Computers in Biology and Medicine 159 (2023) 106962

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

# Enhancing thoracic disease detection using chest X-rays from PubMed Central Open Access

Mingquan Lin [a], Bojian Hou [b], Swati Mishra [c], Tianyuan Yao [d], Yuankai Huo [d], Qian Yang [c], Fei Wang [a], George Shih [e], Yifan Peng [a,*]

[a] Department of Population Health Sciences, Weill Cornell Medicine, New York, USA
[b] Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, USA
[c] Department of Information Science, Cornell University, New York, USA
[d] Department of Computer Science, Vanderbilt University, Nashville, TN, USA
[e] Department of Radiology, Weill Cornell Medicine, New York, USA

ARTICLE INFO

*Keywords:*
PubMed
Artificial intelligence
Chest X-ray

ABSTRACT

Large chest X-rays (CXR) datasets have been collected to train deep learning models to detect thorax pathology on CXR. However, most CXR datasets are from single-center studies and the collected pathologies are often imbalanced. The aim of this study was to automatically construct a public, weakly-labeled CXR database from articles in PubMed Central Open Access (PMC-OA) and to assess model performance on CXR pathology classification by using this database as additional training data. Our framework includes text extraction, CXR pathology verification, subfigure separation, and image modality classification. We have extensively validated the utility of the automatically generated image database on thoracic disease detection tasks, including Hernia, Lung Lesion, Pneumonia, and pneumothorax. We pick these diseases due to their historically poor performance in existing datasets: the NIH-CXR dataset (112,120 CXR) and the MIMIC-CXR dataset (243,324 CXR). We find that classifiers fine-tuned with additional PMC-CXR extracted by the proposed framework consistently and significantly achieved better performance than those without (e.g., Hernia: 0.9335 vs 0.9154; Lung Lesion: 0.7394 vs. 0.7207; Pneumonia: 0.7074 vs. 0.6709; Pneumothorax 0.8185 vs. 0.7517, all in AUC with $p < 0.0001$) for CXR pathology detection. In contrast to previous approaches that manually submit the medical images to the repository, our framework can automatically collect figures and their accompanied figure legends. Compared to previous studies, the proposed framework improved subfigure segmentation and incorporates our advanced self-developed NLP technique for CXR pathology verification. We hope it complements existing resources and improves our ability to make biomedical image data findable, accessible, interoperable, and reusable.

## 1. Introduction

To improve prediction accuracy by artificial intelligence (AI), large chest X-ray (CXR) databases have been collected to train sophisticated deep learning models [1–5]. However, despite their well-documented successes, some critical challenges remain to limit the performance of the algorithms [6–8]. First, most CXR datasets were drawn from single-center studies. For example, NIH-CXR was collected from the NIH Clinical Center [2], and MIMIC-CXR was collected from Beth Israel Deaconess Medical Center [1]. Second, labels collected from existing datasets only focus on a few diseases of interest. A database containing many images for comprehensive concepts classes, including rare diseases, is highly demanded but less studied [9]. Finally, institutional policy often restricts liberal redistribution and reuse of important datasets. A recent example is the COVID-19 CXR dataset. The largest open-source database was launched around eight months after the outbreak of COVID-19 [10]. For these reasons, there is a critical need to quickly create a public CXR database to facilitate the development of advanced image analysis tools and decision support algorithms [11].

To reduce this barrier, we investigate a framework that can accelerate the automatic construction of CXR databases from PubMed Central Open Access Subset (PMC-OA) using a combination of natural language processing and image analysis. PMC-OA is an accessible digital repository that archives full-text scholarly articles published in biomedical

* Corresponding author.
*E-mail address:* yip4002@med.cornell.edu (Y. Peng).

https://doi.org/10.1016/j.compbiomed.2023.106962
Received 26 January 2023; Received in revised form 26 March 2023; Accepted 18 April 2023
Available online 20 April 2023

and life sciences journals. To date, more than three million full-text articles have been added through the PMC [12,13]. Manuscript figures are of paramount interest because they often contain graphical images from CXR, CT, MRI, and ultrasound studies [14,15]. PMC also requires images to be supplied in an uncompressed high-resolution file format [16]. With an estimated 16 million figures (and subfigures) available, PMC promises substantial medical imaging data in various domains. More importantly, rare and new cases are strongly oversampled in the biomedical literature compared to clinical archives [17].

In this study, we hypothesize that we can obtain models with better performance on CXR pathology classification by using the weakly-labeled CXR extracted by our framework as additional training data (called PMC-CXR). Our previous work demonstrated that we could successfully construct a database for radiographs related to COVID-19 from PMC and use it as additional training data to improve deep-learning models for COVID-19 detection [18]. However, the subfigure segmentation model used in the previous study requires improvement as it occasionally struggles to differentiate similar subfigures positioned closely together. For example, in some cases, the model incorrectly identifies the spine as empty space in AP chest X-rays, leading to the erroneous splitting of large figures into two subfigures. Furthermore, the method employed in the previous work to classify the pathology mentioned in the text also requires further refinement. This work extends the previous study but differs in threefold aspects. First, we developed a new and more accurate method for subfigure segmentation and modality detection [19]. Second, we applied Radtext to classify the assertion status of the CXR pathology mentioned in the text [20]. Third, we extensively validated the utility of the automatically generated image database on thoracic disease detection tasks, including Hernia, Lung Lesion, Pneumonia, and pneumothorax. We pick these diseases due to their historically poor performance in the NIH-CXR [2] and MIMIC-CXR [1] datasets.

## 2. Materials and methods

### 2.1. Materials

We measure the Area Under the Curve (AUC) in distinct chest x-ray diagnosis models trained in three datasets: NIH-CXR dataset [2] (112,120 CXR Posterior-Anterior and Anterior-Posterior images from 30,805 individuals), MIMIC-CXR dataset [1] (243,324 CXR Posterior-Anterior and Anterior-Posterior images from 227,827 studies), and newly-created PMC-CXR. More detailed summary statistics for the datasets are listed in Table 1. For a fair comparison, we used the standard training and testing split and added the additional data from PMC-CXR only to the training set. We used a Pair T-test to compare the difference between the two groups.

**Table 1**
Characteristics of NIH-CXR, MIMIC-CXR, and PMC-CXR.

| Disease | Dataset | Training | Test |
|---|---|---|---|
| | | Positive/negative | Positive/negative |
| Hernia | | | |
| | NIH-CXR | 141/86,383 | 86/25,510 |
| | PMC-CXR | 100/- | – |
| | Total | 241/86,383 | 86/25,510 |
| Lung Lesion | | | |
| | MIMIC-CXR | 6,511/233,410 | 121/3,282 |
| | PMC-CXR | 331/- | – |
| | Total | 6842/233,410 | 121/3282 |
| Pneumothorax | | | |
| | NIH-CXR | 2,637/83,887 | 2,665/22,931 |
| | MIMIC-CXR | 11,127/228,794 | 108/3295 |
| | PMC-CXR | 929/- | – |
| | Total | 34,693/312,681 | 2,773/26,226 |
| Pneumonia | | | |
| | NIH-CXR | 876/85,648 | 555/25,041 |
| | MIMIC-CXR | 16,880/223,041 | 342/3,061 |
| | PMC-CXR | 170/- | – |
| | Total | 17,926/308,689 | 897/28,102 |

### 2.2. The pipeline to extract CXR from PMC-OA

Fig. 1 shows the overview of the proposed framework. First, we used the PubMed API (i.e., Entrez Programming Utilities) to retrieve PMC-OA articles with keywords mentioned in the titles and abstracts. Then, we extracted figures and associated captions from the input PMC-OA article and verified that the figure captions contained a positive mention of the given CXR pathology. This step ensures the image is "about" the CXR pathology of interest. Afterward, we separated the compound figures into subfigures and classified the figures into CXR and non-CXR. Finally, we created the PMC-CXR based on the following three criteria: (1) the caption contains a positive mention of the disease, (2) the figure/subfigure is a chest x-ray (CXR), and (3) the subfigure has a width-to-height or height-to-width ratio greater than 0.5. We then trained a deep neural network (DNN) with and without additional PMC-CXR.

#### 2.2.1. PMC articles retrieval

We used the PubMed API (i.e., Entrez Programming Utilities or E-utilities) to retrieve PMC-OA articles with keywords mentioned in the titles and abstracts [21]. For example, the query term for hernia-relevant articles is "Hernia [Title/Abstract]". PubMed uses Medical Subject Headings (MeSH terms) as a controlled vocabulary of biomedical and health-related terms to describe the subject of a journal article [22]. Therefore, the query automatically includes alternate spellings and MeSH terms related to "Hernia", such as "Abdominal Hernia" and "Abdominal Wall Hernias". The E-utilities retrieves PubMed Central identification (PMCID), the unique reference number assigned to every article accepted into PMC-OA. We then used PMCIDs to retrieve full-length articles in the BioC format [12]. The BioC format is a data structure in XML for text sharing and processing to facilitate the automated processing of full-text articles [23].

#### 2.2.2. Figure and text extraction

We parsed the PMC-OA articles in the BioC format to identify figures and their captions. Specifically, each figure block bears a caption, a label such as "Fig. 3", and a figure internal identifier (fig-id). We then used the fig-id to collect the figures. It is worth noting that PMC requires images to be supplied in an uncompressed high-resolution file format [16]. Fig. 2 shows an example of a typical biomedical image in the article "Pneumonia in Normal and Immunocompromised Children: An Overview and Update" [24]. The examples contain figures and a figure caption, and text that describes the case with rich information.

#### 2.2.3. CXR pathology verification

Not all images in a PMC-OA article are "about" the pathology of interest. Here, we hypothesize that the figure caption expresses the interpretation of the image. Therefore, we applied a Natural Language Processing technique to detect the pathology keywords in the figure caption and filter out CXR if its caption has no mention or negative mention of the CXR pathology keywords, such as "Hernia". Specifically, we used the previously reported extraction tool (RadText) [20] developed ourselves. It was evaluated on the MIMIC-CXR dataset with five new disease labels we annotated in our previous work [20], and achieved highly accurate classification performances, with an average precision of 0.91, a recall of 0.94, and an F-1 score of 0.92.

#### 2.2.4. Subfigure separation

The figures extracted from PMC-OA articles are usually compound and must be separated. In this study, we reused a deep learning model with 53 convolutional layers proposed by TY and YH to separate compound figures [19]. In short, the model was pretrained on the

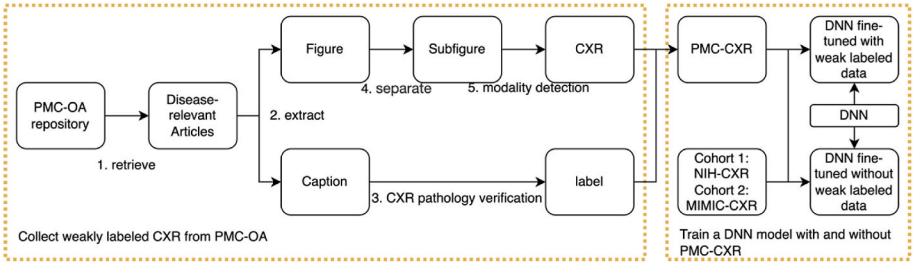


**Fig. 1.** The pipeline of collecting PMC-CXR from PMC-OA and evaluating its contribution to CXR pathology detection. ① Retrieve articles with specific diseases from PMC-OA. ② Extract figures and their associated captions from the PMC-OA articles. ③ Verify if the caption positively mentions the CXR pathology of interest. ④ Separate compound figures into subfigures. ⑤ Detect the figure modality. To evaluate the contribution of PMC-CXR, we train a deep neural network (DNN) using the cohorts or the cohorts plus PMC-CXR.

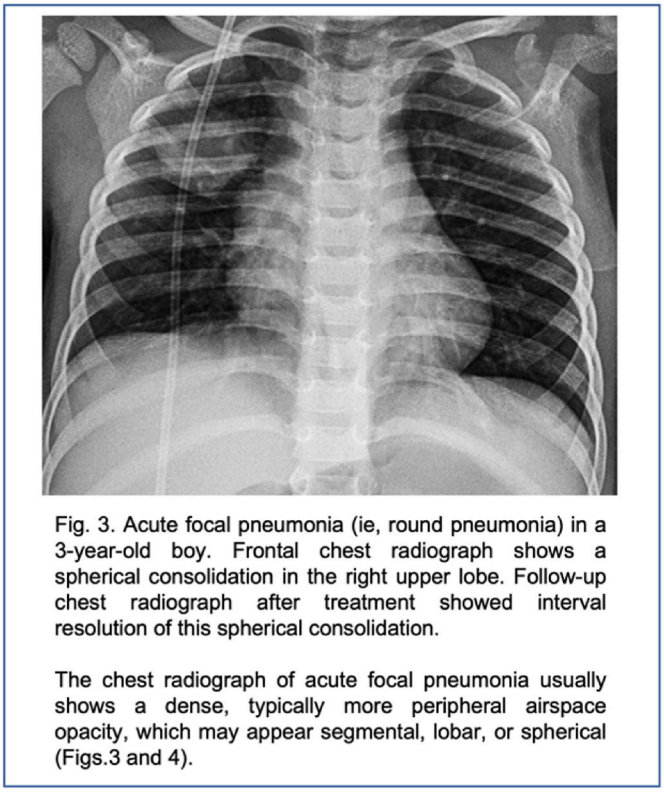

Fig. 3. Acute focal pneumonia (ie, round pneumonia) in a 3-year-old boy. Frontal chest radiograph shows a spherical consolidation in the right upper lobe. Follow-up chest radiograph after treatment showed interval resolution of this spherical consolidation.

The chest radiograph of acute focal pneumonia usually shows a dense, typically more peripheral airspace opacity, which may appear segmental, lobar, or spherical (Figs.3 and 4).

**Fig. 2.** An example of CXR with Pneumonia from the PMC-OA article "Pneumonia in Normal and Immunocompromised Children: An Overview and Update" [24].

Image-CLEF 2016 Medical dataset with an accuracy of 88.9%. We applied this model to the figures obtained by the previous steps. We discard subfigures with a width-to-height or height-to-width ratio of less than 0.5.

### 2.2.5. Image modality classification

Because many figures in PMC-OA articles are not CXR, we need to filter out non-CXR images. In this study, we applied a "double-checking" strategy, a standard practice to improve accuracy. Here, we used two models to check the model modality independently. The first model is a DenseNet-121 model [25] used in the study of Peng et al. [18]. The second is a fine-tuned ResNet-50 [26] on a newly created dataset. This dataset consists of 3901 figures (Table 2). Specifically, we randomly selected 1000 and 200 CXR images from NIH-CXR [2] and Litcovid [27], respectively. The non-CXR images were drawn from DeepLesion [3], Litcovid [27], DocFigure [28], and ImageCLEF 2016 [29].

The ResNet-50 used in this study is pretrained on ImageNet. We replaced the last classification layer with a fully connected layer with a sigmoid operation that outputs the approximate probability that an input image is a CXR or non-CXR. All images are resized to $224 \times 224 \times 3$. The models were implemented by Keras with a backend of TensorFlow. The proposed network was optimized using the Adam optimizer method [30]. The learning rate is $5 \times 10^{-5}$. A stochastic image augmentation was applied to randomly transform a given fundus photograph, resulting in an augmentation view. In this work, we sequentially apply three simple augmentations: (1) random rotation between $0^0$ and $10^0$, (2) random translation: an image was translated randomly along the x- and y-axes by distances ranging from 0 to 10% of

**Table 2**
Characteristics of the imaging modality classification dataset.

| Modality | Data source | Training | Test |
|---|---|---|---|
| CXR | | | |
| | NIH Chest X-ray | 799 | 201 |
| | LitCOVID | 160 | 40 |
| Others | | | |
| | DeepLesion | 815 | 185 |
| | LitCOVID | 160 | 41 |
| | DocFigure | 400 | 100 |
| | PMC-OA | 400 | 100 |
| | ImageCLEF 2016 | 386 | 114 |
| Total | | 3,120 | 781 |

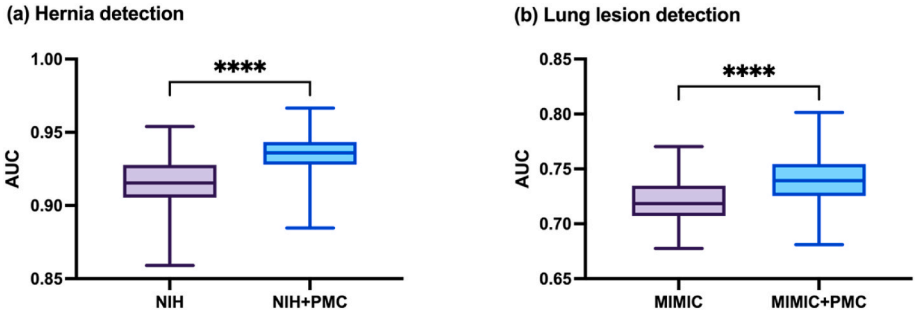


**Fig. 3.** Performance of Hernia and Lung Lesion detection on CXR. (a) Hernia detection on CXR. The models were trained on the NIH-CXR training set (left) and the NIH-CXR training set plus PMC Hernia CXR (right) and tested on the NIH-CXR test set. (b) Lung Lesion detention on CXR. The models were trained on the MIMIC-CXR training set (left) and MIMIC-CXR training set plus PMC Lung Lesion CXR (right) and tested on the MIMIC-CXR test set.

the width or height of the image, and (3) random flipping. The experiments were performed on Intel Core i9-9960 X 16 cores processor and NVIDIA Quadro RTX 6000 GPU.

### 2.3. CXR pathology detection and experimental settings

Our experiments reported the Area Under the ROC curve (AUC). We used 200 bootstrap samples to obtain a distribution of the AUC and reported 95% confidence intervals. For each bootstrap iteration, we sampled n images with replacements from the test set of n images. We used the official release training, validation, and testing datasets for both datasets. We used a DenseNet-201 model pretrained on ImageNet as a classifier for chest X-ray images. The last classification layer was replaced with a fully connected layer with a sigmoid activation function. The images were resized to $224 \times 224 \times 3$ before being fed into the model, which was implemented using Keras with a TensorFlow backend. The network was optimized using the Adam optimization algorithm [30] with a learning rate of $5 \times 10^{-5}$. We trained the model for 15 epochs. A model with the lowest validation loss is selected as the final model. The batch size is set to 96. To augment the data, we applied random rotations, translations, and flips to the images. The rotations were in the range of 0–10°, the translations were up to 10% of the image width or height in either the x or y direction, and the flips were either horizontally or vertically. The experiments were performed on an Intel Core i9-9960 X cores processor and NVIDIA Quadro RTX 6000 GPU.

## 3. Results

### 3.1. The impact of PMC-CXR on CXR pathology detection trained on the single-source dataset

In this work, we hypothesize that the additional training data extracted from biomedical articles can improve the performance of the deep learning model and reduce human effort. We first assess the impact of PMC-CXR on thoracic disease detection on CXR drawn from a single data source. Here, we chose the task of "Hernia" detection on NIH-CXR and "Lung Lesion" detection on MIMIC-CXR, because they had relatively small numbers of CXR in the datasets and are thus challenging to identify [2,31–33].

Fig. 3 shows the performance of Hernia and Lung Lesion detection on CXR. For Hernia detection, the model trained on NIH-CXR plus PMC Hernia CXR is significantly superior to the model trained on NIH-CXR only (0.9335 vs. 0.9154 in AUC, $p < 0.0001$). For Lung Lesion detection, the model trained on MIMIC-CXR and PMC Lung Lesion CXR also achieved higher AUC than its counterpart (0.7394 vs. 0.7207 in AUC, $p < 0.0001$).

### 3.2. The impact of PMC-CXR on CXR pathology detection trained on multi-source datasets

We then assess the impact of PMC-CXR on thoracic disease detection on CXR drawn from multiple data sources. For this purpose, we chose the tasks of Pneumonia and Pneumothorax detection because these two diseases are commonly annotated in NIH-CXR and MIMIC-CXR.

Fig. 4 shows the results of pneumonia and pneumothorax detection tested on the NIH-CXR test set. In both cases, the model trained on the NIH-CXR plus PMC-CXR is superior to the model trained on the NIH-CXR only (pneumonia: 0.6506 vs. 0.6348 in AUC, $p < 0.0001$; pneumothorax: 0.8423 vs. 0.8279 in AUC, $p < 0.0001$). For a fair comparison, we also examined whether the model could reach a higher performance using additional "real" positive CXR. For this purpose, we selected "Pneumonia" images from MIMIC-CXR and added them to the NIH-CXR training set. We observed that the model trained on a combination of the NIH-CXR and MIMIC pneumonia/pneumothorax CXR also had better performance (pneumonia: 0.6484 vs. 0.6348 in AUC, $p < 0.0001$; pneumothorax: 0.8431 vs. 0.8279 in AUC, $p < 0.0001$). More importantly, the model trained on NIH CXR plus PMC Pneumonia CXR is superior to that trained on NIH CXR plus MIMIC Pneumonia CXR on Pneumonia detection, demonstrating the usefulness of our extracted PMC-CXR.

Fig. 5 shows the pneumonia and pneumothorax detection results tested on the MIMIC-CXR test set. The model trained with additional PMC-CXR had better performance than that trained on MIMIC-CXR only (pneumonia: 0.6813 vs. 0.6709 in AUC, $p < 0.0001$; pneumothorax: 0.7670 vs. 0.7517 in AUC, $p < 0.0001$). When we replaced the positive images from PMC-CXR with the positive images from NIH-CXR, the model also had superior performance to those trained on MIMIC-CXR only (pneumonia: 0.7074 vs. 0.6709 in AUC, $p < 0.0001$, pneumothorax: 0.8185 vs. 0.7517 in AUC, $p < 0.0001$) and those trained on MIMIC-CXR plus PMC-CXR.

Finally, we examined whether additional data PMC-CXR is necessary when we already have multi-source data by training two models: one using a combination of NIH-CXR and MIMIC-CXR, and the other using a combination of these datasets plus PMC-CXR. Fig. 6(a–b) shows that, when tested on the NIH-CXR for pneumonia detection, the model trained with additional PMC-CXR had better performance than that trained on the combination of NIH-CXR and MIMIC-CXR training sets (NIH-CXR test set: 0.6788 vs. 0.6640 in AUC, $p < 0.0001$). But when tested on the MIMIC-CXR test set, there is no significant difference. For pneumothorax detection, Fig. 6(c–d) shows that the model trained with additional PMC-CXR had better performance than that trained on the combination of NIH-CXR and MIMIC-CXR training sets (NIH-CXR test set: 0.8720 vs. 0.8669 in AUC, $p < 0.0001$; MIMIC-CXR test set: 0.8202 vs. 0.7972 in AUC, $p < 0.0001$).

### 3.3. Image modality classification

A large portion of figures in the PMC-OA articles is not CXR. Therefore, an image modality classifier is needed to distinguish CXR from non-CXR figures. The full description is available in the Methods. The model achieved high accuracy (0.9987), sensitivity (1.0), and specificity (0.9981).

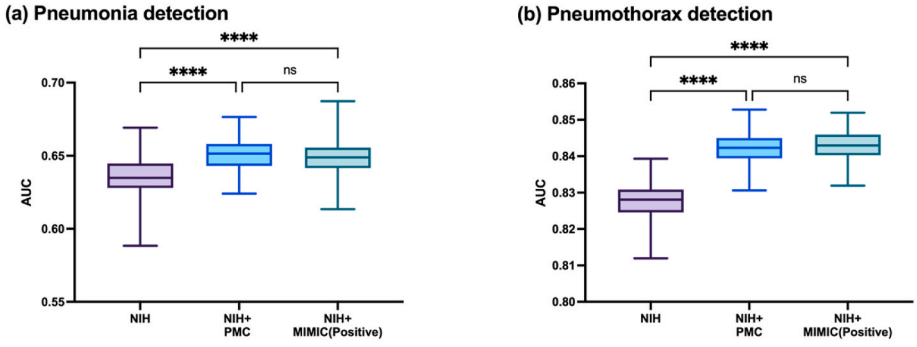


**Fig. 4.** Performance of Pneumonia and Pneumothorax detection on CXR tested on NIH-CXR test set. (a) Pneumonia detection on CXR. The models were trained on the NIH-CXR training set (left), NIH-CXR training set plus MIMIC Pneumonia CXR (middle), and NIH-CXR training plus PMC Pneumonia CXR (right). (b) Pneumothorax detection on CXR. The models were trained on the NIH-CXR training set (left), NIH-CXR training set plus MIMIC Pneumothorax CXR (middle), and NIH-CXR training plus PMC Pneumothorax CXR (right).

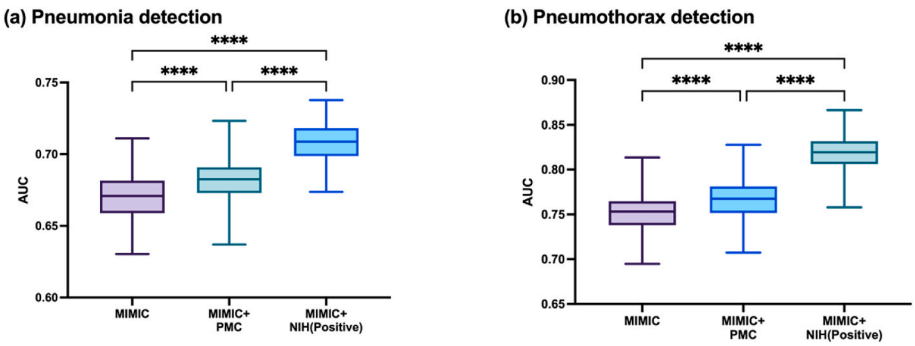


**Fig. 5.** Performance of Pneumonia and Pneumothorax detection on CXR tested on MIMIC-CXR test set. (a) Pneumonia detection on CXR. The models were trained on the MIMIC-CXR training set (left), MIMIC-CXR training set plus NIH Pneumonia CXR (middle), and MIMIC-CXR training plus PMC Pneumonia CXR (right). (b) Pneumothorax detection on CXR. The models were trained on the MIMIC-CXR training set (left), MIMIC-CXR training set plus NIH Pneumothorax CXR (middle), and MIMIC-CXR training plus PMC Pneumothorax CXR (right).

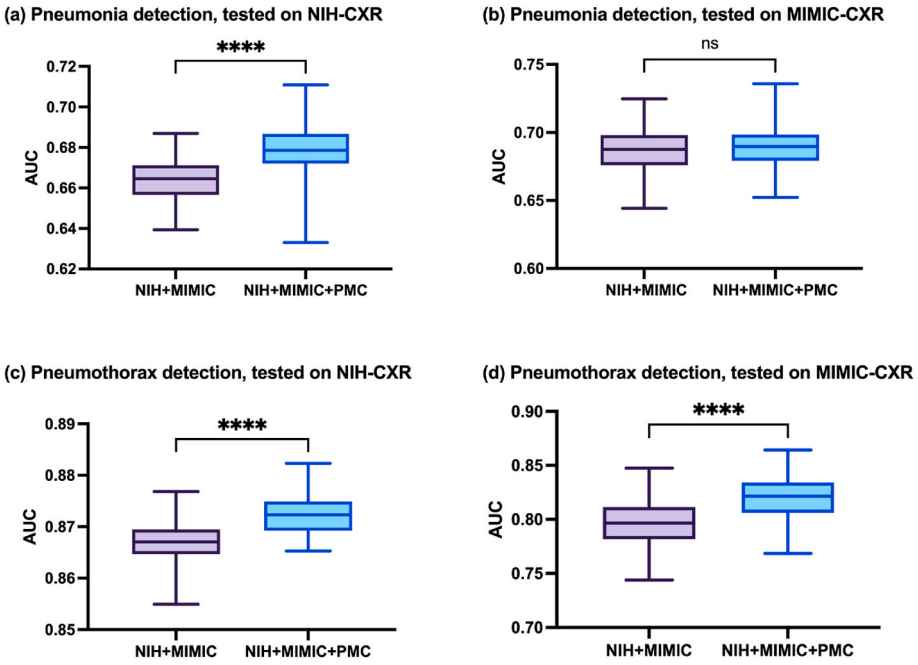


**Fig. 6.** Performance of pneumonia (a–b) and pneumothorax (c–d) detection on CXR. The models were trained on the combination of NIH-CXR and MIMIC-CXR training sets (left) plus PMC Pneumonia/Pneumothorax CXR (right) and tested on the NIH-CXR and MIMIC-CXR test sets, respectively.

## 4. Discussion

There are many studies on mining figures within scientific documents [18,34–36]. However, constructing a large-scale medical imaging database from biomedical literature needs to be better studied, not to mention the associated information [18]. To bridge this gap, we designed a practical framework to extract medical images from PMC. In contrast to previous approaches that relied solely on the manual submission of medical images to the repository, figures and their accompanied figure legends are automatically collected using algorithms that integrate natural language processing and medical image analysis. We have shown that the model trained with PMC-CXR as the additional training data consistently achieved superior performance (Figs. 2–4). More importantly, we observed that, even though the additional data extracted using our proposed framework is less than the real data, the improvement in classification is similar to that obtained by adding the real data. This demonstrates the effectiveness of our flexible proposed framework, which allows for the easy extraction of data from relevant articles. As more relevant articles are published online, we can continue to expand our dataset using this framework.

Beyond these immediate takeaways, there are several topics for further discussion. First, we have seen the success of "ImageNet" in the general domain [37]. ImageNet uses the hierarchical structure of WordNet [38], an extensive lexical database of English where nouns, verbs, adjectives, and adverbs are grouped into sets of cognitive synonyms. Since the launch of ImageNet and the inaugural competitions using the database, many successful efforts have been using deep neural networks in a wide variety of computer vision tasks, providing an appreciation of the importance of hierarchical structures in improving model performance [39]. In the medical domain, there exist large public datasets of CXR [1,2,40]. Nevertheless, few collections consist of diverse CXR pathologies to effectively train a deep neural network for the general purpose of CXR understanding [41–43]. Our approach provides the potential to construct a large-scale CXR database quickly and accurately with comprehensive disease coverage and promote the performance of current CXR analysis algorithms.

Second, we highlight that, owing to the advantages of the Creative Commons license, our dataset will allow researchers to gain findable, assessable, interoperable, and reusable (FAIR) access to huge collections of CXR images along with the descriptive text. PMC provides rich resources with a free and reusable license for secondary analysis [44,45]. To date, more than three million full-text articles have been added through the PMC [12,13]. Manuscript figures are of paramount interest because they often contain graphical images from CXR, CT, MRI, and ultrasound studies [14,15]. With an estimated 16 million figures (and subfigures) available, PMC promises substantial medical imaging data in

various domains. More importantly, rare and new cases are strongly oversampled in the biomedical literature compared to clinical archives [17]. As a result, medical imaging databases built from PMC will provide opportunities for the fast development of image-based models for detecting new diseases such as COVID-19.

Furthermore, NIH-CXR, MIMIC-CXR, and PMC-CXR used the same automatic labeling algorithm. In radiology, we have seen the increasing use of NLP methods to automatically generate labels from text. This results in large annotated CXR datasets such as NIH-CXR, MIMIC-CXR, and CheXpert. While these datasets are widely used for developing DL models, obtaining CXR with rare or new diseases is challenging. To this end, our solution provides an alternative way to quickly harvest cases with CXR pathologies. More importantly, the performance of the labeler has been validated for quality [1,2,40,46] and adopted a reliable ground truth. Experiments show that PMC-CXR significantly improves the CXR pathologies classification across two large chest X-ray datasets.

One limitation of our proposed framework is that the labels for the captions may contain noises. For example, the caption of a compound figure may include several items. Our current method can only provide a unitary label for the whole caption. Fig. 7 shows an example of a typical biomedical image in the article "Congenital Hemidiaphragmatic Agenesis Presenting as Reversible Mesenteroaxial Gastric Volvulus and Diaphragmatic Hernia: A Case Report" [47]. While we can separate the figure into three subfigures, we incorrectly label subfigure A as a CXR with Hernia. This error may degenerate the model training and the testing performance. However, the experimental results demonstrate that such noises do not significantly degenerate the effectiveness, i.e., the additional data generated by our framework still improves the performance. In the future, we will segment the captions into fine-grained ones and align them to each subfigure. This will ensure we can generate more precise labels for subfigures.

The other limitation of using medical literature is the high level of heterogeneity presented in different sources. For example, figures may be collected from various perspectives and locations, resulting in significant differences in their feature domains. In the future, we plan to address this issue by using techniques such as domain adaptation with contrast loss to align domains [48], and improve the effectiveness of the figures in classification tasks.

In conclusion, we have developed an end-to-end framework, a first-of-its-kind, to automatically extract CXR images from PMC-OA. Our proposed framework offers improved subfigure segmentation and incorporates our advanced self-developed NLP technique for CXR pathology verification. The method will avoid the considerable questions of the ownership, control, security, and privacy of biomedical data. We have performed several experiments to prove that creating additional training data from biomedical articles can improve the performance of the deep learning model. While this project focuses on CXR images, we expect that our approach could apply to other specialties like dermatology, ophthalmology, and pathology. We hope this framework can generate additional data to facilitate deep learning model development and evaluation, educate medical students and residents, and help to evaluate findings reported by radiologists. It may also have positive feedback and encourage researchers to dedicate their resources to the open research questions identified and contribute their image data and resources to establishing high-quality benchmarking data sets.

## Data availability

The first dataset is provided by the NIH Clinical Center and is available through the NIH download site: https://nihcc.app.box.com/v/ChestXray-NIHCC. The second dataset MIMIC-CXR is also publicly available on PhysioNet[49,50] https://www.physionet.org/content/mimic-cxr-jpg/.

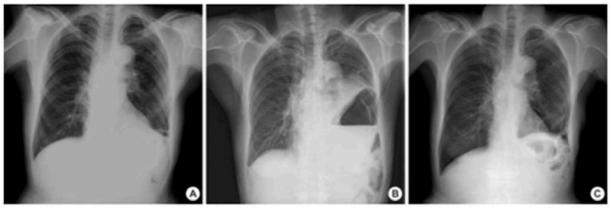

Fig. 1. Chest radiographies. (A) There was no abnormal finding reported on from the local clinic 4 days before admission. (B) On admission, gastric air-fluid and bowel loops were observed in the left thoracic cavity and a coiled nasogastric tube was seen in the stomach. (C) The herniation improved but haziness was still seen at the left lower lobe after surgical intervention.

**Fig. 7.** Examples of CXR that are positive for Hernia. The figures are from the article "Congenital Hemidiaphragmatic Agenesis Presenting as Reversible Mesenteroaxial Gastric Volvulus and Diaphragmatic Hernia: A Case Report" [47].

## Code availability

Codes are available at https://github.com/bionlplab/PMC-CXR.

## Conflict of interest

No conflicting relationship exists for any author.

## Acknowledgments

This work was supported by the National Institutes of Health under Award No. 4R00LM013001 (Peng) and R01DK135597 (Huo), NSF CAREER Award No. 2145640 (Peng), Schmidt Futures' AI2050 Early Career Fellowship (Yang), Cornell Multi-Investigator Seed Grant (Peng, Yang, and Shih), and Amazon Research Award (Peng).

## References

[1] A.E.W. Johnson, T.J. Pollard, N.R. Greenbaum, M.P. Lungren, C.-Y. Deng, Y. Peng, Z. Lu, R.G. Mark, S.J. Berkowitz, S. Horng, MIMIC-CXR-JPG, a Large Publicly Available Database of Labeled Chest Radiographs, 2019 arXiv preprint arXiv: 1901.07042.
[2] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, R.M. Summers, Chest x-ray8: hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2017, pp. 3462–3471.
[3] K. Yan, X. Wang, L. Lu, R.M. Summers, DeepLesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning, J.Medi. Imag. 5 (2018), 036501.
[4] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, Chexpert: a large chest radiograph dataset with uncertainty labels and expert comparison, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2019, pp. 590–597.
[5] N.A. Phillips, P. Rajpurkar, M. Sabini, R. Krishnan, S. Zhou, A. Pareek, N.M. Phu, C. Wang, M. Jain, N.D. Du, CheXphoto: 10,000+ Photos and Transformations of Chest X-Rays for Benchmarking Deep Learning Robustness, Machine Learning for Health, PMLR, 2020, pp. 318–327.
[6] A.S. Adamson, A. Smith, Machine learning and health care disparities in dermatology, JAMA Dermatol. 154 (2018) 1247–1248.
[7] L. Oakden-Rayner, Exploring large-scale public medical image datasets, Acad. Radiol. 27 (2020) 106–112.
[8] Z. Obermeyer, B. Powers, C. Vogeli, S. Mullainathan, Dissecting racial bias in an algorithm used to manage the health of populations, Science 366 (2019) 447–453.
[9] J. Schaefer, M. Lehne, J. Schepers, F. Prasser, S. Thun, The use of machine learning in rare diseases: a scoping review, Orphanet J. Rare Dis. 15 (2020) 145.
[10] Medical Imaging and Data Resource Center (MIDRC), https://data.midrc.org/.
[11] M.J. Willemink, W.A. Koszek, C. Hardell, J. Wu, D. Fleischmann, H. Harvey, L. R. Folio, R.M. Summers, D.L. Rubin, M.P. Lungren, Preparing medical imaging data for machine learning, Radiology 295 (2020) 4–15.

[12] D.C. Comeau, C.-H. Wei, R. Islamaj Doğan, Z. Lu, PMC text mining subset in BioC: about three million full-text articles and growing, Bioinformatics 35 (2019) 3533–3535.
[13] S.R. Choudhury, S. Tuarob, P. Mitra, L. Rokach, A. Kirk, S. Szep, D. Pellegrino, S. Jones, C.L. Giles, A figure search engine architecture for a chemistry digital library, in: Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries - JCDL \textquotesingle13, ACM Press, 2013, pp. 369–370.
[14] L.D. Lopez, J. Yu, C. Arighi, C.O. Tudor, M. Torii, H. Huang, K. Vijay-Shanker, C. Wu, A framework for biomedical figure segmentation towards image-based document retrieval, BMC Syst. Biol. 7 (Suppl 4) (2013) S8.
[15] S. Tsutsui, D.J. Crandall, A Data Driven Approach for Compound Figure Separation Using Convolutional Neural Networks, 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), IEEE, 2017, pp. 533–540.
[16] National Library of Medicine, Image Quality Specifications, https://www.ncbi.nlm.nih.gov/pmc/pub/filespec-images/.
[17] A.K. Dhrangadhariya, O. Jimenez-del-Toro, V. Andrearczyk, M. Atzori, H. Müller, Exploiting biomedical literature to mine out a large multimodal dataset of rare cancer studies, in: T.M. Deserno, P.-H. Chen (Eds.), Medical Imaging 2020: Imaging Informatics for Healthcare, Research, and Applications, SPIE, 2020, p. 9.
[18] Y. Peng, Y. Tang, S. Lee, Y. Zhu, R. Summers, Z. Lu, COVID-19-CT-CXR: a freely accessible and weakly labeled chest X-ray and CT image collection on COVID-19 from biomedical literature, IEEE Trans. Big Data 7 (2021) 3–9.
[19] T. Yao, C. Qu, Q. Liu, R. Deng, Y. Tian, J. Xu, A. Jha, S. Bao, M. Zhao, A.B. Fogo, Compound Figure Separation of Biomedical Images with Side Loss, Deep Generative Models, and Data Augmentation, Labelling, and Imperfections, Springer, 2021, pp. 173–183.
[20] S. Wang, M. Lin, Y. Ding, G. Shih, Z. Lu, Y. Peng, Radiology text analysis system (RadText): architecture and evaluation, IEEE Int Conf Healthc Inform 2022 (2022) 288–296.
[21] E. Sayers, A General Introduction to the E-Utilities, National Center for Biotechnology Information, 2022.
[22] H.J. Lowe, G.O. Barnett, Understanding and using the medical subject headings (MeSH) vocabulary to perform literature searches, JAMA 271 (1994) 1103–1108.
[23] D.C. Comeau, R. Islamaj Doğan, P. Ciccarese, K.B. Cohen, M. Krallinger, F. Leitner, Z. Lu, Y. Peng, F. Rinaldi, M. Torii, A. Valencia, K. Verspoor, T.C. Wiegers, C. H. Wu, W.J. Wilbur, BioC: a minimalist approach to interoperability for biomedical text processing, Database 2013 (2013) bat064.
[24] H.K. Eslamy, B. Newman, Pneumonia in normal and immunocompromised children: an overview and update, Radiol. Clin. 49 (2011) 895–920.
[25] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely Connected Convolutional Networks, 2017, pp. 4700–4708.
[26] K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, 2016, pp. 770–778.
[27] Q. Chen, A. Allot, Z. Lu, Keep up with the latest coronavirus research, Nature 579 (2020) 193.
[28] K.V. Jobin, A. Mondal, C.V. Jawahar, DocFigure: A Dataset for Scientific Document Figure Classification, IEEE, 2019, pp. 74–79.
[29] A. Garcia Seco De Herrera, R. Schaer, S. Bromuri, H. Müller, Overview of the ImageCLEF 2016 Medical Task, CEUR Workshop Proceedings, 2016.
[30] D.P. Kingma, J. Ba, Adam: A Method for Stochastic Optimization, 2014 arXiv preprint arXiv:1412.6980.
[31] Y. Han, C. Chen, L. Tang, M. Lin, A. Jaiswal, S. Wang, A. Tewfik, G. Shih, Y. Ding, Y. Peng, Using radiomics as prior knowledge for thorax disease classification and localization in chest X-rays, AMIA Annu. Symp. Proc. 2021 (2021) 546–555.
[32] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, Others, Chexnet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning, 2017 arXiv preprint arXiv:1711.05225.
[33] L. Seyyed-Kalantari, G. Liu, M. McDermott, I.Y. Chen, M. Ghassemi, CheXclusion: fairness gaps in deep chest X-ray classifiers, Pac. Symp. Biocomput. 26 (2021) 232–243.
[34] Z. Ahmed, S. Zeeshan, T. Dandekar, Mining biomedical images towards valuable information retrieval in biomedical and life sciences, Database (2016) 2016.
[35] N. Siegel, N. Lourie, R. Power, W. Ammar, Extracting scientific figures with distantly supervised neural networks, in: Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries, ACM, 2018, pp. 223–232.
[36] P. Li, X. Jiang, H. Shatkay, Figure and caption extraction from biomedical documents, Bioinformatics 35 (2019) 4381–4388.
[37] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, ImageNet: a large-scale hierarchical image database, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2009, pp. 248–255.
[38] G.A. Miller, WordNet: a lexical database for English, Commun, ACM 38 (1995) 39–41.
[39] K.K. Bressem, L.C. Adams, C. Erxleben, B. Hamm, S.M. Niehues, J.L. Vahldiek, Comparing different deep learning architectures for classification of chest radiographs, Sci. Rep. 10 (2020), 13590.
[40] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, CheXpert: a large chest radiograph dataset with uncertainty labels and expert comparison, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2019, pp. 590–597.
[41] J.P. Cohen, P. Morrison, L. Dao, COVID-19 Image Data Collection, 2020 arXiv: 2003.11597 [cs, eess, q-bio].
[42] X. He, X. Yang, S. Zhang, J. Zhao, Y. Zhang, E. Xing, P. Xie, Sample-efficient Deep Learning for COVID-19 Diagnosis Based on CT, 2020, p. 04, medrxiv,.
[43] K. Zhang, X. Liu, J. Shen, Z. Li, Y. Sang, X. Wu, Y. Zha, W. Liang, C. Wang, K. Wang, L. Ye, M. Gao, Z. Zhou, L. Li, J. Wang, Z. Yang, H. Cai, J. Xu, L. Yang, W. Cai, W. Xu, S. Wu, W. Zhang, S. Jiang, L. Zheng, X. Zhang, L. Wang, L. Lu, J. Li, H. Yin, W. Wang, O. Li, C. Zhang, L. Liang, T. Wu, R. Deng, K. Wei, Y. Zhou, T. Chen, J.Y.-N. Lau, M. Fok, J. He, T. Lin, W. Li, G. Wang, Clinically applicable AI system for accurate diagnosis, quantitative measurements, and prognosis of COVID-19 pneumonia using computed tomography, Cell (2020).
[44] L.L. Wang, K. Lo, Y. Chandrasekhar, R. Reas, J. Yang, D. Eide, K. Funk, R. Kinney, Z. Liu, W. Merrill, P. Mooney, D. Murdick, D. Rishi, J. Sheehan, Z. Shen, B. Stilson, A.D. Wade, K. Wang, C. Wilhelm, B. Xie, D. Raymond, D.S. Weld, O. Etzioni, S. Kohlmeier, CORD-19: the Covid-19 Open Research Dataset, 2020.
[45] Q. Chen, A. Allot, Z. Lu, LitCovid: an Open Database of COVID-19 Literature, Nucleic Acids Res., 2021, 49.D1 D1534-D1540.
[46] Y. Peng, X. Wang, L. Lu, M. Bagheri, R. Summers, Z. Lu, NegBio: a high-performance tool for negation and uncertainty detection in radiology reports, AMIA Jt Summits Transl Sci Proc (2017) 188–196, 2018.
[47] H.Y. Sung, S.H. Cho, S.B. Sim, J.I. Kim, D.Y. Cheung, S.-H. Park, J.-Y. Han, S. M. Lee, C.H. Noh, Y.-B. Park, Congenital hemidiaphragmatic agenesis presenting as reversible mesenteroaxial gastric volvulus and diaphragmatic hernia: a case report, J. Kor. Med. Sci. 24 (2009) 517–519.
[48] S. Motiian, M. Piccirilli, D.A. Adjeroh, others, Unified deep supervised domain adaptation and generalization, in: In Proceedings of the IEEE international conference on computer vision, 2017, pp. 5715–5725.