wuhan Digital Computing Technology Co. Gratitude to the anonymous reviewers and the associate editor for their insightful comments.

Appendix

See Tables A.6–A.8.

Data availability

Data will be made available on request.

References

[1] F. Liu, E. Bugliarello, E.M. Ponti, S. Reddy, N. Collier, D. Elliott, Visually grounded reasoning across languages and cultures, in: M. Moens, X. Huang, L. Specia, S.W. Yih (Eds.), Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021, Association for Computational Linguistics, 2021, pp. 10467–10485, URL: https://doi.org/10.18653/v1/2021.emnlp-main.818.

[2] C. Qiu, D. Oneata, E. Bugliarello, S. Frank, D. Elliott, Multilingual multimodal learning with machine translated text, in: Y. Goldberg, Z. Kozareva, Y. Zhang (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022, Association for Computational Linguistics, 2022, pp. 4178–4193, URL: https://doi.org/10.18653/v1/2022.findings-emnlp.308.

[3] H. Yang, X. Liu, C.D. Wang, FinGPT: Open-source financial large language models, in: FinLLM Symposium at IJCAI 2023, 2023, URL: https://doi.org/10.2139/ssrn.4489826.

[4] F. Liu, T. Zhu, X. Wu, B. Yang, C. You, C. Wang, L. Lu, Z. Liu, Y. Zheng, X. Sun, Y. Yang, L.A. Clifton, D.A. Clifton, A medical multimodal large language model for future pandemics, Npj Digit. Med. 6 (2023) URL: https://doi.org/10.1038/s41746-023-00952-2.

[5] Y. Dan, Z. Lei, Y. Gu, Y. Li, J. Yin, J. Lin, L. Ye, Z. Tie, Y. Zhou, Y. Wang, A. Zhou, Z. Zhou, Q. Chen, J. Zhou, L. He, X. Qiu, EduChat: A large-scale language model-based chatbot system for intelligent education, 2023, CoRR abs/2308.02773. URL: https://doi.org/10.48550/arXiv.2308.02773.

[6] S. Bao, H. He, F. Wang, H. Wu, H. Wang, W. Wu, Z. Wu, Z. Guo, H. Lu, X. Huang, X. Tian, X. Xu, Y. Lin, Z. Niu, PLATO-XL: exploring the large-scale pre-training of dialogue generation, in: Y. He, H. Ji, Y. Liu, S. Li, C. Chang, S. Poria, C. Lin, W.L. Buntine, M. Liakata, H. Yan, Z. Yan, S. Ruder, X. Wan, M. Arana-Catania, Z. Wei, H. Huang, J. Wu, M. Day, P. Liu, R. Xu (Eds.), Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022, Online Only, November 20-23, 2022, Association for Computational Linguistics, 2022, pp. 107–118, URL: https://aclanthology.org/2022.findings-aacl.10.

[7] F. Ren, N. An, Q. Ma, L. Hei, TechGPT: Technology-oriented generative pretrained transformer, 2023, https://github.com/neukg/TechGPT.

[8] A. Healthcare, 2023 Healthcare World Languages Index, Technical Report, AMN Healthcare, 2023.

[9] P. Qiu, C. Wu, X. Zhang, W. Lin, H. Wang, Y. Zhang, Y. Wang, W. Xie, Towards building multilingual language model for medicine, Nat. Commun. 15 (1) (2024) 8384, URL: https://doi.org/10.1038/s41467-024-52417-z.

[10] X. Wang, N. Chen, J. Chen, Y. Hu, Y. Wang, X. Wu, A. Gao, X. Wan, H. Li, B. Wang, Apollo: A lightweight multilingual medical LLM towards democratizing medical AI to 6B people, 2024, CoRR abs/2403.03640. URL: https://doi.org/10.48550/arXiv.2403.03640.

[11] Y. Xu, Deep learning in multimodal medical image analysis, in: H. Wang, S. Siuly, R. Zhou, F. Martín-Sánchez, Y. Zhang, Z. Huang (Eds.), Health Information Science - 8th International Conference, HIS 2019, Xi'an, China, October 18-20, 2019, Proceedings, in: Lecture Notes in Computer Science, 11837, Springer, 2019, pp. 193–200, URL: https://doi.org/10.1007/978-3-030-32962-4_18.

[12] A. Radford, J.W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning transferable visual models from natural language supervision, in: M. Meila, T. Zhang (Eds.), Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event, in: Proceedings of Machine Learning Research, 139, PMLR, 2021, pp. 8748–8763, URL: http://proceedings.mlr.press/v139/radford21a.html.

[13] Y. Zhang, H. Jiang, Y. Miura, C.D. Manning, C.P. Langlotz, Contrastive learning of medical visual representations from paired images and text, in: Z.C. Lipton, R. Ranganath, M.P. Sendak, M.W. Sjoding, S. Yeung (Eds.), Proceedings of the Machine Learning for Healthcare Conference, MLHC 2022, 5-6 August 2022, Durham, NC, USA, in: Proceedings of Machine Learning Research, vol. 182, PMLR, 2022, pp. 2–25, URL: https://proceedings.mlr.press/v182/zhang22a.html.

[14] S. Huang, L. Shen, M.P. Lungren, S. Yeung, Gloria: A multimodal global-local representation learning framework for label-efficient medical image recognition, in: 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021, IEEE, 2021, pp. 3922–3931, URL: https://doi.org/10.1109/ICCV48922.2021.00391.

[15] Z. Wang, Z. Wu, D. Agarwal, J. Sun, Medclip: Contrastive learning from unpaired medical images and text, in: Y. Goldberg, Z. Kozareva, Y. Zhang (Eds.), Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022, Association for Computational Linguistics, 2022, pp. 3876–3887, URL: https://doi.org/10.18653/v1/2022.emnlp-main.256.

[16] A.E. Johnson, T.J. Pollard, S.J. Berkowitz, N.R. Greenbaum, M.P. Lungren, C. ying Deng, R.G. Mark, S. Horng, MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports, Sci. Data 6 (1) (2019) 317.

[17] T. Pires, E. Schlinger, D. Garrette, How multilingual is multilingual bert? in: A. Korhonen, D.R. Traum, L. Màrquez (Eds.), Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers, Association for Computational Linguistics, 2019, pp. 4996–5001, URL: https://doi.org/10.18653/v1/p19-1493.

[18] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised cross-lingual representation learning at scale, in: D. Jurafsky, J. Chai, N. Schluter, J.R. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020, Association for Computational Linguistics, 2020, pp. 8440–8451, URL: https://doi.org/10.18653/v1/2020.acl-main.747.

[19] A. Fan, S. Bhosale, H. Schwenk, Z. Ma, A. El-Kishky, S. Goyal, M. Baines, O. Celebi, G. Wenzek, V. Chaudhary, N. Goyal, T. Birch, V. Liptchinsky, S. Edunov, M. Auli, A. Joulin, Beyond english-centric multilingual machine translation, J. Mach. Learn. Res. 22 (2021) 107:1–107:48, URL: http://jmlr.org/papers/v22/20-1307.html.

[20] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, L. Zettlemoyer, Multilingual denoising pre-training for neural machine translation, Trans. Assoc. Comput. Linguist. 8 (2020) 726–742, URL: https://doi.org/10.1162/tacl_a_00343.

[21] Q. Zheng, X. Xia, X. Zou, Y. Dong, S. Wang, Y. Xue, Z. Wang, L. Shen, A. Wang, Y. Li, T. Su, Z. Yang, J. Tang, Codegeex: A pre-trained model for code generation with multilingual benchmarking on humaneval-x, in: Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2023, pp. 5673–5684, URL: https://doi.org/10.1145/3580305.3599790.

[22] X. Wang, G. Chen, G. Qian, P. Gao, X. Wei, Y. Wang, Y. Tian, W. Gao, Large-scale multi-modal pre-trained models: A comprehensive survey, Mach. Intell. Res. 20 (4) (2023) 447–482, URL: https://doi.org/10.1007/s11633-022-1410-8.

[23] F. Li, H. Zhang, Y. Zhang, S. Liu, J. Guo, L.M. Ni, P. Zhang, L. Zhang, Vision-language intelligence: Tasks, representation learning, and large models, 2022, CoRR arXiv:2203.01922. URL: https://doi.org/10.48550/arXiv.2203.01922.

[24] Y. Du, Z. Liu, J. Li, W.X. Zhao, A survey of vision-language pre-trained models, in: L.D. Raedt (Ed.), Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022, ijcai.org, 2022, pp. 5436–5443, URL: https://doi.org/10.24963/ijcai.2022/762.

[25] X. Qiu, T. Sun, Y. Xu, Y. Shao, N. Dai, X. Huang, Pre-trained models for natural language processing: A survey, Sci. China Technol. Sci. 63 (10) (2020) 1872–1897.

[26] C. Jia, Y. Yang, Y. Xia, Y. Chen, Z. Parekh, H. Pham, Q.V. Le, Y. Sung, Z. Li, T. Duerig, Scaling up visual and vision-language representation learning with noisy text supervision, in: M. Meila, T. Zhang (Eds.), Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event, in: Proceedings of Machine Learning Research, vol. 139, PMLR, 2021, pp. 4904–4916, URL: http://proceedings.mlr.press/v139/jia21b.html.

[27] H. Bao, L. Dong, S. Piao, F. Wei, BEiT: BERT pre-training of image transformers, in: The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022, OpenReview.net, 2022, URL: https://openreview.net/forum?id=p-BhZSz59o4.

[28] L.H. Li, M. Yatskar, D. Yin, C. Hsieh, K. Chang, VisualBERT: A simple and performant baseline for vision and language, 2019, CoRR abs/1908.03557. URL: http://arxiv.org/abs/1908.03557.

[29] W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, J. Dai, VL-BERT: pre-training of generic visual-linguistic representations, in: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020, OpenReview.net, 2020, URL: https://openreview.net/forum?id=SygXPaEYvH.

[30] J. Lu, D. Batra, D. Parikh, S. Lee, Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks, in: H.M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E.B. Fox, R. Garnett (Eds.), Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada, 2019, pp. 13–23, URL: https://proceedings.neurips.cc/paper/2019/hash/c74d97b01eae257e44aa9d5bade97baf-Abstract.html.

[31] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, I. Sutskever, Zero-shot text-to-image generation, in: M. Meila, T. Zhang (Eds.), Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event, in: Proceedings of Machine Learning Research, vol. 139, PMLR, 2021, pp. 8821–8831, URL: http://proceedings.mlr.press/v139/ramesh21a.html.

[32] X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei, Y. Choi, J. Gao, Oscar: Object-semantics aligned pre-training for vision-language tasks, in: A. Vedaldi, H. Bischof, T. Brox, J. Frahm (Eds.), Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXX, in: Lecture Notes in Computer Science, vol. 12375, Springer, 2020, pp. 121–137, URL: https://doi.org/10.1007/978-3-030-58577-8_8.

[33] L. Yao, R. Huang, L. Hou, G. Lu, M. Niu, H. Xu, X. Liang, Z. Li, X. Jiang, C. Xu, FILIP: fine-grained interactive language-image pre-training, in: The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022, OpenReview.net, 2022, URL: https://openreview.net/forum?id=cpDhcsEDC2.

[34] S. Zhang, Y. Xu, N. Usuyama, H. Xu, J. Bagga, R. Tinn, S. Preston, R. Rao, M. Wei, N. Valluri, C. Wong, A. Tupini, Y. Wang, M. Mazzola, S. Shukla, L. Liden, J. Gao, M.P. Lungren, T. Naumann, S. Wang, H. Poon, BiomedCLIP: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs, 2024, arXiv:2303.00915. URL: https://arxiv.org/abs/2303.00915.

[35] Y. Wang, G. Wang, UMCL: unified medical image-text-label contrastive learning with continuous prompt, in: X. Jiang, H. Wang, R. Alhajj, X. Hu, F. Engel, M. Mahmud, N. Pisanti, X. Cui, H. Song (Eds.), IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2023, Istanbul, Turkiye, December 5-8, 2023, IEEE, 2023, pp. 2285–2289, URL: https://doi.org/10.1109/BIBM58861.2023.10386034.

[36] B. Liu, D. Lu, D. Wei, X. Wu, Y. Wang, Y. Zhang, Y. Zheng, Improving medical vision-language contrastive pretraining with semantics-aware triage, IEEE Trans. Med. Imaging 42 (12) (2023) 3579–3589, URL: https://doi.org/10.1109/TMI.2023.3294980.

[37] Z. Lin, E. Bas, K.Y. Singh, G. Swaminathan, R. Bhotika, Relaxing contrastiveness in multimodal representation learning, in: IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2023, Waikoloa, HI, USA, January 2-7, 2023, IEEE, 2023, pp. 2226–2235, URL: https://doi.org/10.1109/WACV56688.2023.00226.

[38] R. Wang, Y. Duan, J. Li, P. Pang, T. Tan, XrayGLM: The first Chinese medical multimodal model that chest radiographs summarization, 2023, https://github.com/WangRongsheng/XrayGLM.

[39] M.R. Costa-jussà, J. Cross, O. Çelebi, M. Elbayad, K. Heafield, K. Heffernan, E. Kalbassi, J. Lam, D. Licht, J. Maillard, A.Y. Sun, S. Wang, G. Wenzek, A. Youngblood, B. Akula, L. Barrault, G.M. Gonzalez, P. Hansanti, J. Hoffman, S. Jarrett, K.R. Sadagopan, D. Rowe, S. Spruit, C. Tran, P. Andrews, N.F. Ayan, S. Bhosale, S. Edunov, A. Fan, C. Gao, V. Goswami, F. Guzmán, P. Koehn, A. Mourachko, C. Ropers, S. Saleem, H. Schwenk, J. Wang, No language left behind: Scaling human-centered machine translation, 2022, CoRR abs/2207.04672. URL: https://doi.org/10.48550/arXiv.2207.04672.

[40] M. Zhang, O. Press, W. Merrill, A. Liu, N.A. Smith, How language model hallucinations can snowball, in: Forty-First International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024, OpenReview.net, 2024, URL: https://openreview.net/forum?id=FPlaQyAGHu.

[41] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: 2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021, IEEE, 2021, pp. 9992–10002, URL: https://doi.org/10.1109/ICCV48922.2021.00986.

[42] A.R. Aronson, F. Lang, An overview of MetaMap: historical perspective and recent advances, J. Am. Med. Inform. Assoc. 17 (3) (2010) 229–236, URL: https://doi.org/10.1136/jamia.2009.002733.

[43] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R.L. Ball, K.S. Shpanskaya, J. Seekins, D.A. Mong, S.S. Halabi, J.K. Sandberg, R. Jones, D.B. Larson, C.P. Langlotz, B.N. Patel, M.P. Lungren, A.Y. Ng, CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison, in: The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, the Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, the Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019, AAAI Press, 2019, pp. 590–597, URL: https://doi.org/10.1609/aaai.v33i01.3301590.

[44] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, D. Krishnan, Supervised contrastive learning, Adv. Neural Inf. Process. Syst. 33 (2020) 18661–18673.

[45] M. Zheng, F. Wang, S. You, C. Qian, C. Zhang, X. Wang, C. Xu, Weakly supervised contrastive learning, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10042–10051.

[46] Z. Wang, J. Sun, Transtab: Learning transferable tabular transformers across tables, Adv. Neural Inf. Process. Syst. 35 (2022) 2902–2915.

[47] T. Rahman, A. Khandakar, Y. Qiblawey, A.M. Tahir, S. Kiranyaz, S.B.A. Kashem, M.T. Islam, S. Al-Máadeed, S.M. Zughaier, M.S. Khan, M.E.H. Chowdhury, Exploring the effect of image enhancement techniques on COVID-19 detection using chest X-ray images, Comput. Biol. Med. 132 (2021) 104319, URL: https://doi.org/10.1016/j.compbiomed.2021.104319.

[48] G. Shih, C.C. Wu, S.S. Halabi, M.D. Kohli, L.M. Prevedello, T.S. Cook, A. Sharma, J.K. Amorosa, V. Arteaga, M. Galperin-Aizenberg, R.R. Gill, M.C. Godoy, S. Hobbs, J. Jeudy, A. Laroia, P.N. Shah, D. Vummidi, K. Yaddanapudi, A. Stein, Augmenting the national institutes of health chest radiograph dataset with expert annotations of possible pneumonia, Radiol.: Artif. Intell. 1 (1) (2019) e180041.

[49] F. Wang, H. Liu, Understanding the behaviour of contrastive loss, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, Virtual, June 19-25, 2021, Computer Vision Foundation / IEEE, 2021, pp. 2495–2504, URL: https://doi.org/10.1109/CVPR46437.2021.00252.

Qiyuan Li received her Bachelor of Engineering degree in Computer Science and Technology from Wuhan University of Science and Technology in 2019. She is a Ph.D. student in the School of Computer Science and Technology at Wuhan University of Science and Technology. Her research interests include medical multimodality, knowledge graphs, and deep learning.

Chen Qiu received the B.Sc. and Ph.D. degrees from the China University of Geosciences, Wuhan, China, in 2014 and 2020, respectively. She is a Lecturer at the School of Science and Technology, Wuhan University of Science and Technology, Wuhan. Her research interests include question answering and multilingual multimodal representation learning.

Haijiang Liu received his Software Engineering bachelor's degree from the Wuhan University of Science and Technology in 2023. Currently, he is a Ph.D. candidate in the Section for Natural Language Processing at Wuhan University of Science and Technology. His research interests include multilingual and cross-cultural natural language processing.

Jinguang Gu received his Ph.D. from Wuhan University in 2005. He is a professor and doctoral supervisor at Wuhan University of Science and Technology. His research interests include distributed computing and knowledge graphs. He has published over 100 papers and chaired more than 10 projects at provincial and ministerial levels.

Dan Luo received his Bachelor's degree from Wuhan University of Science and Technology in 2011, and a Master's degree from Huazhong University of Science and Technology in 2014. He is a Ph.D. candidate from Lehigh University, Bethlehem, USA. His research interests include information retrieval and data mining.