Journal of Digital Imaging
https://doi.org/10.1007/s10278-019-00299-9

ORIGINAL PAPER

Crowdsourcing pneumothorax annotations using machine learning annotations on the NIH chest X-ray dataset

Ross W. Filice[1] · Anouk Stein[2] · Carol C. Wu[3] · Veronica A. Arteaga[4] · Stephen Borstelmann[5] · Ramya Gaddikeri[6] · Maya Galperin-Aizenberg[7] · Ritu R. Gill[8] · Myrna C. Godoy[3] · Stephen B. Hobbs[9] · Jean Jeudy[10] · Paras C. Lakhani[11] · Archana Laroia[12] · Sundeep M. Nayak[13] · Maansi R. Parekh[11] · Prasanth Prasanna[14] · Palmi Shah[6] · Dharshan Vummidi[15] · Kavitha Yaddanapudi[4] · George Shih[16]

Abstract

Pneumothorax is a potentially life-threatening condition that requires prompt recognition and often urgent intervention. In the ICU setting, large numbers of chest radiographs are performed and must be interpreted on a daily basis which may delay diagnosis of this entity. Development of artificial intelligence (AI) techniques to detect pneumothorax could help expedite detection as well as localize and potentially quantify pneumothorax. Open image analysis competitions are useful in advancing state-of-the art AI algorithms but generally require large expert annotated datasets. We have annotated and adjudicated a large dataset of chest radiographs to be made public with the goal of sparking innovation in this space. Because of the cumbersome and time-consuming nature of image labeling, we explored the value of using AI models to generate annotations for review. Utilization of this machine learning annotation (MLA) technique appeared to expedite our annotation process with relatively high sensitivity at the expense of specificity. Further research is required to confirm and better characterize the value of MLAs. Our adjudicated dataset is now available for public consumption in the form of a challenge.

Keywords Artificial intelligence · Machine learning annotations · Public datasets · Challenge · Pneumothorax · Chest radiograph

✉ Ross W. Filice
ross.w.filice@gunet.georgetown.edu

[1] Department of Radiology, MedStar Georgetown University Hospital, 3800 Reservoir Road, NW CG201, Washington, DC 20007, USA

[2] New York, NY, USA

[3] Department of Radiology, University of Texas MD Anderson Cancer Center, 1515 Holcombe Blvd. Houston, Houston, TX 77030, USA

[4] Department of Medical Imaging, University of Arizona, 1501 N. Campbell Ave, Tucson, AZ 85724, USA

[5] UCF College of Medicine, 6850 Lake Nona Blvd, Orlando, FL 32827, USA

[6] Department of Radiology and Nuclear Medicine, Rush University Medical Center, 1653 W Congress Parkway, Chicago, Illinois 60612, USA

[7] Department of Radiology, Perelman School of Medicine, Hospital of the University of Pennsylvania, 3400 Spruce Street, Philadelphia, PA 19104, USA

[8] Department of Radiology, Beth Israel Deaconess Medical Center, Harvard Medical School, 330 Brookline Avenue, Boston, MA 02112, USA

[9] Department of Radiology, University of Kentucky, 800 Rose Street, Lexington, KY 40536, USA

[10] Department of Diagnostic Radiology & Nuclear Medicine, University of Maryland School of Medicine, 22 S Greene Street, Baltimore, MD 21201, USA

[11] Department of Radiology, Thomas Jefferson University Hospital, 111 S 11th St, Philadelphia, PA 19107, USA

[12] Department of Radiology, University of Iowa, 3868 JPP 200 Hawkins Drive, Iowa City, IA 52242, USA

[13] Division of Thoracic Imaging, Department of Diagnostic Radiology, The Permanente Medical Group, Inc., San Leandro, CA 94577, USA

[14] Diagnostic Imaging Associates, 698 12th St SE, Suite 145, Salem, OR 97301, USA

[15] Department of Radiology, University of Michigan Health System, CVC 5581 1500 E Medical Center Drive, Ann Arbor, MI 48109, USA

[16] Department of Radiology, Weill Cornell Medicine, 525 E. 68th St., New York, NY 10065, USA

Published online: 25 November 2019

Background

Pneumothorax is a potentially life-threatening condition that warrants prompt recognition and potentially intervention [1]. Pneumothorax is diagnosed frequently in the ICU setting, with an incidence between 4 and 15% [1] and can also be seen spontaneously, albeit at a lower rate, affecting men more than women [2, 3]. Development of an artificial intelligence (AI) model to detect pneumothorax could help both in the interpretive setting to assist radiologists in detection, quantification, and tracking size over time, and in the non-interpretive setting where a more sensitive model might prioritize and triage radiology examinations that are suspicious for new or enlarging pneumothorax for more prompt review [4]. Such models could also be very useful in international, rural, or other settings where access to expert radiologists is limited.

AI in radiology is a nascent and potentially transformative technology that could augment clinical practice. Facilitating open image analysis competitions is an important mechanism to advance AI in various imaging spaces [5, 6]. Several recent challenges in radiology oriented around bone age [7], fracture detection [8], and pneumonia localization [9] have shown value in promoting collaboration between data scientists and radiologists to facilitate advances in algorithm development. The Society for Imaging Informatics in Medicine (SIIM) has recently opened a similar competition for pneumothorax detection and localization in collaboration with the American College of Radiology (ACR) and with support from members of the Society of Thoracic Radiology (STR) [10].

To create AI models capable of not only classifying exams as positive or negative for pneumothorax but also localizing the possible pneumothorax, typically require detailed bounding box or segmentation labels during training [9, 11] though saliency features or activation heatmaps derived from classification models can also provide localization information [12]. Bounding box or segmentation labels require time-consuming and cumbersome manual labeling by highly trained experts though new methods are being explored to augment this process [13, 14]. We similarly tested a workflow in which AI is implemented in parallel to expedite the labeling process. Approximately two thirds of a training dataset was manually labeled by a group of board-certified diagnostic radiologists, two AI machine learning annotation (MLA) models were developed based on those labels, and the resulting MLAs augmented the manual labeling of the remaining third of the dataset. All the human-approved annotations were then reviewed and adjudicated by subspecialty-trained thoracic radiologists with modification of the pneumothorax contours in a small percentage of cases as deemed appropriate. This annotated and adjudicated dataset is now available to the public [10] in hopes of advancing the translation of useful AI models into practice, and we believe our MLA methodology could further expedite development of additional highly curated AI datasets.

Methods

Chest radiographs were obtained from the National Institutes of Health (NIH) Chest X-ray Dataset of 14 Common Thorax Disease Categories [11]. This dataset includes nearly 111,000 weakly labeled frontal chest radiograph views in Portable Network Graphics (PNG) format. These weak labels were derived from radiology reports using natural language processing techniques with the understanding that there may be inherent labeling inaccuracies. A subset of 15,302 chest radiographs was derived from the weak labels to generate a balanced mix of positive, negative or normal, and intermediate disease states for the pneumothorax use case. A total of 5302 radiographs were weakly labeled with "pneumothorax," 5000 with "no findings," and 5000 that were not labeled pneumothorax or no findings. These PNG images were converted to the Digital Imaging and Communications in Medicine (DICOM) format and imported into an annotation platform.

We aimed to improve on the provided NIH labels by creating more specific and information-rich label definitions [Table 1]. Pneumothorax was defined as a freeform segmentation. "Normal" was an image-level weak label intended for radiographs that had normal cardiopulmonary findings but which still might have benign osseous findings such as degenerative disc disease or common medical instrumentation such as pacemakers or central venous catheters. "No pneumothorax/not normal" was an image-level weak label intended to capture all other abnormal cardiopulmonary findings such as pulmonary nodules, consolidations, and pleural effusions. "Chest tube" was defined as a line segmentation; chest tubes were labeled because their presence is correlated with pneumothorax, so we believed that any AI model derived from this data that is intended to detect *de novo* pneumothoraces would likely need to take this potential bias into account. "Question/exclude" was a weak label for exams that needed further review or possible exclusion such as lateral views or abdominal radiographs.

Six board-certified radiologists from five institutions [Table 2] with a mean of 8.2 years experience (range 1–12) participated in the annotation process. Annotation was

Table 1 Annotation categories for this project

Pneumothorax
Normal
No Pneumothorax/Not Normal
Chest Tube
Question/Exclude

Table 2 Society for Imaging Informatics in Medicine (SIIM) radiologists who participated in the initial annotation process, in alphabetical order by last name

Annotator	Institution	Years experience
SB	University of Central Florida	12
RWF	MedStar Georgetown University Hospital	9
PCL	Thomas Jefferson University Hospital	7
MRP	Thomas Jefferson University Hospital	1
PP	Diagnostic Imaging Associates	8
GS	Weill Cornell Medicine	12

performed using a web-based commercial annotation platform (MD.ai, New York, NY) which allowed for window/level adjustment as well as zoom and pan. Readers were blinded to other readers' annotations. Segmentation annotations (pneumothorax and chest tube) were performed by creating unlimited free-form bounding control points; the pneumothorax segmentations were then comprised of the area enclosed and the chest tube segmentations as the line defined by the control points. The control points could be adjusted and refined after annotation. It should be noted that for a singular class label of pneumothorax or chest tube, there could be multiple segmentations accounting for the possibility of multifocal or bilateral pneumothorax and more than one chest tube.

All the initial radiologists first annotated the same set of 100 randomly selected images during a warm-up period, and this process was reviewed as a group to ensure consistency with the process, familiarization with the program, and to answer any questions. Each radiologist was then assigned 1500–2000 cases for a total of 10,902 (approximately 71% of the entire set). Any image that was labeled as question/exclude was reviewed as a group and either updated with a label using consensus agreement or excluded (i.e., abdominal radiographs, lateral views, or images that excluded the majority of the chest and were not interpretable for this purpose). Forty-five exams (0.3%) were excluded leaving a total of 15,257 in the final dataset.

After annotation of the first 10,902 exams in the dataset, two separate Mask Regions with Convolutional Neural Networks (Mask R-CNN) [15, 16] models were trained on these manual labels to generate candidate MLAs for pneumothorax and chest tubes respectively. For the chest tube MLAs, the resulting mask was post-processed to a freeform line to match the training annotation format. These MLA models were then used to generate candidate annotations for the subsequent 4355 exams left in the dataset [Figs. 1 and 2]. The same six radiologists were assigned the remainder of the exams, divided roughly equally, to complete the annotation process. Candidate labels were displayed on the new images, and the radiologists could choose to approve, modify, or delete these labels and were also allowed to create new labels to finalize the annotations. The weak labels normal, no pneumothorax/not normal, and question/exclude were still applied manually though the presence of pneumothorax or chest tube MLAs could be used for guidance in this process.

Annotation timestamp data was extracted from the annotation platform in an attempt to assess how the MLA process affected workflow. Each MLA was assigned a unique machine learning model identifier so these could be distinguished from manual labels which had their own unique identifier. Some MLAs were updated by humans, and this could be reliably distinguished using the timestamps and metadata from when they were updated.

All annotations were then independently reviewed by 12 subspecialty thoracic radiologists followed by adjudication by an additional subspecialty thoracic radiologist from the STR. These 13 radiologists were from ten institutions with a mean of 11 years of experience (range 2–30) [Table 3]. For cases where both radiologist annotations and MLAs were present, STR radiologists were instructed to agree or disagree with annotations of radiologists and disregard the MLAs. STR radiologists modified pneumothorax contours where appropriate. STR radiologists did not assess the accuracy of chest tube labeling.

Final adjudication of the combined readings from both groups of radiologists was performed by one of the subspecialty thoracic radiologists (CCW) who was not involved with the independent review phase and had prior experience in adjudicating datasets [9]. Approximately 10% of the cases received additional adjudication (1490 exams). These cases either had disagreement between the two groups (i.e., one group labeled a pneumothorax and the other did not) or the STR reader agreed with the annotations on a case but for various reasons it was not clear to what they were agreeing. Annotations were reconciled in this adjudication process, but no further exams were excluded. Any MLAs that were not agreed with, modified, or otherwise owned by a human annotator were also discarded.

Results

A total of 15,302 exams were initially available, 45 were excluded, and 15,257 were annotated; 10,902 manually and 4355 with the assistance of MLAs. A total of 59,642 annotations were created; of these, 34,086 were either pneumothorax or chest tube. The first 10,902 cases were annotated with 11,147 manual pneumothorax and/or chest tube labels over the course of approximately 64 days. There was a break in the manual annotation process of approximately 33 days while the MLA models were generated. A total of 22,297 MLAs, of which 15,950 were pneumothorax and 6347 were chest tube, were generated and added to all 15,302 exams over the course of 2 days. Following the addition of these MLAs, in the

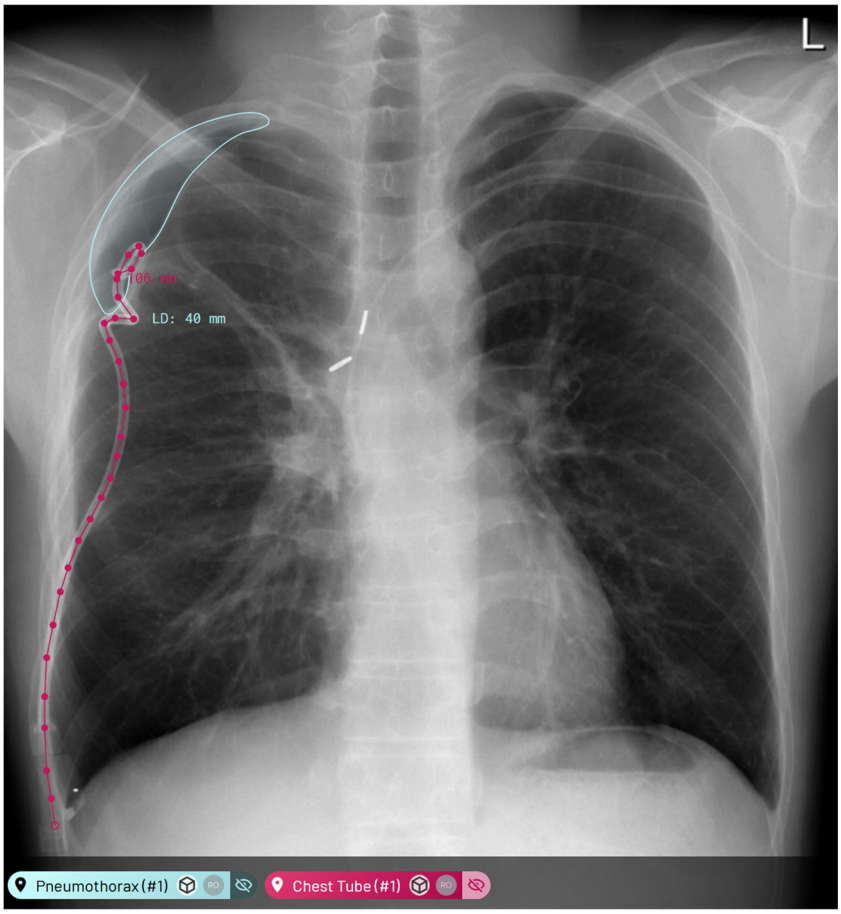

Fig. 1 Examples of proposed machine learning annotations (MLA) in the annotation platform where both chest tube and pneumothorax MLAs are true positives

remaining 4355 exams that were to be annotated, 2087 MLAs were modified (1084 pneumothorax; 1003 chest tube), 642 new labels were created manually (170 pneumothorax and 472 chest tube), with the remainder of the MLAs discarded. While the annotation process of the first 10,902 cases took 64 days, the group was able to complete the second set of 4355 cases in approximately 17 days including the 2 days required to deploy the MLAs (Fig. 3). The group of 12 STR radiologists reviewed all of the annotated cases over 62 days, after which the final STR radiologist reviewed and adjudicated the approximately 10% of cases with disagreements over a period of 5 days.

A focused subset MLA performance assessment found our chest tube MLAs to be more accurate than our pneumothorax MLAs with a mean area precision (mAP) of 77% and 46% respectively for annotations with a bounding box intersection over union (IOU) of greater than 50%. However, performance was lower when compared against radiologist verification. Precision was compared against radiologist-modified annotations and was lower at 46% and 22% for chest tube and pneumothorax, respectively. True positives were counted when MLAs were agreed with, false positives for those that were deleted, true negatives for those where no MLAs were proposed and no annotations were created manually, and false negatives for those where MLAs were not proposed but the radiologist created annotations. Based on these assumptions, sensitivity was much higher for the pneumothorax MLAs at 87% compared to a specificity of 35% while sensitivity and specificity were similar for chest tube MLAs at 76% and 78% respectively [Table 4].

While we did not gather enough data across the entire annotation process to properly calculate comparable performance metrics, we made a limited subjective assessment. A total of 22,297 MLAs were deployed across all 15,302 exams. Assuming a similar distribution across exams, this means approximately 6345 MLAs (4539 pneumothorax and 1806 chest tube) were deployed on the 4355 exams that remained after the manual annotation process. Of these, 2087 were modified (1084 pneumothorax and 1003 chest tube) which means the MLAs were correct or at least partially correct, 642 new annotations were created (170 pneumothorax and 472 chest tube) which means there were certainly some false negative MLAs, and approximately 3625 MLAs were discarded (3285 pneumothorax and 331 chest tube) indicating a relatively high rate of false positives in particular for pneumothorax. If we consider the 2087 modified annotations as true positives, the sensitivity of the MLA process overall is approximately 76%.

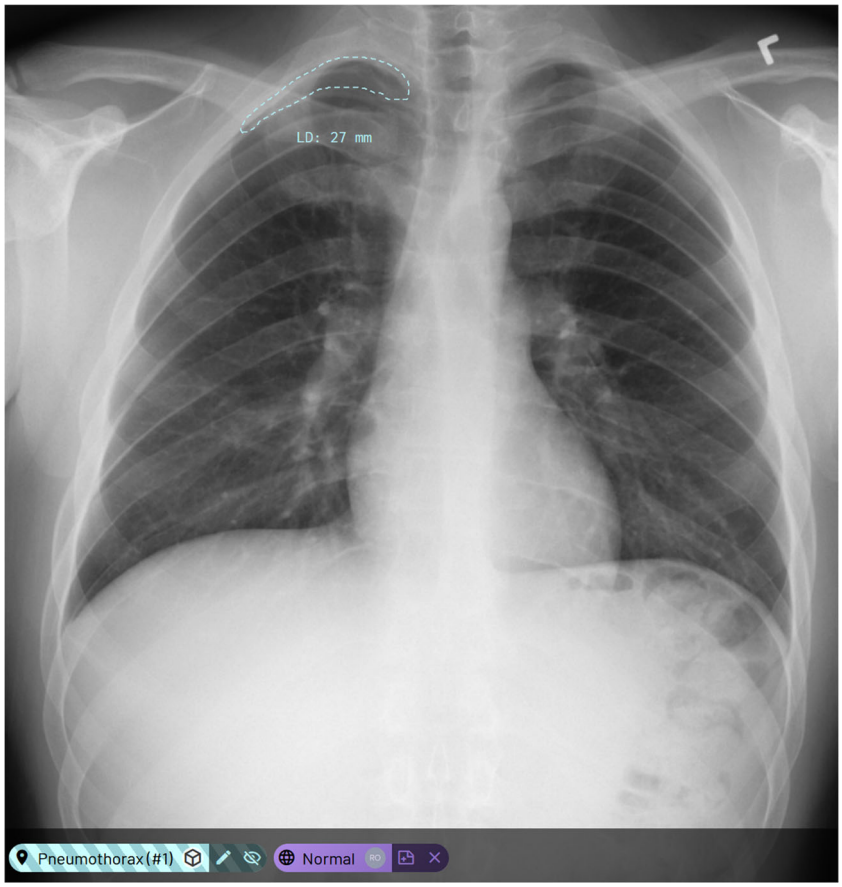

Fig. 2 Examples of a proposed machine learning annotation (MLA) in the annotation platform where the suggested pneumothorax is a false positive

The true negative rate is not definitively known, but even if one assumes a relatively high number of true negatives, the specificity is likely fairly low given the large number of false positives. Pneumothorax MLA sensitivity appears somewhat higher at the expense of specificity given the small number of false negatives and high number of false positives. Chest tube MLAs appear to have higher specificity given the relative low number of false positives. Clearly, there are limitations to this thought experiment, but the approximated performance characteristics are comparable at least in relationship between sensitivity and specificity to the more focused assessment described above.

Table 3 Society of Thoracic Radiology (STR) radiologists who participated in the secondary review and adjudication process, in alphabetical order by last name.

Annotator	Institution	Years experience
VAA	University of Arizona	10
MGA	Perelman School of Medicine at the University of Pennsylvania	7
MCG	UT MD Anderson Cancer Center	14
RG	Rush University Medical Center	2
RRG	Beth Israel Deaconess Medical Center	13
SBH	University of Kentucky	6
JJ	University of Maryland School of Medicine	15
AL	University of Iowa	9
SMN	The Permanente Medical Group, Inc.	30
PS	Rush University Medical Center	12
DV	University of Michigan	8
CCW	UT MD Anderson Cancer Center	12
KY	University of Arizona	5

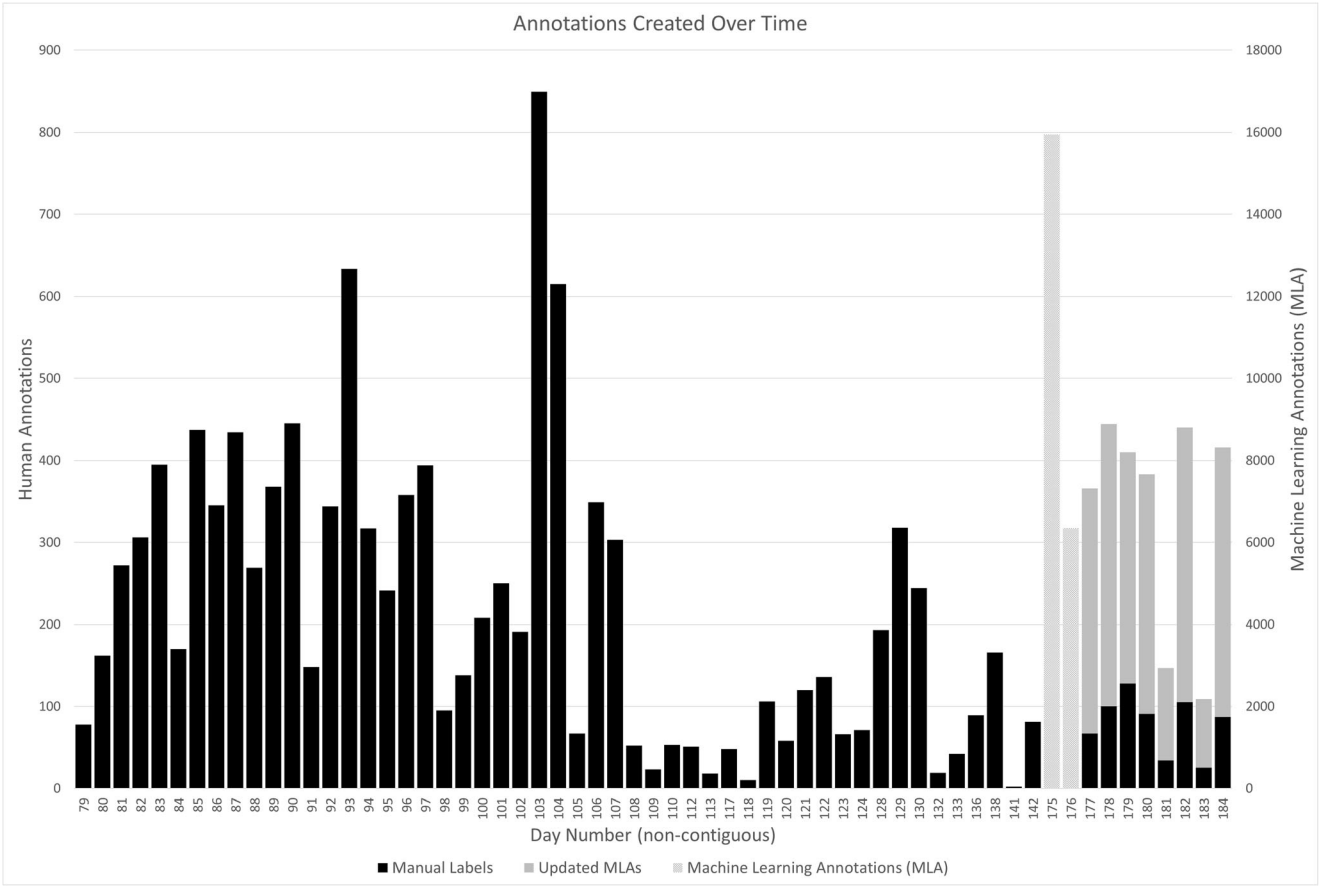

Fig. 3 Time required to complete the annotation process with the manual first half taking substantially longer than the second half after deployment of machine learning annotations (MLA)

Discussion

This annotation project contributes to existing work by extending available datasets with expert annotations and also by making the resulting data public to spur new AI model research and development. We also explore the concept of using AI model development during the course of an annotation project to expedite the remaining annotations.

The generated MLAs exhibit relatively high sensitivity at the expense of specificity based on our focused subset analysis. A similar pattern is suggested over the entire annotation process, but this evaluation is substantially limited as described above due to lack of needed detail. A prospective, carefully controlled analysis of MLA performance characteristics over the course of an entire annotation project would be an interesting and useful future direction; a similar assessment

Table 4 Performance characteristics of the Mask Regions with Convolutional Neural Networks (Mask R-CNN) machine learning annotations (MLA) based on a focused subset MLA analysis.

		Pneumothorax mask R-CNN Model	Chest tube mask R-CNN Model
PR values based on bounding box with IOU > 50%	mAP	0.46	0.77
	Sensitivity	0.47	0.77
PR values based on generated mask	Precision	0.22	0.46
	Sensitivity	0.87	0.76
	Specificity	0.35	0.78
	TP	0.15	0.15
	FP	0.53	0.18
	FN	0.29	0.62
	TN	0.02	0.05

could be performed to assess human performance by manually labeling a dataset followed by MLA deployment and adjudication. We believe relatively sensitive MLAs that sacrifice some specificity could be preferable for candidate annotation generation in this setting, but this has not been formally studied to determine how this affects label accuracy or quality. Finally, it appears more clear that interval MLA generation speeds the subsequent annotation process, but time must still be allocated to generate the MLA models and we did not control for other factors such as annotators becoming more facile with the annotation process over time.

Additional limitations of this work include the fact that the dataset is comprised of chest radiographs from a single source (NIH) and only frontal views were annotated; a more robust and generalizable dataset from additional institutions with annotation of lateral views could be valuable. Now that the dataset is public, this could facilitate the addition of more heterogeneous images and could allow further adjudication and refinement.

Conclusion

Publicly available training and testing datasets with highly curated and detailed annotations are important to further artificial intelligence in radiology. Machine learning annotations appear to expedite the rate-limiting, expensive, and cumbersome annotation process but further study is required to conclude this more broadly.

Acknowledgements Anna Zawacki from the Society of Imaging Informatics in Medicine (SIIM) for administrative support during the STR review process.

Compliance with ethical standards

Conflict of Interest The annotation platform used for this work was provided by MD.ai at no cost. Two authors (Anouk Stein, M.D. and George Shih, M.D., M.S.) serve as stakeholders and/or consultants for MD.ai.

References

1. Yarmus L, Feller-Kopman D: Pneumothorax in the critically ill patient. *Chest* 141(4):1098–1105, 2012
2. Gupta D, Hansell A, Nichols T, Duong T, Ayres JG, Strachan D: Epidemiology of pneumothorax in England. *Thorax*. 55:666–671, 2000
3. Onuki T, Ueda S, Yamaoka M, Sekiya Y, Yamada H, Kawakami N, Araki Y, Wakai Y, Saito K, Inagaki M, Matsumiya N. Primary and secondary spontaneous pneumothorax: prevalence, clinical features, and in-hospital mortality. *Can Respir J*. 2017
4. Lakhani P, Prater AB, Hutson RK, Andriole KP, Dreyer KJ, Morey J, Prevedello LM, Clark TJ, Geis JR, Itri JN, Hawkins CM: Machine learning in radiology: applications beyond image interpretation. *J Am Coll Radiol* 15(2):350–359, 2018 Feb
5. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg AC, Fei-Fei L. ImageNet large scale visual recognition challenge. arXiv: 1409.0575v3 [cs.CV] 30 Jan 2015.
6. Prevedello LM, Halabi SS, Shih G, Wu CC, Kohli MD, Chokshi FH, Erickson BJ, Kalpathy-Cramer J, Andriole KP, Flanders AE. Challenges related to artificial intelligence research in medical imaging and the importance of image analysis competitions. *Radiology: Artificial Intelligence*. 2019 Jan;1(1).
7. Halabi SS, Prevedello LM, Kalpathy-Cramer J, Mamonov AB, Bilbily A, Cicero M, Pan I, Pereira LA, Sousa RT, Abdala N, Kitamura FC, Thodberg HH, Chen L, Shih G, Andriole K, Kohli MD, Erickson BJ, Flanders AE: The RSNA pediatric bone age machine learning challenge. *Radiology*. 290(2):498–503, 2019
8. Rajpurkar P, Irvin J, Bagul A, Ding D, Duan T, Mehta H, Yang B, Zhu K, Laird D, Ball RL, Langlotz C, Shpanskaya K, Lungren MP, Ng AY. MURA: a large dataset for abnormality detection in musculoskeletal radiographs. *arXiv*: 1712.06957v4 [physics.med-ph] 22 May 2018.
9. Shih G, Wu CC, Halabi SS, Kohli MD, Prevedello LM, Cook TS, Sharma A, Amorosa JK, Arteaga V, Galperin-Aizenberg M. Augmenting the National Institutes of Health chest radiograph dataset with expert annotations of possible pneumonia. *Radiology: Artificial Intelligence*. 2019 Jan;1(1).
10. SIIM-ACR Pneumothorax Segmentation Challenge. https://www.kaggle.com/c/siim-acr-pneumothorax-segmentation/overview
11. Wang X, Peng Y, Lu L. ChestX-Ray8: hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases *arXiv*:1705.02315v5 [cs.CV] Dec 2017.
12. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A. Learning deep features for discriminative localization. *arXiv*:1512.04150 [cs.CV] 14 Dec 2015.
13. Bach S, Rodriguez D, Liu Y, Luo C, Shao H, Xia C, Souvik S, Ratner A, Hancock B, Al Borzi H, Kuchkal R, Re C, Malkin R. Snorkel Drybell: a case study in deploying weak supervision at industrial scale. *arXiv*:1812.00417v1 [cs.LG] 2 Dec 2018.
14. Dunnmon J, Ratner A, Khandwala N, Saab K, Markert M, Sagreiya H, Goldman R, Lee-Messer C, Lungren M, Rubin D, Re C. Cross-modal data programming enables rapid medical machine learning. *arXiv*:1903.11101 [cs.LG] 26 Mar 2019.
15. Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. *arXiv*: 1311.2524v5 [cs.CV] Oct 2014.
16. He K,Gkioxari G, Dollar P. Mask R-CNN *arXiv*:1703.06870v3 [cs.CV] Jan 2018.