Automated Labeling of Chest X-ray Images using a Quantitative Explainable Atlas-Based AI Model

Doyun Kim
Massachusetts General Hospital

Joowon Chung
https://orcid.org/0000-0003-1572-8056

Jongmung Choi
Massachusetts General Hospital

Marc Succi
Massachusetts General Hospital

John Conklin
Massachusetts General Hospital

Maria Gabriela Figueiro Longo
Massachusetts General Hospital

Jeanne Ackman
Massachusetts General Hospital https://orcid.org/0000-0001-5022-9530

Brent Little
Massachusetts General Hospital

Milena Petranovic
Massachusetts General Hospital

Mannudeep Kalra
Massachusetts General Hospital https://orcid.org/0000-0001-9938-7476

Michael Lev
Massachusetts General Hospital https://orcid.org/0000-0003-0236-7319

Synho Do (✉ sdo@mgh.harvard.edu)
Massachusetts General Hospital https://orcid.org/0000-0001-6211-7050

Article

Keywords: labeling, chest x-ray images, open-access medical imaging datasets, AI models

DOI: https://doi.org/10.21203/rs.3.rs-830690/v1

License: © This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License
Automated Labeling of Chest X-ray Images using a Quantitative Explainable Atlas-Based AI Model

Abstract

The inability to accurately, efficiently label large, open-access medical imaging datasets limits the widespread implementation of artificial intelligence models in healthcare. There have been few attempts, however, to automate the annotation of such public databases; one approach, for example, focused on labor-intensive, manual labeling of subsets of these datasets to be used to train new models. In this study, we describe a method for standardized, automated labeling based on similarity to a previously validated, explainable AI model (xAI), using an atlas-based approach, for which the user can specify a quantitative threshold for a desired level of accuracy, the “probability-of-similarity” (pSim) metric. We showed that our xAI model, by calculating the pSim values for each feature based on comparison to its training-set derived reference atlas, could automatically label the external datasets to a user-selected, high level of accuracy, equaling or exceeding that of human experts.
Introduction

The implementation of medical artificial intelligence (AI) into clinical practice in general, and radiology practice in particular, has in large part been limited by the time, cost, and expertise required to accurately label very large imaging datasets, which can serve as “platinum level” ground truth for training clinically relevant AI models. The ability to automatically and efficiently annotate large external datasets, to a user-selected level-of-accuracy, may therefore be of considerable value in developing impactful, important, medical AI models that bring added value to, and are widely accepted by, the healthcare community. Such an approach not only has the potential to benefit re-training to improve the accuracy of existing AI models, but also – through using explainable, atlas-based methodology [1] - may help to standardize labeling of open-source datasets [2-5], for which the provided labels can be noisy, inaccurate, or absent. Such standardization may, in turn, reduce the number of datapoints required for accurate model building, facilitating training and re-training from initial small but well annotated datasets [1, 6].

In this study, we develop and demonstrate a method for standardized, automated labeling based on similarity to a previously validated explainable AI model (xAI), using an atlas-based approach for which the user can specify a quantitative threshold for a desired level of accuracy (the “probability-of-similarity”, or “pSim” metric). The pSim values range from a “baseline” likelihood of similarity (pSim=0, least selective) to a “maximal” likelihood of similarity (pSim=1, most selective); pSim is computed by comparison between test-set derived image features and image features “remembered”/retrieved from the model’s reference atlas (i.e., “library”). This atlas is constructed during model building (Fig. 1a) from the training set cases (Fig. 1b). The calculated pSim value reflects the harmonic mean between two model related parameters, the “patch similarity” and the “confidence” (Methods, Figs. 1b,c).
Specifically, we applied our existing AI model for detection of five different chest X-ray (CXR) imaging features (cardiomegaly, atelectasis, pulmonary edema, pneumonia, and pleural effusion), to three large open-source datasets – CheXpert [2], MIMIC [3], and NIH [4] - and compared the resulting labels to those of 7 human expert radiologists. Of note, there is an inverse relationship between the selected pSim threshold values and the number of cases identified (i.e., “captured”) by the model from the external dataset; in other words, the higher the threshold for likelihood of similarity, the fewer cases that will be identified from the external database as “similar” to the model labeled cases.

We showed that our xAI model, by calculating the pSim values for each feature based on comparison to the model’s training-set derived reference atlas, could automatically label the external datasets to a user-selected, arbitrarily high level of accuracy, equaling or exceeding that of human experts. Although the pSim threshold value required to achieve “maximal” similarity varies by feature, once that value is identified - based on comparison of model labels to expert-labeled ground truth - it can then be applied to the remaining external dataset, to identify cases likely to be positive for that feature at a pre-determined, high confidence level.
Results

Fig. 1 | System overview. Standardized, automated labeling method, based on similarity to a previously validated five-feature CXR detection xAI model, using an explainable atlas-based approach. 

a, b The xAI model calculates “patch similarity” and “confidence” probabilities, based on class activation mapping (CAM) [7, 8] and predicted probability from the model, for each feature. c, The harmonic mean between the patch similarity and confidence xAI model outputs are then used to calculate a “probability of similarity” (pSim) for each feature.
System design. We developed an xAI model for detection of the following five different features on posterior-anterior (PA) projection CXRs: cardiomegaly, atelectasis, pulmonary edema, pneumonia, and pleural effusion (see Methods). As per previous reports, our model featured atlas creation and prediction-basis calculation modules, for explainability (Fig. 1) [1]. The prediction-basis was used to calculate a "patch similarity" value (a probability between 0 and 1). Our model also included a "confidence" probability calculation module (Figs. 1a,b). The harmonic mean between the patch similarity and confidence model outputs were used to calculate a quantitative "probability of similarity" (pSim) value, between 0 and 1, for each feature studied (Fig. 1c).

xAI model development. CXR examinations performed at our institution from February 2015 through February 2019 were identified from our RIS (Radiology Information System) and PACS (Picture Archiving and Communication System), resulting in a dataset of 440,852 studies. Examinations were excluded if there was no associated radiology report, no view position information (e.g. anteroposterior projection, "portable", etc.), or no essential patient identifiers (including but not limited to medical record number, age, or gender). A total of 400,886 CXR images from 267,180 examinations, representing 117,195 patients, together with their corresponding radiology reports, were collected retrospectively (Supplementary Figure 1). Using a rule-based Natural Language Processing (NLP) model (Supplementary Table 1), we automatically extracted 20 pathological features from the radiology reports, which were assigned one of the following three labels: "positive", "negative", or "ignore". After automated NLP data mining and clean-up, we archived 151,700 anteroposterior (AP) CXR views from 49,096 patients (58% male, mean age 62±18 years) and 90,023 posteroanterior (PA) CXR views from 69,404 patients (50% male, mean age 57±19 years). We randomly selected 1000 images for each view position as a test set; the remaining examinations, from non-overlapping patients, were separated into training and validation sets (Supplementary Fig. 1). The labels for the training and validation
sets were determined exclusively from the automated NLP assignments, whereas those for the
test set were determined by consensus of three U.S. board-certified radiologists at our institution
(further details provided in Supplementary Table 1), using the “Mark-it” tool
(https://markit.mgh.harvard.edu, MA, USA) for annotation [9]. This xAI model achieved a mean
Area Under the Receiver Operating Characteristic (AUROC) curve [10] of 0.95 ± 0.02 for detection
of the five features (Supplementary Table 2) in our initial, independent test set (Methods).

**Auto-labeling model performance applied to three open-source datasets.** We applied our
xAI CXR auto-labeling model to the available PA CXR images from three large open-source
datasets; CheXpert (n=29,420 PA CXR’s), MIMIC (n=71,223), and NIH (n=67,310) [2-4]. To
assess labeling accuracy, we randomly selected a subset of “positive” and “negative” cases as
determined by the model for each of the five features, distributed equally in each of ten pSim
value ranges (0-0.1, 0.1-0.2, 0.2-0.3, …, 0.9-1.0), for expert review (Fig. 2). Ground truth (GT)
was defined as the majority consensus of 7 expert sub-specialist radiologists (three with 12-25
years’ experience in thoracic radiology and four with 1-6 years’ experience in emergency
radiology); GT and individual ratings of each reader, for each feature (cardiomegaly, atelectasis,
pulmonary edema, pneumonia, and pleural effusion), in each of the pSim value ranges, are shown
in panel 1 of Fig. 2 (upper left). In Fig. 2 panel 2 (upper right), we graph the relationship between
the pSim value applied for the model’s auto-labeling (x-axis) and both the (i) “positive predictive
value” and “negative predictive value” of the model’s ratings, versus ground truth; and the (ii)
model’s “true positive capture rate” and “true negative capture rate”, defined respectively as the
total *true positive* (by GT) divided by the total *positive* (by GT), and the total *true negative* (by GT)
divided by the total *negative* (by GT). In panels 3 (lower left) and 4 (lower right), respectively, the
number of false positive (by GT) and false negative (by GT) cases rated by the model at each
pSim threshold value (x-axis), are shown, stratified by dataset (CheXpert, MIMIC, or NIH), with
the “optimal”, lowest pSim threshold achieving 100% PPV or NPV, indicated. Of note, the lowest
possible pSim threshold required for 100% PPV or NPV, corresponds to the maximal “correct
capture rate”, as shown in panel 2.

Also, as shown in the text boxes in Fig. 2 panels 3 and 4, as well as in Fig. 3, model accuracy
compared favorably to that of the available pooled public labels of the external, open-source
datasets. Figure 3 additionally shows that the automated-labeling model’s AUROC performance,
compared favorably to that of the individual expert radiologists, for each feature, at both the
pSim=0 “baseline” value labeling threshold and the “optimal” pSim value labeling threshold (i.e.,
the lowest pSim value achieving 100% accuracy, as per Fig. 2 panels 3 and 4).

Sample auto-labeled CXR images that had complete agreement between all 7 expert radiologists
and the xAI model, positive for each of the five features studied, are shown in Fig. 4. The pSim
threshold values applied by the model for each image and the number/percent of PA CXR
examinations with total agreement for each feature, are also shown. Of note, there were only 14
positive examinations identified by the model as “pneumonia” that had full agreement with each
reader, of 50 total examinations labeled as positive for pneumonia (28%). The percent positive
labels with complete agreement for the other four features, as shown in the figure, were
cardiomegaly 78% (39/50), atelectasis 46% (23/50), pulmonary edema 43% (17/40), and pleural
effusion 78% (39/50).

In Table 1, we applied our automated-labeling model to the three complete public, open source
CXR datasets: CheXpert (n=29,420), MIMIC (n=71,223), and NIH (n=67,310); in order to
demonstrate the magnitude of the number of cases “captured”, at the optimized pSim threshold
value for maximal accuracy for each feature (PPV, NPV = 1; as per Fig. 2). Pooling the model’s
labels for the three full public datasets (Table 1, far right) resulted in a “capture rate” of 80% for
cardiomegaly (134,076/167,953), 28% for atelectasis (47,436/167,953), 27% for pulmonary
edema (45,660/167,953), 20% for pneumonia (33,308/167,953), and 68% for pleural effusion (114,230/167,953). It is noteworthy that the model's mean CXR “capture rates” for the pooled results from the three public datasets, closely corresponded to those shown in the graphs of Fig. 2.2 (a-e), for the subset of examinations (n=90-100) labeled by both the model and the expert radiologists.

**Summary comparison of labeling efficiency, confidence metrics for the 5 auto-labeled features.** For each of the five auto-labeled features (Fig. 5), we compared: (i) the percent of positively auto-labeled CXR’s “captured” from the three pooled, full public datasets (from Table 1); (ii) the percent of cases with complete agreement between the model and all 7 expert readers (from Fig. 4); (iii) the lowest pSim value such that PPV=1 (graphed as “1-pSim@PPV1”; from Fig. 2 panel 3), and (iv) the lowest pSim value such that NPV=1 (graphed as “1-pSim@NPV1”; from Fig. 2 panel 4). Features with higher values of these parameters (e.g. cardiomegaly, pleural effusion) corresponded to greater model auto-labeling efficiency and confidence; features with lower values (e.g. pneumonia, pulmonary edema) corresponded to lesser model auto-labeling efficiency and confidence. Of note, for atelectasis, “1-pSim@PPV1” was higher than “1-pSim@NPV1”, indicating greater confidence that the model is correct in “ruling-in” this feature (i.e. correctly auto-labeling true-positives) than in “ruling-out” this feature (i.e. correctly auto-labeling true-negatives). This relationship was reversed for the other four features (e.g. greater confidence that the model can correctly “rule-out” than “rule-in” pneumonia or pulmonary edema).

The pairwise kappa statistics estimating inter-observer variability among the 7 expert radiologists are shown in Fig. 6, for each of the five auto-labeled features. The ranges for these values are as follows: cardiomegaly 0.82-0.92, pleural effusion 0.78-0.94, atelectasis 0.47-0.78, pulmonary edema 0.57-0.86, and pneumonia 0.38-0.80. The distribution of these ranges correlates well with the model’s per feature auto-labeling efficiency and confidence metrics, shown in Fig. 5, with
cardiomegaly and pleural effusion showing the most inter-rater agreement, and pneumonia, pulmonary edema, and atelectasis showing the least.
c. Pulmonary Edema (total n=180)

d. Pneumonia (total n=100)

---

182

---

183
Fig. 2 | Automated-labeling model performance applied to three open-source CXR datasets, compared to consensus ground truth of 7 expert radiologists. We applied our xAI CXR auto-labeling model to three large open-source datasets; CheXpert, MIMIC, and NIH. For each of the five features – a cardiomegaly, b atelectasis, c pulmonary edema, d pneumonia, and e pleural effusion – we randomly selected a subset of “positive” and “negative” cases as determined by the model, distributed equally in each of ten pSim value ranges (0-0.1, 0.1-0.2, 0.2-0.3, …, 0.9-1.0), for expert review.

In panel 1 (upper left), the positive (light red) and negative (light blue) ratings for each of the 7 individual readers (columns A-G) are displayed graphically, with the consensus ground truth (GT, determined by majority) shown in the last column (bold red or bold blue).

In panel 2 (upper right), the “positive predictive values” (PPV=[true positive by GT]/[total positive by model], solid red triangles, y-axis left) and “negative predictive values” (NPV=[true negative by GT]/[total negative by model], solid blue circles, y-axis left), of the model’s ratings, are graphed versus the pSim threshold value that was applied by the model (x-axis). Also displayed in panel 2 (y-axis right) are the model’s “true positive capture rate” (TPCR, dotted red triangles) and “true negative capture rate” (TNCR, dotted blue circles), defined respectively as TPCR=[true positive (TP) by GT]/[total positive by GT (number bold red from panel 1)] and TNCR=[true negative (TN) by GT]/[total negative by GT (number bold blue from panel 1)].

In panels 3 (lower left) and 4 (lower right), respectively, the number of false positive (FP by GT) and false negative (FN by GT) cases rated by the model at each pSim threshold value (x-axis), are shown stratified by dataset (CheXpert, MIMIC, or NIH; total number cases positive or negative...
by the model in parentheses), with the “optimal”, lowest pSim threshold achieving 100% PPV or NPV, as indicated (bold green triangles).
Fig. 3 | AUROC performance of automated-labeling model at two different pSim threshold values, compared to sensitivity, specificity of individual expert radiologists and pooled public labels from three open-source CXR datasets. AUROC performance of our xAI CXR auto-labeling model applied to the CheXpert, MIMIC, and NIH open source datasets, is shown for each of the five labeled imaging features - a cardiomegaly, b atelectasis, c pulmonary edema, d pneumonia, and e pleural effusion. Comparison is to the performance of the individual expert radiologists (A-G, red circles), as well as to the performance of the pooled external annotations (blue squares, n=number available labeled external cases per feature). ROC curves (y-axis sensitivity, x-axis 1-specificity) are shown for both the “baseline” pSim=0 threshold (magnified box) and the “optimal” pSim threshold (i.e., the lowest pSim threshold achieving 100% accuracy, as per Fig. 2 panels 3 and 4).
a. Cardiomegaly: 39/50 (78%) positive cases with complete agreement

b. Atelectasis: 23/50 (46%) positive cases with complete agreement

c. Pulmonary Edema: 17/40 (43%) positive cases with complete agreement
d. Pneumonia: 14/50 (28%) positive cases with complete agreement

e. Pleural Effusion: 39/50 (78%) positive cases with complete agreement

Fig. 4 | Sample PA CXR images of positive examinations identified by the automated-labeling model, which had complete agreement between all 7 expert radiologists and the xAI model output, for each of the five features studied. Shown are fourteen sample PA CXR images with positive findings, for each of the five features studied (cardiomegaly a, atelectasis b, pulmonary edema c, pneumonia d, and pleural effusion e), for which there was complete agreement between both the xAI model output and the 7 expert radiologist interpretations. The pSim threshold values used by the model for auto-labeling feature detection are shown (a-e). For pneumonia, d, there were only 14 positive examinations, out of 50 total positive examinations identified by the model, which had complete agreement with all readers (14/50=28%), as per Fig. 2.1; percent agreement for the other features is as shown in the above captions (a-c, e).
Fig. 5 | Comparison of labeling efficiency/confidence metrics for each of the 5 features. For each of the five auto-labeled features—cardiomegaly (blue), pleural effusion (orange), atelectasis (gray), pulmonary edema (green), and pneumonia (yellow)—we compared: (i) the percent of positively auto-labeled CXR’s “captured” from the three pooled, full public datasets (i.e. “Pooled Capture%”, from Table 1, far right); (ii) the percent of cases with complete agreement between the model and all 7 expert readers (i.e. “Full Agree%”, from Fig. 4); (iii) the lowest pSim value such that PPV=1 (graphed as “1-pSim”, from Fig. 2, panel 3), and (iv) the lowest pSim value such that NPV=1 (graphed as “1-pSim”, from Fig. 2, panel 4). Features with higher y-axis values (e.g. cardiomegaly, pleural effusion) correspond to those with greater model auto-labeling efficiency/confidence; features with lower y-axis values (e.g. pneumonia, pulmonary edema) correspond to those with lesser model auto-labeling efficiency/confidence. Of note, in the graph for atelectasis, “1-pSim@PPV1” is higher than “1-pSim@NPV1”, which can be interpreted as greater confidence that the model is correct in “ruling-in” the feature (i.e. correctly auto-labeling true-positives) than in “ruling-out” the feature (i.e. correctly auto-labeling true-negatives); this relationship is reversed for the other 4 features (e.g. greater confidence that the model can correctly “rule-out” than “rule-in” pneumonia or pulmonary edema).
Fig. 6 | Pairwise kappa statistics between the 7 expert radiologists, for each of the 5 features. For each of the five auto-labeled features – cardiomegaly, pleural effusion, atelectasis, pulmonary edema, and pneumonia – the pairwise kappa statistics estimating inter-observer variability are shown in the respective color-coded matrices [11].
Table 1 | Automated-labeling model CXR “capture rate” applied to the three complete public datasets, at the optimized pSim threshold value for maximal accuracy per feature (PPV, NPV = 1; Fig. 2). Our automated-labeling model was applied to the three external open source CXR datasets: CheXpert, MIMIC, and NIH. Dataset composition and number of positively labeled, negatively labeled, and unlabeled PA CXR’s for each of the five features (cardiomegaly, atelectasis, edema, pneumonia, and pleural effusion) are shown (Table 1, left). The minimal “optimal” pSim threshold value, such that PPV, NPV = 1 for labeling accuracy, was selected as per Fig. 2 (Table 1, middle column). It is noteworthy that the model’s mean CXR “capture rates” for the pooled results from the three full public datasets (Table 1, far right), closely correspond to those shown in the graphs of Fig. 2.2 (a-e), for the subset of examinations (n=90-100) labeled by both the model and the expert radiologists.
Accurate, efficient annotation of large medical imaging datasets is an important limitation in the training, and hence widespread implementation, of artificial intelligence (AI) models in healthcare [12-21]. To date, however, there have been few attempts described in the literature to automate the labeling of such large, open-access databases [2-6]. One approach, for example, focused on developing new AI models using labor-intensive, manually annotated subsets of the external datasets, and applying these models to the remaining database [6]. In this study, we demonstrate a method for standardized, automated labeling based on similarity to a previously validated explainable AI model (xAI), using an atlas-based approach, for which the user can specify a quantitative threshold for a desired level of accuracy, the “probability-of-similarity” (pSim) metric.

Specifically, we applied our existing AI model for detection of five different CXR features (cardiomegaly, pleural effusion, atelectasis, pulmonary edema, and pneumonia), to three large public open-source datasets (CheXpert, MIMIC, and NIH), and compared the resulting labels to those of 7 human expert radiologists.

We showed that our xAI model, by calculating the pSim values for each feature based on comparison to its “remembered” training-set derived reference atlas, could automatically label a subset of the external data at a user-selected, arbitrarily high level of accuracy, equaling or exceeding that of the human experts (Fig. 3).

As shown in Fig. 2, the pSim value used for annotation reflects a trade-off between the accuracy of image labeling (i.e., the higher the pSim value, the more accurate the labels) and the efficiency of image labeling (i.e., the higher the pSim value, the fewer examinations that the model selects for annotation).
To evaluate the efficiency of our automated-labeling approach, we applied our xAI model to the three full public datasets, and compared the five auto-labeled features according to the following parameters: (a) the percent of positively auto-labeled CXR’s from the three pooled public datasets (i.e., the “capture rate”), (b) the percent of cases with complete agreement between the model and all 7 expert readers, (c) the lowest pSim value for annotation such that all positive cases captured are “true positive” (i.e., “optimal” pSim for PPV=1), and (d) the lowest pSim value for annotation such that all negative cases captured are “true negative” (i.e., “optimal” pSim for NPV=1). We found a strong correlation between the magnitude of these parameters for each of the annotated features, as shown in Fig. 5. It is noteworthy that the positive “capture rates” from the three pooled public datasets also strongly correlated with the “capture rates” graphed in Fig. 2.2, for the subset of examinations (n=90-100) labeled by both the model and the radiologist experts. Moreover, the parameter values reported for each feature corresponded well with the kappa values for inter-observer variability shown in Fig. 6.

Together, these results suggest that the overall accuracy and efficiency of the auto-labeling model, applied to the full public datasets at the “optimal” pSim for each feature, is similar to the accuracy and efficiency of the model as applied to the subset of examinations annotated by the 7 expert radiologists. These results also suggest greater auto-labeling efficiency, with higher confidence in label accuracy, for “cardiomegaly” and “pleural effusion” - two of the more objective findings in CXR interpretation - and lesser auto-labeling efficiency, with lower confidence in label accuracy, for “pneumonia” and “pulmonary edema” - two of the more subjective assessments in CXR interpretation. Indeed, the larger the quantity “1-pSim_{optimal}” for a given feature (where 0≤pSim≤1 and pSim_{optimal} = the minimum pSim value such that PPV/NPV=1), the more reliable and robust is the labeling for that feature, based on similarity to the “remembered” reference atlas derived from the model’s NLP training set.
A noteworthy aspect of our approach relates to system deployment. We can apply the pSim value threshold to each class independently, selecting a “low” pSim value for high conspicuity features with high inter-rater agreement, and selecting a “high” pSim value for noisier, more subjective non-specific features with lower inter-rater agreement, the latter at the cost of generating fewer labeled examinations (i.e., lower “capture rate”). Employing pSim values helps quantify which features of the AI model are most reliably annotated and which need to be improved, making it possible to measure system robustness. Deploying the xAI system is also HIPAA compliant, as no patient identifiable source data need be stored, since the mode selection (Fig. 1) uses only the encoded predicted probability distributions for categories and the compressed information from the UMAP transformation [22] for the atlas.

Another technical capability of our system is the “re-annotation” mode, which was not used in the current study, but which may be of value in real-world clinical settings. This mode is a component of the “explainability” functionality of our model, by which the system can prompt or query a human user if the pSim value for feature detection falls below a pre-selected threshold. More generally, the re-annotation mode has the potential to be applied to other medical AI models as a safety feature, alerting users that there is a measurable, quantitative probability of interpretive error in the AI model’s output.

Other current approaches to auto-labeling have involved semi-supervised [6, 23] and self-supervised [24-27] learning. Because these approaches assume low correlation between classes, however, their performance has not been validated for multi-label CXR classification models with high interclass correlation. Transfer learning and fine-tuning have also been attempted to improve performance when independently developed models are applied to external datasets [28-30], however, these methods are often impractical because different institutions are likely to use different definitions for similar categories, and capturing data with external labels based on even
slightly different definitions can introduce considerable noise when such data is used for training or re-training new models. Our approach, however, allows for generation of standardized labels, with a user defined probability of similarity to that of established models. Our atlas-based approach, which simplifies the computational issues by focusing on small patch regions with lower interclass and higher intraclass correlations, was capable of achieving high accuracy and efficiency for auto-labeling three large public open source CXR datasets, similar to or exceeding that of human experts.

Our auto-labeling AI model reflects several characteristics of human intelligence [31] in general, and radiologist-mimicking behavior in particular. Specifically, our system is “smart”, in that it can access its “memory” of examination features present in the training set, and quantitatively estimate their similarity to features in the new, external examination data. The “1-pSimoptimal” metric for each feature provides a measure of the “intelligence” of the system for efficient accurate labeling, and its value (between 0 and 1) reflects the quality (i.e., ground-truth accuracy) of the NLP-derived dataset used for initial training. The model can also provide feedback to users through its “explainability” functionality, by displaying examples of the features under consideration from its reference atlas together with their associated pSim value; this interaction offers the user an additional level of confidence that the model is doing what it’s supposed to do. In this regard, our system can be viewed as an “augmented” intelligence tool to improve the accuracy and efficiency of medical imagers.

Indeed, one limitation of our model is that its labeling accuracy and efficiency is directly proportional to the quality of the initial training set. This may help explain why cardiomegaly and pleural effusion - two high-conspicuity features routinely correctly described in the radiology reports identified by NLP for model training - have higher efficiency metrics (Fig. 5) than pulmonary edema and pneumonia, which are more non-specific and variably assessed by
different radiologists. This also may help explain why the “1-\text{pSim}_{\text{optimal}} values for NPV=1” in Fig.5 are higher than the “1-\text{pSim}_{\text{optimal}} values for PPV=1”, for all features except “atelectasis”, since atelectasis is a lower conspicuity, more non-specific feature typically noted in CXR radiology reports only when it is present, but not mentioned when it is absent (i.e., the model “learned” from its NLP derived training set to have a higher level of certainty, and hence a higher “1-\text{pSim}_{\text{optimal}}" value, when atelectasis is present, then when it is absent). Pulmonary edema and pneumonia, on the other hand, are typically described in CXR reports with a higher level of certainty when they are definitely absent (e.g., “no evidence of pulmonary edema or pneumonia”), than when they are possibly present (e.g., “cannot exclude pulmonary edema or pneumonia”).

Another limitation of our model is that our proposed xAI system requires substantial computational resources and storage space to provide the prediction basis and to operate the mode selection module. Because the explainable modules have been designed to operate independently, however, we can differentially deploy the xAI system of adjusted capabilities according to the specification of a given server.

In summary, we have developed and demonstrated an explainable AI model for automated-labeling of five different CXR imaging features, at a user selected quantitative level of confidence, based on similarity to the reference atlas library of an existing, validated model. The ability to automatically, accurately, and efficiently annotate large medical imaging datasets may be of considerable value in developing important, high-impact AI models that bring added value to, and are widely accepted by, the healthcare community. This approach might not only benefit re-training to improve the accuracy of existing AI models, but also help to standardize labeling of open-source datasets, for which the provided labels can be noisy, inaccurate, or absent. Such standardization may, in turn, reduce the amount of data required for accurate model building, facilitating training and re-training from initial small, but well annotated, datasets.
Methods

This study was compliant with the Health Insurance Portability and Accountability Act and was approved by the Institutional Review Board of the Massachusetts General Hospital.

Retrospective collection of the development and test datasets. The development dataset contained CXR images acquired between February 2015 and February 2019. All DICOM (digital imaging and communications in medicine) images were de-identified before data analyses. To make a consistent dataset, we chose only examinations that had associated radiology reports, view position information (e.g. AP/PA projections, “portable”, etc.), and essential patient identifiers (including but not limited to medical record number, age, or gender). If an examination had multiple CXR images, only a single CXR image was included. We randomly selected 1000 images for each view position as a test set; the remaining examinations, from non-overlapping patients, were separated into training and validation sets (Supplementary Fig. 1).

Labeling of the development and test datasets. The labels for the training and validation sets were determined exclusively from the automated NLP assignments, whereas those for the test set were determined by consensus of three U.S. board-certified radiologists at our institution (further details provided in Supplementary Table 1) using the “Mark-it” tool (https://markit.mgh.harvard.edu, MA, USA) for annotation [9].

Network training. Densely Connected Convolutional Network (DenseNet-121) [32], which connects each layer to all other layers in a feed-forward method, was selected to develop the 20 pathologic features detection and classification system. The pretrained model, available from the official repository in Pytorch [33, 34], was fine-tuned with our training dataset after the last fully connected layer with 1000 outputs and the first convolutional layer were replaced with 21 outputs.
The network topology was optimized using AdamW [35], where we used a batch size of 144, a learning rate of $1 \times 10^{-4}$, beta-1 of 0.9, beta-2 of 0.999, epsilon of $1 \times 10^{-8}$, and weight decay of $1 \times 10^{-5}$. In the training step, real-time data augmentation was performed by applying geometric transformations: rotation from -10 to 10, scaling to 110 %, random crop to 512x512, random horizontal flip with 1 % probability. All experiments were conducted on four GPUs of Tesla V100 SXM 32 GB [NVIDIA DGX, CA, USA], and all deep-learning models were implemented with Pytorch (v.1.2.0).

**Weighted loss function.** The Binary Cross-Entropy (BCE) loss function was weighted by the ratios of positive and negative samples for each class label ($\alpha_P^c$ and $\alpha_N^c$), in a similar fashion as described previously [4], for multi-label classification. We considered two additional weights: the first weight had to reflect the ratio of the number of effective samples ($\alpha^c_s$, the maximal sum number between positive and negative labels among 20 features divided by that of the c-th feature) to train because of consideration of ignore labels for each feature. When training the AI model, we experimentally found that using samples with the other view position as well as those with a targeted view position can improve the generalization performance of the model, so we added the second weight ($\alpha(v)$) in the loss function to relatively control the impact of samples with the target view position. The weighted BCE loss function is given by the equation (1):

$$L_{W-BCE}(x, y, t, v) = -\alpha(v) \sum_{c=1}^{J} \alpha^c_s \{\alpha_P^c t^c \ln y^c + \alpha_N^c (1 - t^c) \ln (1 - y^c)\},$$  

(1)

where $x$ denotes CXR images, the model’s output is $y = \{y^1, y^2, ..., y^J\}$ that indicates the predicted probability of $J$ classes, $v$ is a view position of the image, and $t = \{t^1, t^2, ..., t^J\}$ means the labels of features extracted by NLP. In addition, $\alpha^c_s$ is defined as $(|P^m_c| + |N^m_c|)/(|P^c| + |N^c|)$ in order to make fairness among classes with different numbers of effective samples which consider only ‘0’ and ‘1’, not ‘-1’. Here, $|P^c|$ and $|N^c|$ are the total numbers of ‘1’s and ‘0’s in labels
for c feature, and m means the class index having the maximum total number of both ‘1’s and ‘0’s
\( m = \arg \max_c (|P^c| + |N^c|) \). We also define \( \alpha^c_P = \frac{|P^c| + |N^c|}{|P^c|} \) and \( \alpha^c_N = \frac{|P^c| + |N^c|}{|N^c|} \) for solving the imbalance between positive and negative; \( \alpha(\nu) \) is set to \( \omega \) if \( \nu \) is the targeted view, 1 for the others.

**Design overview for quantitative, explainable, atlas-based system.** Our automated dataset labeling, based on similarity to a validated CXR AI model, requires calculation of two quantitative atlas-based parameters, the “patch similarity” and “confidence” probabilities (values between 0 and 1), as per Fig. 1. For the “patch similarity” computation, a patch atlas is generated based on class activation mapping (CAM) [7, 8]; for the “confidence” computation a distribution atlas is generated based on predicted probabilities (Fig. 1a,b). The harmonic mean between the patch similarity and confidence values are then used to calculate a “probability of similarity” (\( pSim \)) for each feature (Fig. 1c).

**Predicted probabilities, model ensemble, and distribution atlas creation.** To improve the robustness of the entire system, an ensemble of six DenseNet-121 models is composed using unweighted averaging, such that the final probability is determined as an average of probabilities predicted by the six models [36]. Those six models are constructed by independently training with three weights (i.e., \( \omega = 1.1, 1.5, \) and \( 2.0 \) in \( \alpha(\nu) \)) for PA view, then selecting two models maximized by AUROC and accuracy, respectively. To create the Distribution-atlas, we do inference with the trained AI model on a full training dataset, to obtain two probability distributions of positive and negative samples for the training dataset. These probability distributions are saved as the Distribution-atlas for each feature.

**Patch atlas creation based on CAM ensemble method.** To improve the localization performance of our class activation mapping, we developed an ensemble method as follows: by
removing noise components of a single CAM, adding only significant components, and normalizing it in equation (2), the ensemble CAM was able to highlight sharply the overlapping regions among the single CAMs.

\[
CAM_E^c = \text{Normalize}\left( \sum_{s=1}^{S} CAM_s^c \odot U_{\tau} \right),
\]

where \( CAM_E^c \) means the ensemble CAM matrix, \( CAM_s^c \) is a CAM matrix for the c-class generated from s-th single model, and \( S \) denotes the number of models. \( U_{\tau} \) denotes a matrix with the component of \( u_{i,j} = u(CAM_s^c(i,j) - \tau) \) to determine CAM values less than \( \tau \) as noise components and to remove them. \( u \) is a unit step function, \( \odot \) means the Hadamard product, and \( \text{Normalize} \) is a linear scale for converting into a standard range between 0 and 1.

To create the patch atlas, we search for main contours on a high-resolution CAM (512x512) generated from a CAM for each class, select a bounding box to include the outline, define it as the patch, and save it (one or two patches from a CAM are considered in this study). For each feature, patches are saved as typical, representative patterns from only the CXR images with the AI model’s predicted probability of being greater than or equal to 0.9. We train a cosine metric based UMAP model using the patches for all features [22]. The UMAP model transforms the patches into coordinates in two-dimensional embedding space, such that the smaller the Euclidean distance in this space, the higher the cosine similarity. For the automated labeling method, therefore, the patch atlas consists of coordinates for all patches in the two-dimensional embedding space and the UMAP model (Fig. 1b).

**Patch similarity value calculation.** To calculate the patch similarity as shown in Fig. 1b, we need to extract the Prediction-basis \( (\psi_{pb}^c) \) for the c-th feature by calculating Euclidean distance between the UMAP transformed coordinate of the test image and the Patch-atlas, and then by selecting K-basis with the minimum distance as equation (3):
where $\Omega_{pb}^c(k)$ denotes the patch with the $k$-th minimum Euclidean distance among the Patch-atlas, and the Euclidean distance is calculated by $\left\| f_{UMAP}^c(y_p^c) - A_{p-UMAP}^c(i) \right\|_2$ for $i = 1, \ldots, n(A_{p-UMAP}^c)$. Moreover, $f_{UMAP}^c$ is the trained UMAP model for c-class, $y_p^c$ is a 1024-dimensional patch vector calculated by a test image, $A_{p-UMAP}^c$ is the Patch-atlas, and $n(A_{p-UMAP}^c)$ is the size of the Patch-atlas. The patch similarity is proposed to enable the AI model to interpret the new patch based on the prediction-basis ($\Psi_{pb}^c$), as a quantitative metric. The metric is calculated by a percentile of how close a patch of a test image is on a prediction-basis of $K$ patches in the embedding space.

$$\text{patch similarity} = 1 - f_D^c \left( \frac{1}{K} \sum_{m=1}^{K} \left\| f_{UMAP}^c(y_p^c) - f_{UMAP}^c(\Omega_{pb}^c(m)) \right\|_2 \right),$$

where $f_D^c$ denotes a function calculating a percentile for the mean Euclidean distance of $K$-nearest patches for the test image, based on a distribution of the mean Euclidean distance for all patches of the Patch-atlas.

Confidence value calculation. As per Fig. 1b, we propose the confidence metric, based on the distribution atlas, as a measure of the trust level between the positive and negative predicted probabilities for a feature. This quantitative metric is simply defined with equations (5) and (6) for positive and negative predicted features, as follows:

$$Confidence_P = \max(f_P^c(y^c) - (1 - f_N^c(y^c)), 0)$$

$$Confidence_N = \max((1 - f_N^c(y^c)) - f_P^c(y^c), 0)$$

Assuming that a predicted probability is $y^c$ for c-class, we calculate a percentile ($f_D^c(y^c)$) in the positive Distribution-atlas and a percentile ($1 - f_N^c(y^c)$) in the negative Distribution-atlas. Then, the difference between two percentiles is calculated as the confidence. Because the predictive ability of the xAI model for each feature is related to the shape and degree of intersection of the
two probability density curves (positive and negative) on the distribution-atlas, the confidence metric, as defined based on equations (5) and (6), provides a quantitative measure analogous to a p-value between different statistical distributions. In other words, the higher the confidence value for a label, the higher the likelihood that the test image is mapping to the correct label, and the lower the likelihood of incorrect mapping. Moreover, this metric has the ability to quantify different levels of confidence according to different distributions of feature characteristics on the distribution atlas for each class of the model, even at the same predicted probabilities.

**pSim calculation, pSim threshold selection.** Our automated dataset labeling method calculates the pSim value using a harmonic mean between confidence and patch similarity ($pSimilarity$ in equation 7) for each test image.

$$pSim = \frac{2 \cdot confidence \cdot pSimilarity}{(confidence + pSimilarity)}$$  

(7)

The pSim threshold for each feature is chosen by the lowest pSim values that can achieve 100% PPV and NPV, as per Fig. 2.

An additional functionality of our model design includes a “mode selection” algorithm, which, using the selected pSim threshold value, can determine both the image label (positive, negative, or unlabeled) and a “mode” - self-annotation or re-annotation - or each test image, as per Fig. 1 and Supplementary Table 3. The “re-annotation” mode, which was not applied in this current study but can be used as part of the “explainability” functionality of the model, to prompt the system to alert or query a human user if the pSim value for a feature falls below a selected threshold.

**Statistical analyses.** To assess the statistical significance of the AUROC’s, we calculated 95% CIs using a non-parametric bootstrap approach via the following process: first, 1000 cases were randomly sampled from the test dataset of 1000 cases with replacement, and the DCNN models were evaluated on the sampled test set. After running this process 2,000 times, 95% CIs were
obtained by using the interval between 2.5 and 97.5 percentiles from the distribution of AUROCs. The 95% CIs of percentage accuracy, sensitivity and specificity of the models at the selected operating point were calculated using binomial proportion CIs.
References

[1] Lee, H., Yune, S., Mansouri, M., Kim, M., Tajmir, S.H., Guerrier, C.E., Ebert, S.A., Pomerantz, S.R., Romero, J.M., Kamalian, S., Gonzalez, R.G., Lev, M.H., Do, S. An explainable deep-learning algorithm for the detection of acute intracranial haemorrhage from small datasets. *Nature biomedical engineering* **3**, 173-182, (2019).

[2] Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K. and Seekins, J. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence* **33**, 590-597 (2019).

[3] Johnson, A.E., Pollard, T.J., Greenbaum, N.R., Lungren, M.P., Deng, C.Y., Peng, Y., Lu, Z., Mark, R.G., Berkowitz, S.J. and Horng, S. MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. Preprint at [http://arXiv:1901.07042](http://arXiv:1901.07042) (2019).

[4] Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., & Summers, R. M. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2097-2106 (2017).

[5] Bustos, A., Pertusa, A., Salinas, J.M. and de la Iglesia-Vayá, M. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical image analysis* **66**, 101797, (2020).

[6] Kim, T.K., Paul, H.Y., Hager, G.D. and Lin, C.T. Refining dataset curation methods for deep learning-based automated tuberculosis screening. *Journal of Thoracic Disease*, **12**(9), 5078-5085, (2020).

[7] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A. and Torralba, A. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2921-2929, (2016).
[8] Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D. and Batra, D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE international conference on computer vision, 618-626, (2017).

[9] Witowski, J., Choi, J., Jeon, S., Kim, D., Chung, J., Conklin, J., Longo, M.G.F., Succi, M.D. and Do, S. MarkIt: A Collaborative Artificial Intelligence Annotation Platform Leveraging Blockchain For Medical Imaging Research. Blockchain in Healthcare Today, (2021).

[10] Powers, D. Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation. Journal of Machine Learning Technologies, 2(1), 37–63 (2008).

[11] Stehman, S.V. Selecting and interpreting measures of thematic classification accuracy. Remote sensing of Environment, 62(1), 77-89, (1997).

[12] Lakhani, P., and Sundaram, B. Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. Radiology, 284(2), 574-582, (2017).

[13] Baltruschat, I. M., Nickisch, H., Grass, M., Knopp, T., and Saalbach, A. Comparison of deep learning approaches for multi-label chest X-ray classification. Scientific reports 9(1), 1-10, (2019)

[14] Pasa, F., Golkov, V., Pfeiffer, F., Cremers, D., & Pfeiffer, D. Efficient deep network architectures for fast chest X-ray tuberculosis screening and visualization. Scientific reports 9(1), 1-9, (2019).

[15] Wang, L., Lin, Z. Q., and Wong, A. Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images. Scientific Reports 10(1), 1-12, (2020).

[16] Rajpurkar, P., et al. CheXaid: deep learning assistance for physician diagnosis of tuberculosis using chest x-rays in patients with HIV. NPJ digital medicine 3(1), 1-8, (2020).

[17] Oh, Y., Park, S., and Ye, J. C. Deep learning covid-19 features on cxr using limited training data sets. IEEE Transactions on Medical Imaging 39(8), 2688-2700, (2020).
[18] Nam, J.G., Park, S., Hwang, E.J., Lee, J.H., Jin, K.N., Lim, K.Y., Vu, T.H., Sohn, J.H., Hwang, S., Goo, J.M. and Park, C.M. Development and validation of deep learning–based automatic detection algorithm for malignant pulmonary nodules on chest radiographs. *Radiology* **290**(1), 218-228, (2019).

[19] Sim, Y., Chung, M.J., Kotter, E., Yune, S., Kim, M., Do, S., Han, K., Kim, H., Yang, S., Lee, D.J. and Choi, B.W. Deep convolutional neural network–based software improves radiologist detection of malignant lung nodules on chest radiographs. *Radiology* **294**(1), 199-209, (2020).

[20] Sung, J., Park, S., Lee, S.M., Bae, W., Park, B., Jung, E., Seo, J.B. and Jung, K.H. Added Value of Deep Learning–based Detection System for Multiple Major Findings on Chest Radiographs: A Randomized Crossover Study. *Radiology*, 202818, (2021).

[21] Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J., & Oermann, E. K. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS medicine* **15**(11), e1002683, (2018).

[22] McInnes, L., Healy, J. and Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. Preprint at [http://arXiv:1802.03426](http://arXiv:1802.03426) (2018).

[23] Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., and Raffel, C. A. Mixmatch: A holistic approach to semi-supervised learning. In *Advances in Neural Information Processing Systems*, 5050–5060, (2019).

[24] He, K., Fan, H., Wu, Y., Xie, S. and Girshick, R. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9729-9738, (2020).

[25] Chen, T., Kornblith, S., Norouzi, M. and Hinton, G. A simple framework for contrastive learning of visual representations. In *Proceedings of International conference on machine learning*, 1597-1607, (2020).
[26] Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P. and Joulin, A. Unsupervised learning of visual features by contrasting cluster assignments. In Proceedings of Advances in Neural Information Processing Systems (NeurIPS), (2020).

[27] Hadsell, R., Chopra, S. and LeCun, Y. Dimensionality reduction by learning an invariant mapping. In Proceedings of the IEEE conference on computer vision and pattern recognition, 1735-1742, (2006).

[28] Apostolopoulos, I. D., & Mpesiana, T. A. Covid-19: automatic detection from x-ray images utilizing transfer learning with convolutional neural networks. Physical and Engineering Sciences in Medicine 43(2), 635-640, (2020).

[29] Shin, H.C., Roth, H.R., Gao, M., Lu, L., Xu, Z., Nogues, I., Yao, J., Mollura, D. and Summers, R.M. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE transactions on medical imaging 35(5), 1285-1298, (2016).

[30] Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. How transferable are features in deep neural networks?. Preprint at http://arXiv:1411.1792. (2014)

[31] Kolb, D. A. Experiential learning: Experience as the source of learning and development (FT press, 2014)

[32] Huang, G., Liu, Z., Van Der Maaten, L. and Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, 4700-4708, (2017).

[33] Deng, J., Dong, W., Socher, R., Li, L.J., Li, K. and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition, 248-255, (2009)

[34] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L. and Desmaison, A. Pytorch: An imperative style, high-performance deep learning library. Preprint at http://arXiv:1912.01703 (2019).
[35] Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. Preprint at http://arXiv:1711.05101 (2017).

[36] Ju, C., Bibaut, A. and van der Laan, M. The relative performance of ensemble methods with deep convolutional neural networks for image classification. Journal of Applied Statistics 45, 2800-2818, (2018).

[37] Langlotz, C. P. RadLex: a new method for indexing online educational materials. http://radlex.org/ (2006).

[38] Bird, S., Klein, E. and Loper, E. Natural language processing with Python: analyzing text with the natural language toolkit. (O'Reilly Media, Inc., 2009).

[39] Honnibal, M., & Montani, I. spacy 2: Natural language understanding with bloom embeddings. convolutional neural networks and incremental parsing, 7(1), (2017).
Supplementary Information

Supplementary Table 1 | Detailed information about our institutional datasets. We gathered examinations for both AP and PA view positions, and automatically generated three labels (positive, negative, ignore) by our natural language processing (NLP) tool which had been developed to convert terms used in radiology reports into the labels. This tool is a rule-based approach that has a dictionary of radiology terms, based on Radex Lexicon [37], and a mapping table of 20 pathological features and the key terms. Label generation of the NLP, using NLTK [38] and Spacy [39], employs two-staged processing for each examination: The first stage extracts radiology terms from a radiology report of the examination using the dictionary while considering both negation and double negation. In the second stage, the NLP determines labels from the terms by a mapping rule defining key terms for 20 pathological features. "Ignore" labels are assigned to features when NLP cannot classify the label if there are any conflicting terms in the second stage. The tables show labels' information for the training, validation, and test datasets which were exclusively divided by patients.

| Gender | Male / Female | Age average | std | Train set | Validation set | Test set |
|--------|---------------|-------------|-----|-----------|----------------|---------|
| AP view |               |             |     | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore | Negative | Positive | Ignore |Negative| Supplementary Information | 37 |
Supplementary Table 2 | Model performance. xAI model performance for 20 features was evaluated on PA test datasets of our institution (PA view in Supplementary Table 1) based on both NLP and human labels. The 95% CIs on the metrics were provided in parentheses (Statistical analyses in Methods). Out of 20 features, in this study, we selected 5 typical features that were commonly defined in three public datasets.

| Class                      | Test (NLP)          | Test (Human)        |
|----------------------------|---------------------|---------------------|
| Fracture                   | 0.77 (0.67, 0.86)   | 0.85 (0.77, 0.91)   |
| Non-fracture               | 0.91 (0.85, 0.96)   | 0.99 (0.98, 1.00)   |
| Diaphragm                  | 0.93 (0.90, 0.96)   | 0.97 (0.94, 0.99)   |
| Foreign body               | 0.92 (0.90, 0.94)   | 0.95 (0.93, 0.97)   |
| Aorta                      | 0.91 (0.88, 0.94)   | 0.94 (0.92, 0.96)   |
| Cardiomegaly               | 0.93 (0.91, 0.95)   | 0.97 (0.95, 0.98)   |
| Hilar area                 | 0.82 (0.58, 0.98)   | 0.94 (0.90, 0.98)   |
| Mediastinum                | 0.92 (0.88, 0.97)   | 0.98 (0.95, 1.00)   |
| Cavity/Cyst                | 0.87 (0.76, 0.95)   | 0.94 (0.97, 1.00)   |
| Emphysema                  | 0.94 (0.90, 0.98)   | 1.00 (0.99, 1.00)   |
| Atelectasis                | 0.90 (0.86, 0.91)   | 0.94 (0.93, 0.96)   |
| Nodule/mass                | 0.68 (0.60, 0.70)   | 0.80 (0.71, 0.88)   |
| Other interstitial opacity | 0.83 (0.78, 0.87)   | 0.94 (0.89, 0.98)   |
| Pulmonary edema            | 0.95 (0.91, 0.97)   | 0.98 (0.97, 0.99)   |
| Pneumonia                  | 0.77 (0.73, 0.82)   | 0.50 (0.88, 0.93)   |
| Decreased lung volume      | 0.92 (0.89, 0.94)   | 0.98 (0.96, 0.99)   |
| Increased lung volume      | 0.91 (0.86, 0.95)   | 0.98 (0.96, 0.99)   |
| Other pleural lesions      | 0.85 (0.76, 0.92)   | 0.98 (0.94, 1.00)   |
| Pleural effusion           | 0.97 (0.96, 0.98)   | 0.98 (0.97, 0.99)   |
| Pneumothorax               | 0.89 (0.78, 0.97)   | 0.95 (0.91, 0.98)   |

Supplementary Table 3 | Mode selection for the automated labeling method. The mode selection algorithm determines both a label (positive, negative, or unlabeled) and a mode (self-annotation or re-annotation) for each test image by using uses pSim and a selected pSim threshold values, where \( TH_{pos} \) was set to 0.5 in this study.

Mode selection algorithm

**Input:** predicted probability for c-class \( (y^c) \), Confidence\(_p\), Confidence\(_N\), and patch similarity

%[step-1] To divide into two groups by \( y^c \) and \( TH_{pos} \): positive or negative candidates

If \( y^c \geq TH_{pos} \): then

%[step-2] To decide mode and annotation for the positive candidates

% Probability of Similarity, \( pSim \)

\[ pSim = 2 \cdot \text{Confidence}_p \cdot \text{pSim} / (\text{Confidence}_p + \text{pSim}) \]

If \( pSim \geq pSim \text{ threshold value (PPV, NPV=1)} \): then

Mode = Self-annotation mode
Label = 1 %Positive label

Else
Mode = Re-annotation mode
Label = -1 %unlabeled

Else

%[step-2] To decide mode and annotation for the negative candidates

\( pSim = \text{Confidence}_N \)

If \( pSim \geq pSim \text{ threshold value (PPV, NPV=1)} \): then

Mode = Self-annotation mode
Label = 0 %Negative label

Else
Mode = Re-annotation mode
Annotation = -1 %unlabeled
Supplementary Figure 1 | CXR data collection for the XAI model development. Each dataset for AP and PA view positions collected and mapped one-on-one to the annotations extracted by our NLP tool from the corresponding radiological reports. For each view position, training, validation, and test datasets were divided without overlapped patients or duplicated cases.