 doi:10.20944/preprints202108.0564.v1

Article

Initial experience in developing AI algorithms in medical imaging based on annotations derived from an e-learning platform

Maurice Henkel[1], Christian Breit[1#], Patricia Wiesner[1], Jakob Wasserthal[1], Victor Parmar[1], Thomas Weikert[1], Verena Hofmann[1], Sebastian Eiden[1], Lena Schmuelling[1], Konrad Appelt[1], David Winkel[1], Fabiano Paciolla[1], Christian Lechtenböhmer[1], Moritz Vogt[1], Laurent Binsfeld[1], Raphael Sexauer[1], Christian Wetterauer[2], Kirsten D. Mertz[3], Alexander Sauter[1], Bram Stieltjes[1]

[1]Department of Radiology, University Hospital Basel, Switzerland
[2]Deparment of Urology, University Hospital Basel, Switzerland
[3]Deparment of Pathology, Kantonsspital Baselland, Switzerland
* Correspondence: author: Christian Breit, Radiology Department, University Hospital Basel, Spitalstrasse 8-12 4031 Basel; hanns-christian.breit@usb.ch

Abstract: Introduction. Development of supervised AI algorithms requires a large amount of labeled images. Image labelling is both time-consuming and expensive. Therefore, we explored the value of e-learning derived annotations for AI algorithm development in medical imaging. **Methods.** We have developed an e-learning platform that involves image-based single click labelling as part of the educational learning process. Ten radiology residents, as part of their residency training, trained the recognition of pneumothorax on 1161 chest X-rays in posterior-anterior projection. Using this data, multiple AI algorithms for detecting pneumothorax were developed. Classification and localization performance of the models was tested on an independent internal testing dataset and on the public NIH ChestX-ray14 dataset. **Results.** The AI models F1 scores on the internal and the NIH dataset were 0.87 and 0.44, respectively. Sensitivity was 0.85 and 0.80 for classification and specificity 0.96 and 0.48 for classification. F1 scores were 0.72 and 0.66, sensitivity 0.72 and 0.72. False positive rate was 0.36 and 0.32 for localisation. **Conclusion.** Our results demonstrated that e-learning derived annotations are a valuable data source for algorithm development. Further work is needed to include additional parameters such as user performance, consensus of diagnosis, and quality control in the development pipeline.

Keywords E-learning derived annotations; Pneumothorax; Artificial intelligence; Crowdsourcing; Educational data mining

Introduction

Development of AI models capable of not only classifying exams as positive or negative but also localizing possible findings like pneumothorax requires large amounts of labelled training data [1]. Image labelling is a time-consuming and laborious task that limits the amount of training data available [2]. The creation of training data by designated experts is neither a cost-effective nor a scalable approach. Crowdsourcing, the process of outsourcing a task to a network of many people, is a promising approach for the labelling of medical images that could drastically improve the ability to create large amounts of training data in a short time [3]. Previous work has demonstrated that labelling by non-experts is an alternative for the creation of large amounts of training data [1]. However, another study highlighted that the crowds' lack of medical knowledge, misunderstanding of the task or monetary incentives resulted in low quality of labels [4].

However, instead of including untrained users, the recognition and labelling of diseases could be an opportunity to attract medical students and residents to train their radiological skills while bringing a certain level of education to the annotation process. Due

 doi:10.20944/preprints202108.0564.v1

to an extensive theory-based education obtained in medical schools, students and residents are an enormous resource of medical knowledge. They understand the purpose, reason, limitations, and estimated results of a medical examination. However, they lack tools that enable them to translate their theoretical knowledge into practical skills [5-8]. By aligning practical training for students and residents with the labelling of diseases, an environment can be created that both helps them to improve their practical skills and at the same time creates high-quality datasets with the methods of crowdsourcing. Such an approach, derived from educational data mining (EDM), a research field that is concerned about gaining valuable information from educational environments, could transform the medical knowledge of students and residents into a useful resource for image annotation [9]. While the usage of educational data is not new, digitalization in education has boosted the number of available data points. Analysis of educational data provides valuable insights not only about the student but also about the content [10]. Educational data is already being used to train machine learning algorithms to personalize learning experience [11], continuously improve educational content [12] and advance e-learning technologies [13]. An e-learning application that involves image annotation could be used to locate and measure findings, information that is subsequently used for training of AI algorithms.

The objective of this study is to explore the value of e-learning derived annotations for the development of AI algorithms in medical imaging, exemplified on a chest X-Ray dataset of pneumothoraces.

Methods

E-learning platform

The e-learning platform is an in-house development of our radiology research group. The system is designed to expose users to large volumes of images in a short period of time and rapidly train their ability to recognize a given finding. Due to the high amount of effort required in the annotation of findings, a crowdsourcing approach to provide feedback was applied. Following this concept, users' responses are used to solve a case. Medical students in German-speaking countries are familiar with this way of solving cases. Most student councils have developed their own platform or use commercially available platforms to solve old exam questions via crowdsourcing [14].

The developed platform incorporates a web based DICOM viewer, Figure 1. The pseudonymized data is stored in a SQL database. The web-application is split into a UI and a server component. The front end provides an overview of learning sessions, a navigation bar, an image upload feature, and a “create session” feature. The server component contains the system architecture. The web-based DICOM viewer uses the open-source cornerstone.js library, which is developed by the Open Health Imaging Foundation (OHIF). Users annotate images as part of training sessions. The labels are stored in a SQL database along with the images. The software was deployed on a research server inside the hospital's IT network.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 August 2021 doi:10.20944/preprints202108.0564.v1

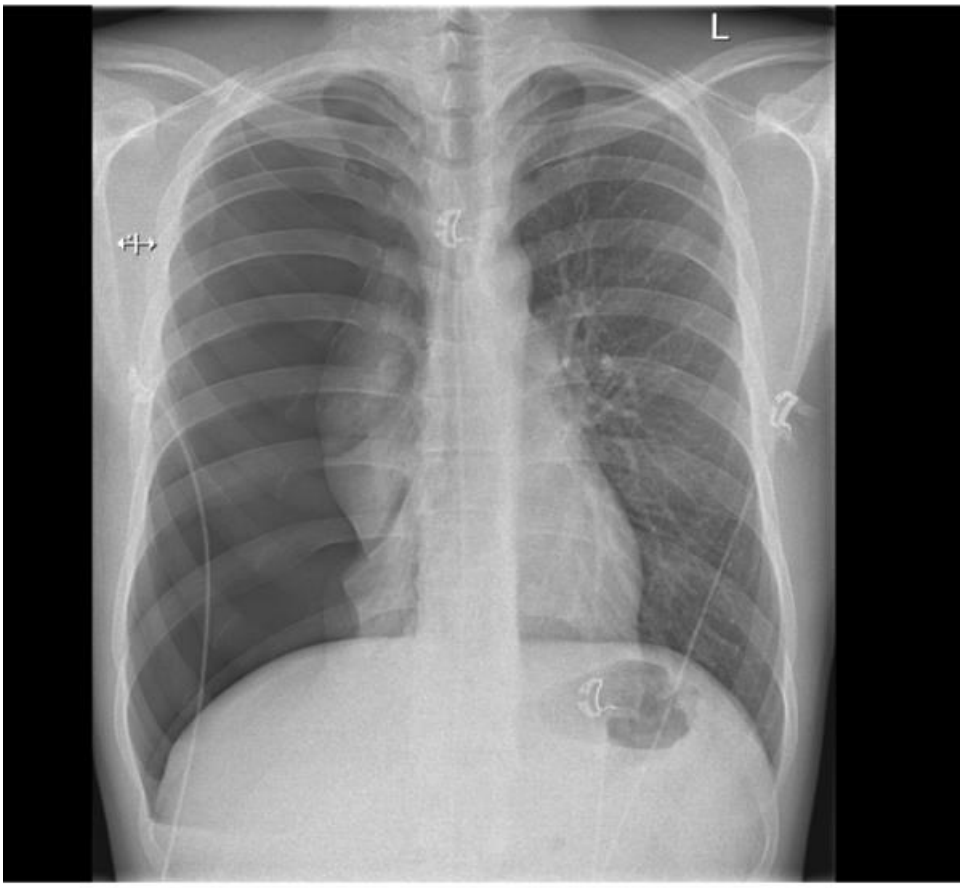

Figure 1. Screenshot of a learning session showing a pneumothorax case. The black area is an embedded, fully functional DICOM viewer. The viewer supports zooming, panning, greyscale, measuring HU and distances (ROI tool) and scrolling (sectional imaging).

E-learning experience

A learning session using the above platform contains collections of images to train the recognition of a specific finding (e.g., pneumothorax). During a session, trainees review images and are requested to label the decisive feature for diagnosis of the finding (e.g., visceral pleural edge for pneumothorax). Images are shown using the web based DICOM viewer that includes zoom and grey scaling functionalities. The labels are single-click annotations with a fixed size of 100x100 pixels, Figure 2. Thus, the labels do not represent a complete semantic segmentation, but what the individual user considers the most important area for detection of the finding. The user annotations are then stored in the database. Once a user has annotated an image, a heatmap calculated on the basis of all annotations of previous users is shown as an overlay of the original image. The more labels overlap, the higher is the value of an area in the heatmap, Figure 3. The overlay of the examination image with the heatmap allows users to compare their suggestions with that of the crowd. Of note, the users are blinded to other users' annotations as the heatmap gets calculated only after the individual users' final annotation. Users of the e-learning platform in our study are radiology residents with a mean working experience of 3 years (range 1 - 5 years) in diagnostic imaging.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 August 2021 doi:10.20944/preprints202108.0564.v1

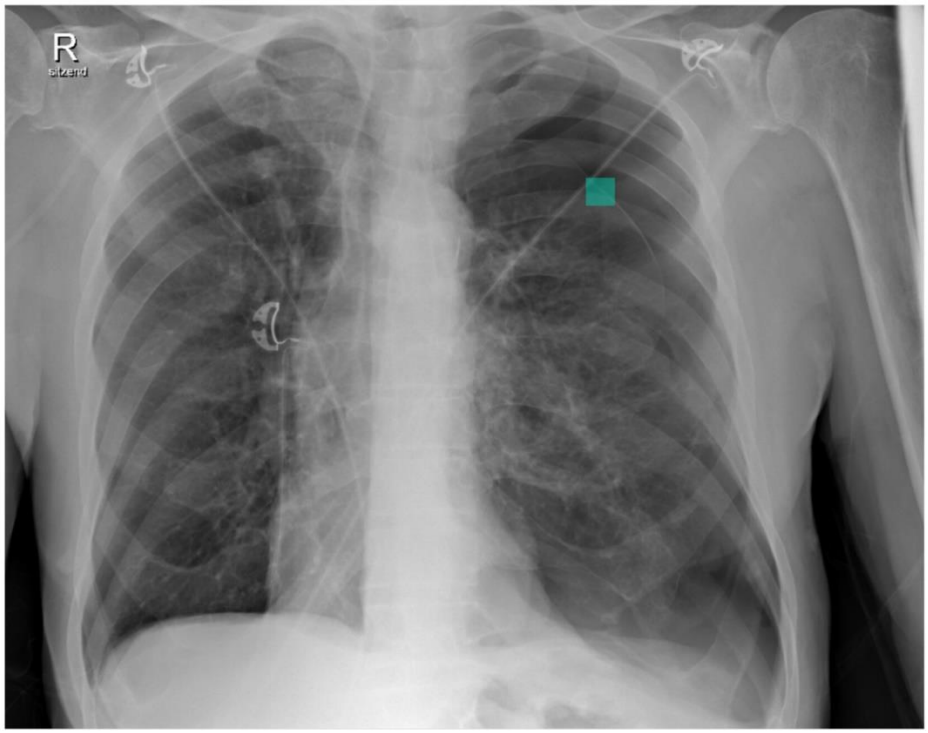

Figure 2. - Screenshot of a chest X-ray showing left sided pneumothorax and a single user's click label on pleura visceralis (rectangle).

Figure 3. - Screenshot of a chest X-Ray showing left-sided pneumothorax and the heatmap in plasma colour scale. Blue = zero annotations, yellow = maximum overlapping annotations, and purple all overlapping in between. The middle third of the visceral pleura of the left upper lobe received the most annotations, indicated in yellow, and therefore appeared to be the most informative area for detecting pneumothorax among users.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 August 2021 doi:10.20944/preprints202108.0564.v1

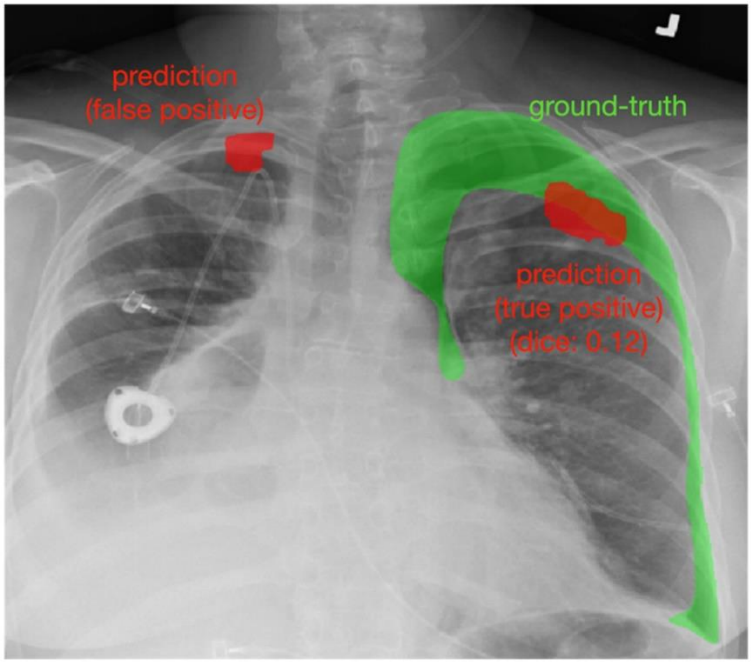

Figure 4. Example of ground-truth (green) and predictions (red) by our model on the external data set. The prediction can make up only a small part of the ground-truth prediction since the model was trained to predict the area where users thought the pneumothorax to be most visible.

All users received an introduction to the novel e-learning platform including opening a session, annotating the revealing finding, and interpreting the meaning of the calculated heatmap.

Internal data set

Images were identified by searching our research database using the RIS report and DICOM metadata. Chest X-rays of adult patients were obtained between 1st of January 2018 to 31st of December 2019.

Images for inclusion in the positive group were identified by using the search terms "pneumothorax", "standing position" and "posterior-anterior projection" on the written reports. To minimize false-positive selections, the terms "no pneumothorax", "pneumothorax absent" and "no sign of pneumothorax" were excluded.

Images for inclusion in the negative group were identified by searching for negative phrases such as "no pneumothorax", and no exclusion criteria were applied.. The images were visually reviewed for the presence and size of pneumothorax by a resident with 4 years of imaging experience. Due to the large image size and the limit of GPU memory, it was necessary to reduce the image batch size. We evaluated downsampling to 25% and 50% of the original image resolution. As a consequence, cases with less than 5 mm width of the finding were excluded.

External data set

A subset of the National Institutes of Health (NIH) Chest X-ray Dataset [15] was used to evaluate the AI model. The NIH dataset is a publicly available dataset of over 112,000 frontal chest radiographs accompanied by labels extracted from radiology reports using natural language processing. Here, we used the dataset that the group of Filice et al

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 August 2021 doi:10.20944/preprints202108.0564.v1

created consisting of 12,047 labelled frontal chest radiographs with 2669 fully segmented pneumothoraces [16].

AI model and training

The internal data set was randomly divided into 80% training and 20% testing. In a first step, two models were trained for the prediction of the presence of pneumothorax in the image (classification models). In a second step, four models were trained to predict where the pneumothorax is located in the image (localization model). The localization model was only trained on the samples of the internal data set which have a pneumothorax (using again a split into 80% training and 20% testing set).

In our study, the annotation did not represent a complete segmentation, but the most revealing area of the positive finding for the user. Overlapping annotations meant that more users recognized the finding in that particular area and this area might have a higher value for pneumothorax detection. Therefore, we trained half of the localization models on the entire annotation area, and the other half only on the area of overlapping annotations. To estimate the effects of down sampling on the ability for pneumothorax detection, we compared localization models trained on half and one-quarter of the original resolution. For our best localization model, we trained a 5-fold cross-validation on the internal dataset. The resulting 5 models were then combined as an ensemble by averaging their predictions. This further improved the predictions on the external dataset.

Classification Model

For the classification of the presence of pneumothorax an EfficientNet-b0 [17] pretrained with noisy student self-learning according to Xie et al. [18] was used. The batch size was set to 16, dropout was applied with a probability of 0.4, the model was trained for 100 epochs, the initial learning rate was set to 5e-4 and reduced over the training with a cosine scheduler. For data augmentation, the following transformations were applied to the images during training and a copy added with a probability of 0.3. The intensity of the transformation was also randomly sampled from the interval given for each transformation: Zoom (factor [0.8, 1.2]), contrast (gamma [0.5,1.5]), gaussian noise (mean 0, standard deviation 100), gaussian smoothing (sigma x/y [0.1, 0.8]), rotation (degree 90), mirroring. Images were resampled to the same spacing as the external NIH dataset and cropped to 1024x1024 pixels (same as the external NIH dataset). Before feeding the images to the model they were normalized to [-1, 1]. The model was implemented using Pytorch-lightning [19] and TIMM [20].

Localisation Model

For the localization of pneumothorax, we trained a nnU-Net [21] on our internal training dataset. The nnU-Net is a medical segmentation framework that automatically configures the data preprocessing as well as the hyperparameters for training a U-Net [22]. By optimising their segmentation pipeline across a range of several different medical segmentation challenges, the authors of the nnU-Net were able to derive heuristics for optimally setting the data pre-processing (e.g. normalization and resampling) as well as the U-Net configuration (e.g. number of layers and batch size) based on the characteristics of the input dataset. The automatically configured U-Net surpasses most submissions on over 23 public challenges. Thus, the nnU-Net has become the best solution for medical image segmentation. The annotations derived from our e-learning platform are no precise segmentation but indicate the location of the most revealing feature for detection of the pneumothorax. Therefore, it does not make sense to evaluate the pixel-wise segmentation performance. Instead, we evaluate the ability of the model to localize the pneumothorax ("detection"). The nnU-Net returns segmentation maps. By taking the following approach we used these segmentation maps to localize pneumothorax: The binary segmentation

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 August 2021 doi:10.20944/preprints202108.0564.v1

was post-processed by dilating it by 10mm and then eroding it by 7mm. This helped to fill small holes in the segmentation. Then connected component analysis was applied to convert the binary segmentation into instance segmentation since a subject can have multiple disconnected areas where a pneumothorax is visible. Small instances with a volume below 80mm² (this responds roughly to a diameter of 10mm) were removed to reduce the number of false positives.

A pneumothorax is counted as detected if the predicted segmentation overlaps at least 10% (in terms of dice score) with the ground-truth segmentation. A low value of 10% was chosen since our model does not segment the entire pneumothorax but only the area where the annotators rated it as most visible. This is only a subset of the entire pneumothorax as can be seen in the example in Figure 3.

Evaluation

The evaluation of our AI models is divided into two categories. First, the classification of radiographs between normal and abnormal CXRs with pneumothorax. Second, the localization of the area in the image containing the revealing feature (pleura visceralis). Evaluation of the models was performed on the internal testing dataset and validated with the NIH dataset. Since the result of the classification model is the binary decision between the presence or absence of pneumothorax, it was evaluated on cases with and without pneumothorax. For the internal evaluation the 20% split of the full data set was used. For external evaluation the full NIH data set was used. Since the results of the localization models is the prediction of the pneumothorax area, they were evaluated only on images with available annotations. For the internal evaluation the 20% split of annotated images were used. For external evaluation the segmented images from the Filice data set were used. We used sensitivity, specificity, and F1 score, which can be interpreted as a weighted average between sensitivity and specificity and makes it easier to compare models based on one metric. For localization the specificity is not defined as the number of true negatives can not be determined: In localization, we are not looking at single pixels but at objects and there is no meaningful way to derive negative objects (e.g., it is not clear how to define an object showing non-pneumothorax). Thus, true negatives cannot be determined. Instead, we used the average number of false positives per case.

Results

A total of 4394 pa chest radiographs, including 1161 with pneumothorax, were selected from our internal database. Since participants were asked to complete at least 1000 cases, the e-learning platform recorded 10769 annotations during learning sessions.

Classification performance

No relevant difference was found between the models trained to 25% and 50% image resolution. However, when applied to the external data set, a significant decline in performance was observed. The F1 score of both models dropped from 0.87 and 0.86 respectively when evaluated on the internal data set to 0.44 and 0.42 respectively when evaluated on the external data set.

Localization performance

The best performing localisation model was the one trained on full annotation area at 25% image resolution. The model showed a F1 score of 0.72, a sensitivity of 0.72, and false positive rate of 0.36 on the internal data set and a F1 score of 0.66, a sensitivity of 0.72, and false positive rate of 0.32 on the NIH data set. The annotation area was found to

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 August 2021 doi:10.20944/preprints202108.0564.v1

have a greater impact on the performance of the model than the image resolution. The difference in annotation area showed a up to 0.12 higher F1 score in favour of the models trained on the full annotation area. The difference in image resolution showed a up to 0.05 higher F1 score in favour of the models trained on 25% image resolution. Further training of the best performing model in 5-fold cross-validation did not substantially improve the F1 score. All results are summarized in tables 1 and 2.

Table 1. - Results of the classification models.

Classification Model						
	F1		Sensitivity		Specificity	
	Internal	External	Internal	External	Internal	External
25% resolution	0.870	0.441	0.849	0.796	0.963	**0.484**
50% resolution	0.855	0.423	0.810	0.901	0.969	0.330

Table 2. - Results of the localisation models.

Localisation Model						
	F1		Sensitivity		Avg False Positive	
	Internal	External	Internal	External	Internal	External
25% resolution All annotations	**0.724**	**0.660**	0.716	0.718	0.356	0.323
25% resolution Overlapping annotations	0.605	0.446	0.608	0.444	0.261	0.248
50% resolution All annotations	0.710	0.630	0.767	0.772	0.575	0.886
50% resolution Overlapping annotations	0.638	0.501	0.690	0.532	0.480	0.429
25% resolution All annotations Ensemble over Cross.Val.	-	0.669	-	0.686	-	0.345

Discussion

During the past two decades, educational data mining has emerged as an important resource to improve learning activities, educational content, and learning technologies [11]. Crowdsourcing through this process offers a novel, peer-generated approach for

 doi:10.20944/preprints202108.0564.v1

cost- and time-efficient generation of large-scale, high-quality data [23]. While much research on e-learning has focused on developing AI models to enhance the learning experience, little research has explored the application of e-learning derived AI models to real-world problems, e.g. in the field of medical diagnostics. In this study, we developed and validated AI models for classification and localisation of pneumothorax in chest radiographs trained on labels from an e-learning platform.

The classification models in our study showed a sensitivity of 0.85 and a specificity of 0.97, when evaluated on the internal data set and a sensitivity of 0.9 and a specificity of 0.48 when evaluated on the NIH ChestX-ray14 data set. This decline in specificity in the evaluation on the NIH dataset has already been reported in previous studies [24] and is partly attributed to the poor quality of the report generated labels [25]. Compared to previous studies, our models showed lower specificity in the evaluation on the NIH dataset. In the study of Taylor et al., classification models reached a specificity of 0.85 [24]. One explanation might be the size of the training data set, which was three times larger in Taylor et al. with 13,292 CXR. The best performing localization model in our study was the one trained on the full annotation area at a resolution of 25%. This model showed similar performance on the internal and NIH data sets, with F1 scores of 0.72 and 0.66, respectively, sensitivity of 0.72 and 0.72, respectively, and average false-positive rates of 0.36 and 0.32, respectively. A direct comparison to most previous studies evaluated on the NIH dataset is not possible, due to missing assessability of the Dice score in our approach. In comparison to similar studies, our model outperformed the model of Taylor et al. which showed a sensitivity of 0.49 [24] and the model of Wang et al. which showed an average false positive rate of 0.52 of [15]. In the sub analysis of our trained localization models, the annotation area seems to have a higher impact on the models' overall performance than the image resolution. The models trained on the full annotation area showed a higher F1 score, higher sensitivity and lower average false positive rate compared to the models trained on the overlapping annotation area. These results may not be surprising, as the models trained on the entire annotation space had more training data of the feature to be identified. Although these results emphasize the status of full-feature segmentation as the gold standard, our models trained on much less elaborate training data showed a reasonable performance. The comparison of the models trained at image resolutions of 25% and 50% showed that the image resolution seemed to have only a marginal impact on F1 score and sensitivity, while the average false positive rate was considerably lower at the 25% resolution. Considering that image resolution is a critical factor in the detection of pneumothorax, these results seem to be confusing [26]. However, the detection area is twice as large at 50%, thus increasing the probability of false positive detection.

Limitations

Our work has several limitations. First, It is important to restate that our intention is not to develop an algorithm that achieves the best performance. Currently, the greatest challenge in the development of AI models is the high effort involved in the preparation of training data. Our approach is intended to provide new opportunities to gain training data for development of AI models with reasonable performance as shown here. These algorithms could be used supportively to prioritize for expedited review. Second, currently the training data are based on data from a single institution. Although performance on the external NIH dataset was similar, it is difficult to predict exactly how well the algorithms would transfer to other institutions. Third, the "ground truth" in this study is based on the consensus opinion of the residents performing the annotation of the images; no expert review other than the approved report has taken place to confirm the diagnoses. Finally, this is a retrospective study, the model’s performance have not yet been prospectively evaluated in a clinical environment.

Conclusion

Overall, our initial experiences have shown that it is possible to train AI models based on e-learning derived annotations and that these models are able to achieve reasonable performance. These results are promising that educational data mining can be a valuable source to gain training data for machine learning. This approach could be a win-win for academic institutions as well as medical students and residents. By providing examinations in a radiological work environment, medical students and residents can train their practical skills in diagnostic imaging. In return, academic institutions receive valuable data for research activities such as the development of AI algorithms. In addition, the developed approach can open new ways to enhance the development of AI in diagnostic imaging. By integrating additional parameters, such as the times users spend on each image, assessment of individual user ability to identify the requested feature, and user's consensus on the feature, multiparametric models could be trained. Further work is needed to examine the value of these additional parameters for development of multiparametric AI.

Author Contributions: MH: VP developed the e-learning software. MH, JW developed the AI models. MH, PW created the data sets. CB, TW, VH, SE, LS, KA, DW, FP, CL, MV, LB, RS gener-ated the labels. CW, KDM, AS, BS advised in the conceptual design und reviewed the manuscript.

Funding: This study was funded by Innosuisse - Schweizerische Agentur für Innovationsförderung, Einsteinstrasse 2, 3003 Bern.

Institutional Review Board Statement: "Not applicable." This study does not investigate dis-eases or the human body and is therefore not subject to the Human Research Act. The study is ap-proved by the Ethikkommission Nordwest- und Zentralschweiz (EKNZ) and exempt from ethi-cal committee approval. All patient data used in this study were completely anonymized and are therefore not subject to the Data Protection Act.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Mehta, P., et al. *Segmenting The Kidney On CT Scans Via Crowdsourcing*. in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. 2019.
2. Rajpurkar, P., et al., *Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning.* arXiv preprint arXiv:1711.05225, 2017.
3. Iqbal, T., A. Shaukat, and U. Akram, *Automatic Diagnosis of Pneumothorax from Chest Radiographs: A Systematic Literature Review.* arXiv preprint arXiv:2012.11214, 2020.
4. Heim, E., et al., *Large-scale medical image annotation with crowd-powered algorithms.* Journal of Medical Imaging, 2018. **5**(3): p. 034002.
5. Bain, P., A. Wareing, and I. Henderson, *A review of peer-assisted learning to deliver interprofessional supplementary image interpretation skills.* Radiography, 2017. **23**: p. S64-S69.
6. Bhogal, P., et al., *Radiology in the undergraduate medical curriculum — Who, how, what, when, and where?* Clinical Radiology, 2012. **67**(12): p. 1146-1152.
7. Kourdioukova, E.V., et al., *Analysis of radiology education in undergraduate medical doctors training in Europe.* European Journal of Radiology, 2011. **78**(3): p. 309-318.
8. Mirsadraee, S., et al., *Radiology curriculum for undergraduate medical studies—A consensus survey.* Clinical Radiology, 2012. **67**(12): p. 1155-1161.

9. Romero, C. and S. Ventura, *Educational data mining: a review of the state of the art.* IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 2010. **40**(6): p. 601-618.
10. Charitopoulos, A., M. Rangoussi, and D. Koulouriotis. *Educational data mining and data analysis for optimal learning content management: Applied in moodle for undergraduate engineering studies*. in *2017 IEEE Global Engineering Education Conference (EDUCON)*. 2017.
11. Romero, C. and S. Ventura, *Educational data mining and learning analytics: An updated survey.* Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2020. **10**(3): p. e1355.
12. Colchester, K., et al., *A survey of artificial intelligence techniques employed for adaptive educational systems within e-learning platforms.* Journal of Artificial Intelligence and Soft Computing Research, 2017. **7**(1): p. 47-64.
13. Baker, R.S. and K. Yacef, *The state of educational data mining in 2009: A review and future visions.* JEDM| Journal of Educational Data Mining, 2009. **1**(1): p. 3-17.
14. Tobias Odendahl, C.S., Christian Rubbert. *Kreuzmich*. unknown; Available from: https://kreuzmich.de/.
15. Wang, X., et al., *ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases.* 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017: p. 3462-3471.
16. Filice, R., et al., *Crowdsourcing pneumothorax annotations using machine learning annotations on the NIH chest X-ray dataset.* Journal of Digital Imaging, 2019. **33**: p. 490-496.
17. Tan, M. and Q. Le, *EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks*, in *Proceedings of the 36th International Conference on Machine Learning*, C. Kamalika and S. Ruslan, Editors. 2019, PMLR: Proceedings of Machine Learning Research. p. 6105--6114.
18. Xie, Q., et al. *Self-training with noisy student improves imagenet classification*. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020.
19. William Falcon, J.B.A.W.N.E.J.S.J.J.N.S.I.d.V.B.E.H.T.M.P.Y.S., *PyTorchLightning/pytorch-lightning: 0.7.6 release (Version 0.7.6).* 2020, Zenodo.
20. Ross Wightman ; Chris Ha; Mike; Csaba Kertész; Dushyant Mehta; Kushajveer Singh; Andrew Lavin; Matthijs Hollemans; Vyacheslav Shults; Yusuke Uchida; Zhun Zhong; Kim, T.m., *rwightman/pytorch-image-models: v0.4.5. Lots of models. NFNets (& NF-ResNet, NF-RegNet), GPU-Efficient Nets, RepVGG, VGG.* Zenodo, 2021.
21. Isensee, F., et al., *nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation.* Nature Methods, 2021. **18**(2): p. 203-211.
22. Ronneberger, O., P. Fischer, and T. Brox. *U-Net: Convolutional Networks for Biomedical Image Segmentation*. in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. 2015. Cham: Springer International Publishing.
23. Alenezi, H.S. and M.H. Faisal, *Utilizing crowdsourcing and machine learning in education: Literature review.* Education and Information Technologies, 2020. **25**(4): p. 2971-2986.
24. Taylor, A.G., C. Mielke, and J. Mongan, *Automated detection of moderate and large pneumothorax on frontal chest X-rays using deep convolutional neural networks: A retrospective study.* PLoS medicine, 2018. **15**(11): p. e1002697.
25. L., O.-R. *Exploring the ChestXray14 dataset: problems.* 2017; Available from: https://lukeoak-denrayner.wordpress.com/2017/12/18/the-chestxray14-dataset- problems/.

 doi:10.20944/preprints202108.0564.v1

26. Yarmus, L. and D. Feller-Kopman, *Pneumothorax in the Critically Ill Patient.* CHEST, 2012. **141**(4): p. 1098-1105.