ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1. Available from http://www.i6doc.com/en/.

Improving Intensive Care Chest X-Ray Classification by Transfer Learning and Automatic Label Generation

Helen Schneider*[1], David Biesner*[1,2], Sebastian Nowak[3], Yannik C. Layer[3], Maike Theis[3], Wolfgang Block[3], Benjamin Wulff[1], Alois M. Sprinkart[3], Ulrike I. Attenberger[1,3], and Rafet Sifa[1]

[1] Fraunhofer IAIS - Department Media Engineering
[2] University of Bonn - Department of Computerscience
[3] University Hospital Bonn - Department of Diagnostic and Interventional Radiology
* First authors, equal contribution

Abstract. Radiologists commonly conduct chest X-rays for the diagnosis of pathologies or the evaluation of extrathoracic material positions in intensive care unit (ICU) patients. Automated assessments of radiographs have the potential to assist physicians by detecting pathologies that pose an emergency, leading to faster initiation of treatment and optimization of clinical workflows. The amount and quality of training data is a key aspect for developing deep learning models with reliable performance. This work investigates the effects of transfer learning on public data, automatically generated data labels and manual data annotation on the classification of ICU chest X-rays of the University Hospital Bonn.

1 Introduction and Background

Patients in critical condition often require intensive monitoring due to a variety of organ dysfunctions. Chest X-rays assist ICU physicians in identifying pathological findings and assessing the location of extrathoracic materials. Automated analysis using image-based deep learning methods has the potential to optimize clinical workflows by automatically identifying potentially critical conditions immediately after image acquisition, potentially speeding treatment initialization. However, creating reliable models requires large image datasets with accurate annotations for training. This can be challenging, primarily due to the need for experts (i.e. radiologists) to handle the time-consuming manual task of annotating new image data. There are multiple possibilities to reduce the amount of manually annotated 'gold' data needed to train image classification models.

To decrease the annotation effort, one can use the patient reports associated with the images. The patient reports were written by experts and already include a classification of the image. Manual annotation of text which carries over to image data can then be realized with the help of medical students. In order to be able to implement automatic annotation of patient reports, these must be available in structured text form, which is rarely the case. However rule-based text labeling systems can automatically classify patients reports and therefore create silver labels for the corresponding images on which image based model can be trained. The reliability of 'silver' labels is highly dependent on the

ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1. Available from http://www.i6doc.com/en/.

performance of the annotation systems, so findings with variable descriptions might not be labeled accurately and classification performance of trained image models might suffer.

A second way to increase the amount of training data is using transfer learning on publicly available annotated chest X-ray datasets [1, 2]. However, many concerns on the data quality of these datasets have been raised [3] and the data distribution of the public dataset is likely to differ from the target application in terms of available labels or data domain.

A reasonable way to alleviate these data issues is to combine transfer learning on public data sets with learning on automatically generated labels. The models are pre-trained on public and silver datasets, and the manually annotated gold labels are used to fine-tune the weights. Within this study we investigate whether models for the classification of thoracic diseases and the location of extrathoracic material in ICU chest X-rays can be improved by these strategies, whether manual annotation is necessary for training sufficiently performative models and how the number of manual annotations affects the classification performance.

2 Related Work

The analysis of chest X-rays for automated detection of diseases is an ongoing field of study. The publicly available CheXpert dataset, based on automatic generated labels, has enabled the development of various machine learning methods for the classification of chest diseases, e.g. implementation of deep neural networks or label smoothing methods [1, 4]. Different labeler approaches were implemented, and their influence on the classification of X-rays was studied [5]. None of the above investigates the improvement of silver label training through gold finetuning. Additionally, all studies based on the CheXpert dataset reside in a different data domain from ICU chest X-rays, in which images are taken in a more controlled environment. See Section 3 for details.

The potential of deep learning models for the evaluation of ICU chest X-rays was investigated in [6, 7], but no automatically annotated labels and their influence on performance were considered. Few studies investigate the impact of transfer learning and automatically generated labels on the analysis of ICU chest X-rays, taking into account not only disease detection but also extrathoracic material location assessment. We would like to fill this gap with our contribution. In addition this is the first attempt to categorize the ICU chest X-ray dataset provided by the University Hospital Bonn.

3 Data

We investigate the training performance of our models on in-house and public datasets.

For the in-house dataset, written informed consent was waived due to the retrospective nature of the study with institutional review board approval (AZ 411/21). A total of 19 059 text-based reports, which had been created by ra-

ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1. Available from http://www.i6doc.com/en/.

diologists in the course of the treatment, were processed by two research assistants who linked the findings written in text to the associated image data, therefore creating gold labels for training. The annotated targets consist of the findings 'pulmonary infiltrates', 'pleural effusion', 'pulmonary congestion' and 'pneumothorax', as well as the foreign object locations 'regular location of central venous catheter (CVC)' and 'misplaced location of CVC' for a multilabel classification on 6 classes. From the gold labeled data set, 1527 images are randomly selected to form a validation set, 1697 scans serve as a test set and 15 835 are considered during training. For additional evaluation, a test set with 483 samples is available, which was manually annotated by a radiology resident.

For an additional 56 324 reports silver labels were created by a rule-based text classification system. None of the silver labeled images are integrated into the validation or test sets. We evaluate the performance of the automatic labeling on the gold labeled test set and present the F1-scores for the individual indications in Table 2. We see that the quality of the automatic labels varies from very good for most (e.g. 95% F1 for 'Pneumothorax') to rather poor (36.9% F1 for 'Misplaced CVC Location'). There are no patient overlaps between the training, validation and test sets.

The CheXpert dataset was used for pretraining on a public in-domain dataset [1]. CheXpert consists of $220k$ images of chest X-rays, with a total of 14 observations labeled positive, negative, uncertain or not mentioned for each image, generated automatically by rule-based labeling tools. See [1, 3, 8] for further details and an evaluation on the quality of the generated labels. For pretraining we train on the training data split, removing all 'lateral' view images for a total of 191 027 training images. Regarding the uncertainty mapping, the approach of [9] was adopted, in which the uncertain labels are mapped to either positive or negative depending on the indication class. Note that while CheXpert, like our in-house dataset, consists of images of thorax X-rays, the domain differs: While CheXpert consists thoracic scans of patient staying upright in a standardized imaging setting, our dataset contains only anterior-posterior thorax images acquired in the ICU, which are taken with a portable chest radiograph while the patient is lying down in a hospital bed.

4 Models and Training

All experiments in this study are based on the same model architecture DenseNet 121 [10] with a single layer sigmoid classifier. The images of the public and in-house dataset are resized to 224×224 pixels and normalized using the variance and mean of the ImageNet dataset. We achieve a validation AUROC of 88.0% on CheXpert, in comparison state-of-the-art methods which achieve 89.0% for single models or 94.0% for ensemble models [4]. We therefore consider the architecture appropriate to investigate influence of training and data on classification performance.

To investigate the effect of label quality and data domain on classification performance, we train a variety of different models.

ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1. Available from http://www.i6doc.com/en/.

In the first approach, only images with manually annotated gold labels are taken into account during training, the resulting model is referred to as M_{G}. To evaluate the absence of manual labels, we train the models M_{S} and $M_{\mathrm{C+S}}$ on the automatically generated silver labels, where $M_{\mathrm{C+S}}$ is additionally pretrained on the public CheXpert data. To investigate the different data domains, we train the models $M_{\mathrm{S+G}}$, which is pretrained on the silver labels and finetuned on the gold labels, and $M_{\mathrm{C+G}}$, which is pretrained on CheXpert and finetuned on the gold labels. Finally we train a model $M_{\mathrm{C+S+G}}$ on all available data, first pretraining on CheXpert, then on the silver labels and finally finetuning on gold labels.

For training we additionally restrict the gold training data in regular intervals between $N = 500$ and the full $N = 15\,835$ images to simulate the effect of less manual annotation.

When pretraining on CheXpert and finetuning on the in-house datasets, we keep the architecture and weights of the convolutional layers of the model and re-initialize the classifier for the new number of classification targets. When moving from silver to gold labels for finetuning, we keep the entire model.

For each model training we apply the Adam optimizer, with learning rate scheduling dependent on the training data. For initial training, a one cycle learning rate scheduler with a maximum learning rate of 1.0e−02 is used. For finetuning, a reduce-on-plateau learning rate scheduler is applied with an initial learning rate of 1.0e−03. More detailed information will be found in the training code available at `https://github.com/fraunhofer-iais/ICU-Chest-X-ray-evaluation` upon publication. Each pretraining and finetuning is run until convergence of the training loss, we use early stopping to select the model with the optimal validation set AUROC score for evaluation or further finetuning.

5 Results

The results of the different learning approaches are presented in Table 1. Looking at the supervised learning approach, it can be seen that as the number of images increases, so does the performance of the model M_{G}. For all N except the smallest we see a clear improvement by including pretraining on the CheXpert dataset with model $M_{\mathrm{C+G}}$. We also see that $N = 500$ images are not enough in-domain data for successful training.

For the training without gold labels, model M_{S} achieves an AUROC score of 75.3% , which can be improved to 75.5% by transfer learning in model $M_{\mathrm{C+S}}$. Table 2 shows the results of model $M_{\mathrm{C+S}}$ for the individual targets. We oberve that the performance of training only on silver labels is comparable to the training on only gold labels. The higher volume of training data compensates for the poorer average label quality. If the number of silver images is reduced to 15 835, the score drops to 73.5%.

The results of the gold label finetuning are shown in the right columns of table 1. We see that the smaller the size of the gold label data set, the more the model benefits from additional silver labeled data compared to public data,

ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1. Available from http://www.i6doc.com/en/.

since the classification layer can already be trained on the final label set during pretraining. A slight improvement in performance can be detected by combining the automatic generated labels with transfer learning.

The over-all best model score is achieved by combining the full gold label data set with pretraining on CheXpert. A 77.9% test set AUROC is still relatively low compared to models predicting the CheXpert dataset (e.g. [4] with 89%) but fits the numbers reported by [6], who achieve a AUROC score of 76.8% on a similar set of labels (without 'CVC Location' labels). This seems to adhere to clinical intuition that ICU chest X-rays are harder to classify than non-emergency X-rays taken in controlled conditions.

If all gold samples are available, the highest score is achieved by transfer learning. Table 2 shows the AUROC depending on the respective findings for the best models, evaluated on the test set labeled by a radiology resident. There are significant performance differences between the individual targets. For high label quality targets, such as 'Pleural Effusion', no perfomance improvement can be achieved by transfer learning or finetuning on gold labels compared to the silver training. Targets with erroneous silver labels can be improved by using gold labels. The use of silver labels may be superior to the transfer learning approach, shown by the targets 'Pulmonary Infiltrates' and 'Pulmonary Congestions'. However, when the automatically generated labels reach a high error level, the classification is worsened by the silver pretraining. A negative learning effect takes place, which cannot be compensated by the gold training implemented, this behavior is recognizable for the target 'Misplaced CVC location'. In this case, pure transfer learning can achieve the best classification.

N	M_{G}	$M_{\mathrm{C+G}}$	$M_{\mathrm{S+G}}$	$M_{\mathrm{C+S+G}}$
500	63.5	65.3	74.6	**75.4**
2500	71.1	73.2	75.3	**75.8**
5000	72.2	75.4	75.6	**75.8**
7500	73.3	**76.3**	75.8	76.1
15835	75.7	**77.9**	76.2	76.3

Table 1: Comparison of classification performance on the hold-out test set labeled by research assistants depending on training data strategies and number of manually annotated training data N. For each model we evaluate the AUROC-Score (in %) over all 6 classes, weighted by positive ratio (see Table 2).

6 Conclusion

In this paper, the influences of automatically generated labels and transfer learning and their potential to improve automatic ICU chest X-ray analyses were investigated. If only a small amount of manually annotated gold labels is available, automatically generated labels improve significantly the performance. This effect is enhanced in combination with transfer learning on public data. However,

ESANN 2022 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 5-7 October 2022, i6doc.com publ., ISBN 978287587084-1. Available from http://www.i6doc.com/en/.

Finding	$M_{\mathrm{C+S}}$	$M_{\mathrm{C+G}}$	$M_{\mathrm{C+S+G}}$	$\mathrm{F1_S}$	%pos
Pulmonary Infiltrates	74.8	74.0	77.5	69.7	41.6
Pleural Effusion	85.1	84.0	85.1	94.3	60.0
Pneumothorax	75.9	69.3	73.3	95.0	5.0
Pulmonary Congestion	75.9	75.8	77.0	87.8	36.0
Misplaced CVC location	56.4	74.5	64.7	36.9	23.4
Regular CVC location	66.9	75.8	68.0	67.0	20.0

Table 2: Comparison of classification performance of best-performing models on the hold-out test set. For reference, we provide the F1-score of the text classification model (silver labels) on the gold label test set and the ratio of positive examples in the entire dataset.

for larger gold data sets, the improvement of the silver training by finetuning on gold labels depends on the quality of the silver labels. If automatically generated labels are too erroneous, transfer learning is a useful alternative. Future work will also use these findigs to improve an ICU chest X-ray classifier beyond the scope of this study, applying better automatic labeling systems (e.g. deep learning based text classification models) and further training techniques (e.g. oversampling of underrepresented classes) which were not considered in this study to isolate the effects of label quality and data domains.

References

[1] Jeremy Irvin et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597, 2019.

[2] Alistair Johnson et al. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs, 2019.

[3] Akshay Smit et al. Chexbert: Combining automatic labelers and expert annotations for accurate radiology report labeling using bert. *arXiv preprint arXiv:2004.09167*, 2020.

[4] Hieu H Pham et al. Interpreting chest x-rays via cnns that exploit hierarchical disease dependencies and uncertainty labels. *Neurocomputing*, 437:186–194, 2021.

[5] Saahil Jain et al. Effect of radiology report labeler quality on deep learning models for chest x-ray interpretation. *arXiv preprint arXiv:2104.00793*, 2021.

[6] K Tanaka et al. Superiority of supervised machine learning on reading chest x-rays in intensive care units. *Front Med (Lausanne)*, 2021.

[7] Daniel Gourdeau et al. Deep learning of chest x-rays can predict mechanical ventilation outcome in icu-admitted covid-19 patients. *Scientific Reports*, 12, 04 2022.

[8] Matthew BA McDermott et al. Chexpert++: Approximating the chexpert labeler for speed, differentiability, and probabilistic output. In *Machine Learning for Healthcare Conference*, pages 913–927. PMLR, 2020.

[9] Zhuoning Yuan et al. Large-scale robust deep auc maximization: A new surrogate loss and empirical studies on medical image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3040–3049, 2021.

[10] Gao Huang et al. Densely connected convolutional networks. *CoRR*, abs/1608.06993, 2016.