Language model-based labeling of German thoracic radiology reports

Sprachmodellbasiertes Labeling Deutscher Röntgenthoraxbefunde

Authors

Alessandro Wollek[1, 2], Philip Haitzer[1, 2], Thomas Sedlmeyr[1, 2], Sardi Hyska[3], Johannes Rueckel[3, 4], Bastian O. Sabel[3], Michael Ingrisch[3], Tobias Lasser[1, 2]

Affiliations

1 Munich Institute of Biomedical Engineering, Technical University of Munich, Garching near Munich, Germany
2 School of Computation, Information and Technology, Technical University of Munich, Garching near Munich, Germany
3 Department of Radiology, Ludwig-Maximilians-University Hospital Munich, Germany, Munich, Germany
4 Institute of Neuroradiology, Ludwig-Maximilians-University Hospital Munich, Munich, Germany

Keywords

annotation, deep learning, chest X-ray, chest radiograph, CheXpert, label extraction

received 19.7.2023
accepted after revision 9.3.2024
published online 2024

Bibliography

Fortschr Röntgenstr
DOI 10.1055/a-2287-5054
ISSN 1438-9029

Georg Thieme Verlag KG, Rüdigerstraße 14, 70469 Stuttgart, Germany

Correspondence

Alessandro Wollek
Munich Institute of Biomedical Engineering, Technical University of Munich, Boltzmannstr. 11, 85748 Garching near Munich, Germany
alessandro.wollek@tum.de

ABSTRACT

Purpose The aim of this study was to explore the potential of weak supervision in a deep learning-based label prediction model. The goal was to use this model to extract labels from German free-text thoracic radiology reports on chest X-ray images and for training chest X-ray classification models.

Materials and Methods The proposed label extraction model for German thoracic radiology reports uses a German BERT encoder as a backbone and classifies a report based on the CheXpert labels. For investigating the efficient use of manually annotated data, the model was trained using manual annotations, weak rule-based labels, and both. Rule-based labels were extracted from 66071 retrospectively collected radiology reports from 2017–2021 (DS 0), and 1091 reports from 2020–2021 (DS 1) were manually labeled according to the CheXpert classes. Label extraction performance was evaluated with respect to mention extraction, negation detection, and uncertainty detection by measuring F1 scores. The influence of the label extraction method on chest X-ray classification was evaluated on a pneumothorax data set (DS 2) containing 6434 chest radiographs with associated reports and expert diagnoses of pneumothorax. For this, DenseNet-121 models trained on manual annotations, rule-based and deep learning-based label predictions, and publicly available data were compared.

Results The proposed deep learning-based labeler (DL) performed on average considerably stronger than the rule-based labeler (RB) for all three tasks on DS 1 with F1 scores of 0.938 vs. 0.844 for mention extraction, 0.891 vs. 0.821 for negation detection, and 0.624 vs. 0.518 for uncertainty detection. Pretraining on DS 0 and fine-tuning on DS 1 performed better than only training on either DS 0 or DS 1. Chest X-ray pneumothorax classification results (DS 2) were highest when trained with DL labels with an area under the receiver operating curve (AUC) of 0.939 compared to RB labels with an AUC of 0.858. Training with manual labels performed slightly worse than training with DL labels with an AUC of 0.934. In contrast, training with a public data set resulted in an AUC of 0.720.

Conclusion Our results show that leveraging a rule-based report labeler for weak supervision leads to improved labeling performance. The pneumothorax classification results demonstrate that our proposed deep learning-based labeler can serve as a substitute for manual labeling requiring only 1000 manually annotated reports for training.

Key Points

- The proposed deep learning-based label extraction model for German thoracic radiology reports performs better than the rule-based model.
- Training with limited supervision outperformed training with a small manually labeled data set.
- Using predicted labels for pneumothorax classification from chest radiographs performed equally to using manual annotations.

Citation Format

- Wollek A, Haitzer P, Sedlmeyr T et al. Language model-based labeling of German thoracic radiology reports. Fortschr Röntgenstr 2024; DOI 10.1055/a-2287-5054

ZUSAMMENFASSUNG

Ziel Das Ziel dieser Studie war es, das Potenzial der schwachen Supervision in einem auf Deep Learning basierenden Modell zur Extraktion von Labels zu untersuchen. Die Motivation bestand darin, dieses Modell zu verwenden, um Labels aus deutschen Freitext-Thorax-Radiologie-Befunden zu extrahieren und damit Röntgenthorax-Klassifikationsmodelle zu trainieren.

Material und Methoden Das vorgeschlagene Modell zur Label-Extraktion für deutsche Thorax-Radiologie-Befunde verwendet einen deutschen BERT-Encoder als Grundlage und klassifiziert einen Befund basierend auf den CheXpert-Labels. Um den effizienten Einsatz von manuell annotierten Daten zu untersuchen, wurde das Modell mit manuellen Annotationen, regelbasierten Labels und beidem trainiert. Regelbasierte Labels wurden aus 66.071 retrospektiv gesammelten Radiologie-Befunden von 2017 bis 2021 (DS 0) extrahiert, und 1091 Befunde von 2020 bis 2021 (DS 1) wurden gemäß den CheXpert-Klassen manuell annotiert. Die Leistung der Label-Extraktion wurde anhand der Erfassung von Erwähnungen, der Erkennung von Negationen und der Erkennung von Unsicherheiten anhand von F1-Scores bewertet. Der Einfluss der Label-Extraktionsmethode auf die Röntgenthorax-Klassifikation wurde anhand eines Pneumothorax-Datensatzes (DS 2) mit 6434 Thoraxaufnahmen und entsprechenden Befunden evaluiert. Hierbei wurden DenseNet-121-Modelle, die mit manuellen Annotationen, regelbasierten und durch Deep Learning-basierten Label-Vorhersagen sowie öffentlich verfügbaren Daten trainiert wurden, verglichen.

Ergebnisse Der vorgeschlagene auf Deep Learning basierende Labeler (DL) zeigte im Durchschnitt für alle drei Aufgaben auf DS 1 eine bedeutend bessere Leistung als der regelbasierte Labeler (RB) mit F1-Scores von 0,938 gegenüber 0,844 für die Erwähnungserkennung, 0,891 gegenüber 0,821 für die Negationserkennung und 0,624 gegenüber 0,518 für die Unsicherheitserkennung. Das Vortraining auf DS 0 und das Feintuning auf DS 1 lieferte bessere Ergebnisse als nur das Training auf entweder DS 0 oder DS 1. Die Klassifikationsergebnisse für Pneumothorax auf Röntgenthoraces (DS 2) waren am besten, wenn sie mit DL-Labels trainiert wurden, mit einer Fläche unter der ROC-Kurve (AUC) von 0,939, im Vergleich zu RB-Labels mit einer AUC von 0,858. Das Training mit manuellen Labels war etwas schlechter als das Training mit DL-Labels mit einer AUC von 0,934. Das Training mit einem öffentlichen Datensatz führte zu einer AUC von 0,720.

Schlussfolgerung Unsere Ergebnisse zeigen, dass die Nutzung eines regelbasierten Labelers für schwache Supervision zu einer verbesserten Labeling-Leistung führt. Die Klassifikationsergebnisse für Pneumothorax zeigen, dass unser vorgeschlagener auf Deep Learning basierender Labeler ein möglicher Ersatz für manuelles Labeling ist und nur 1000 manuell annotierte Befunde für das Training benötigt.

Kernaussagen

- Das vorgeschlagene, Deep Learning basierende Modell zur Label-Extraktion für deutsche Thorax-Radiologie-Befunde schneidet besser ab als das regelbasierte Modell.
- Das Training mit limitierter Supervision schnitt besser ab, als das Training mit einem kleinen manuell annotierten Datensatz.
- Die Verwendung vorhergesagter Annotationen für die Pneumothorax-Klassifikation auf Röntgenthoraces schnitt gleich gut ab gegenüber der manuellen Annotation.

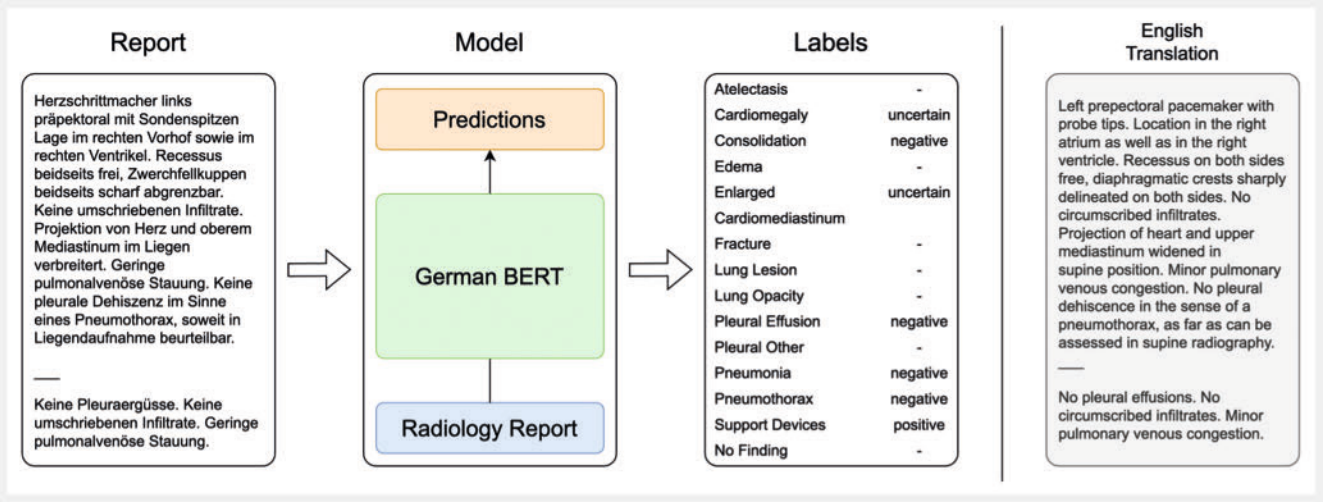
Introduction

Radiologists are in short supply worldwide [1, 2, 3, 4], for example, due to an aging population [5], and deep learning models hold promise for addressing this shortage, for example, as part of clinical decision-support systems [6, 7]. However, training such models often requires large data sets [8, 9] that are expensive and time-consuming to manually label [10, 11]. To reduce the amount of time for obtaining labeled data sets, automatic label extraction from radiology reports is a compelling option. Unfortunately, label extraction from radiology reports itself is a challenging task, for example, due to missing annotated data [12].

Recent developments in the natural language processing (NLP) domain have proposed models that generate dense word vector representations [13, 14, 15, 16], which have been shown to be effective in training deep learning models for a wide range of tasks such as translation [17] or named entity recognition [18]. Similar to the computer vision domain, these language models can be pretrained on a general, large corpus and then fine-tuned on a target corpus that might be otherwise too small for training [19].

In the medical domain, language models have been successfully applied to extract labels from unstructured radiology reports. Smit et al. improved upon their rule-based labeler for English radiology reports by using a BERT [15] language model as a backbone [20]. Similarly, Nowak et al. investigated the use of BERT for German radiology reports [11]. They compared a rule-based labeler to a deep learning model, trained with 18000 manually annotated reports, rule-based extracted labels, and a combination of both.

In this study, we explore the potential of weak supervision of a deep learning-based label prediction model, using a rule-based labeler. The general label extraction pipeline is illustrated in ▶ **Fig. 1**. Our proposed label extraction model takes a German free-text thoracic radiology report and extracts the corresponding labels. In contrast to Nowak et al., we focus on the classes of the CheXpert data set [21], allowing for comparison with previous studies and pooling of data sets for future studies. More importantly, we study the effect of extracted labels on downstream image classification training. Our study builds upon previous work that used rule-based strategies to extract labels [26]. We



▶ **Fig. 1** Automatic label prediction from German thoracic radiology reports. A report is processed by the BERT-based labeler and converted to 14 labels, motivated by the categories in the CheXpert data set. A class is labeled as positive, negative, or uncertain. If the class was not mentioned, it is classified as blank (–).

▶ **Table 1** Data sets used in this study. Data set 0 (DS 0) was labeled with a rule-based labeler [26], data set 1 (DS 1) was manually annotated solely based on radiological reports, and data set 2 (DS 2) was labeled based on the chest radiographs (CXR) and radiological reports.

Split	Data Set 0 (DS 0)	Data Set 1 (DS 1)	Data Set 2 (DS 2)
Training	60071	810	4507
Validation	1000	203	660
Test	5000	78	1267
Total	66071	1091	6434
Annotations	Automatic	Report, manual	CXR + report, manual

conduct extensive experiments on a data set of internal radiology reports and our results demonstrate the effectiveness of our approach.

Our contributions are: (1) We propose a deep learning-based label extraction model for German thoracic radiology reports based on the classes of the CheXpert data set. (2) We demonstrate that our labeler outperforms a rule-based label extraction model with respect to label extraction and utility for downstream applications. (3) We show that a pneumothorax classifier trained with automatically extracted labels performs equivalently to a model trained on manual annotations.

Our code is publicly available at https://gitlab.lrz.de/IP/german-lm-radiology-report-labeler.

Materials and Methods

Data Collection

Data splits and annotation methods of all data sets used throughout this study are reported in ▶ **Table 1**. We retrospectively identified 66071 thoracic radiology reports from 2017 to 2021 in our institutional PACS (DS 0). As the purpose of this study is to investigate automatic labeling of clinical reports, the collected reports represent an unfiltered sample in terms of sex and age. Due to this, information on sex and age were not extracted. Additionally, we used 1091 thoracic radiology reports from 2020–2021 that were manually annotated by a first-year radiology resident from LMU Klinikum in a previous study [26]. In the following, we refer to the manually annotated reports as data set 1 (DS 1).

The training and test set label distributions of DS 1 are reported in ▶ **Table 2**. Since annotated "no finding" reports describe normal appearing chest radiographs, there are no negative or uncertain annotations available for this class.

To increase the number of training samples, we favored test samples with multiple non-blank annotations. We selected 78 of the 1091 reports of DS 1 for testing. Our selection process ensured that each class was mentioned by at least five reports, whenever available. In cases where the entire data set contained less than five samples for a specific class, half of the samples were designated for testing. None of the 78 DS 1 reports used for testing were part of DS 0.

To further test our model, we utilized another internal data set consisting of 6434 chest radiographs with corresponding reports [26]. We refer to this data set in the following as data set 2 (DS 2).

▶ **Table 2** Label distributions of manually annotated data sets used in this study. Data set 1 class annotations were labeled based on free text reports [26]. Data set 2 class annotations were based on reports and radiographs [26]. Enlarged Cardiom. = Enlarged Cardiomediastinum, P = Positive, U = Uncertain, N = Negative.

Data Set	Data Set 1 (DS 1)						Data Set 2 (DS 2)					
Split	Development			Test			Training		Validation		Test	
Class	P	U	N	P	U	N	P	N	P	N	P	N
Atelectasis	220	54	2	12	13	1	–	–	–	–	–	–
Cardiomegaly	184	368	266	16	25	25	–	–	–	–	–	–
Consolidation	205	45	627	23	6	41	–	–	–	–	–	–
Edema	297	9	521	24	5	34	–	–	–	–	–	–
Enlarged Cardiom.	223	295	305	22	19	26	–	–	–	–	–	–
Fracture	63	3	79	9	2	8	–	–	–	–	–	–
Lung Lesion	44	7	8	5	5	5	–	–	–	–	–	–
Lung Opacity	278	41	565	28	6	35	–	–	–	–	–	–
No Finding	1	0	0	1	0	0	–	–	–	–	–	–
Pleural Effusion	455	45	451	29	11	32	–	–	–	–	–	–
Pleural Other	57	16	1	7	5	0	–	–	–	–	–	–
Pneumonia	52	173	649	5	16	45	–	–	–	–	–	–
Pneumothorax	83	7	871	5	5	66	1122	3385	204	456	326	941
Support Devices	590	1	107	43	1	12	–	–	–	–	–	–

This data set, in contrast to DS 1, contains only binary pneumothorax annotations. However, the annotations are based on both report and chest radiograph providing a higher label quality. In the data set, 1568 samples have been labeled as pneumothorax.

Architecture

Based on Smit et al. [20], we used a pre-trained BERT [15] model as the backbone for our label extraction model. The objective of the model is to predict the fourteen CheXpert [21] labels: atelectasis, cardiomegaly, consolidation, edema, enlarged cardiomediastinum, fracture, lung lesion, lung opacity, pleural effusion, pleural other, pneumonia, pneumothorax, support devices, and "no finding" given a German radiology report.

The architecture is illustrated in ▶ **Fig. 2**. The model receives the report as an input and assigns one of the classes: blank, positive, negative, or uncertain to each of the 13 categories, mirroring a manual annotation. The blank classification represents no mention of the class in the report. For the special case "no finding", which corresponds to a normal report, the labeler must predict only blank or positive.

We modified the BERT architecture by using 14 linear heads, as illustrated in ▶ **Fig. 2**. Each head is dedicated to capture one of the 14 labels. For transfer learning, we use the pre-trained "bert-base-German-cased" BERT model[1] trained on German texts, such as the German Wikipedia corpus, with a sequence length of 512 tokens.

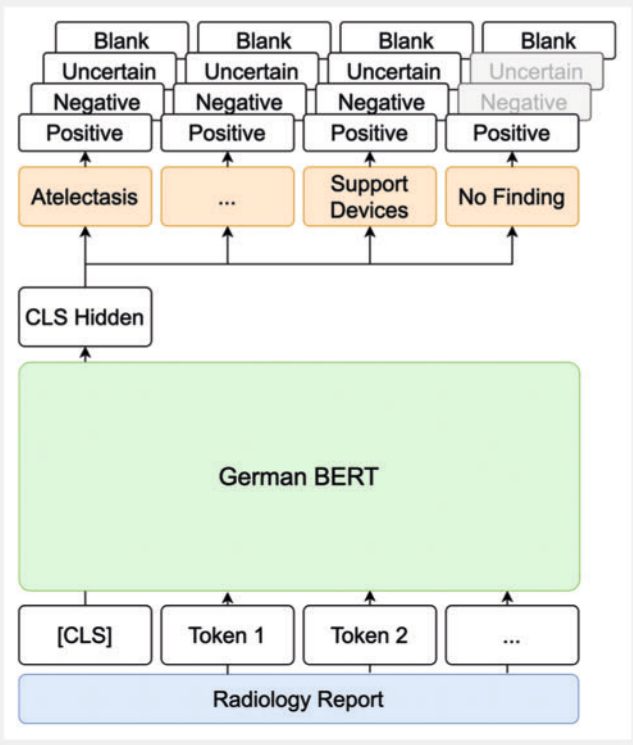
To predict the classes of the 14 findings, the radiology reports were tokenized first. Of all tokenized reports, a single report in the training data, and none in the test data consisted of more than 512 tokens. The overflowing report consisted of 579 tokens and described multiple images. We considered only the first 512 tokens of this report. After tokenization, the reports were processed by the model. Subsequently, the hidden state of the class (CLS) token from the final layer was used as the input for each of the 14 linear heads, predicting the class of each finding via a softmax.

The model was fine-tuned on a NVIDIA GeForce GTX 1080 for 3 epochs using cross-entropy loss, AdamW [22] optimization with default parameters ($\beta_1 = 0.9$, $\beta_2 = 0.999$), a learning rate of 2e-5, and a batch size of 8. The individual cross-entropy losses for the 14 observations were aggregated before calculating the final loss. To monitor model performance, we periodically evaluated the model on the validation set and selected the best checkpoint according to the validation cross-entropy loss across all 14 observations.

Label Extraction (DS 1)

We evaluated our deep learning-based labeler on the three tasks proposed by the original CheXpert data set: mention extraction, negation detection, and uncertainty detection. Following the original CheXpert experimental setup, findings labeled as "blank" were considered as negative for the mention extraction task and the other classes ("positive", "negative", or "uncertain") as positive. Regarding negation detection, only the "negative" classifica-

1 https://huggingface.co/bert-base-german-cased.



▶ **Fig. 2** Deep learning-based German radiology report labeler. The model extracts CheXpert labels from free-text radiology reports.

tion was considered positive, and for uncertainty detection, only the "uncertain" class was considered positive.

To assess the importance of manually and automatically extracted annotations, we designed three experiments: training only with manually annotated reports (supervised), DS 1, automatically extracted labels (weakly supervised), DS 0, and all available data (hybrid), DS 0 + DS 1. As a baseline, we trained the model solely on manually annotated reports (DS 1) (supervised approach).

We investigated the benefit of automatically created weak labels on label extraction performance. The labels were created using the rule-based model proposed in a previous study [26] based on the CheXpert labeler. For validation, we randomly sampled 1000 reports, and for internal testing 5000 reports, without patient overlap, from the total 66071 reports of DS 0 (▶ **Table 1**). We used the remaining 60071 reports for training. For final testing, we used the manually labeled test reports of DS 1.

To leverage all available data, we fine-tuned the weakly supervised model on the manually annotated reports (DS 1) (hybrid approach). Again, we trained the model on increasing fractions of DS 1, as reported in ▶ **Table 3**.

Pneumothorax Classification (DS 2)

To address the limitation of the small test data set of DS 1, we tested the labeler on the larger data set 2. Since the data set contains only binary pneumothorax annotations, we considered uncertain predictions as positive, and blank predictions as negative.

Furthermore, as the goal of label extraction is the training of image classification models, we trained a chest X-ray classifier to predict the presence of a pneumothorax based on manual and extracted labels.

Our pneumothorax classification pipeline utilized a DenseNet-121 [23] pre-trained on ImageNet as a backbone. We replaced the final fully connected layer with a one-dimensional version for fine-tuning on DS 2. The final softmax activation was replaced by a sigmoid. Training involved 10 epochs with AdamW with default parameters ($\beta_1 = 0.9$, $\beta_2 = 0.999$), a learning rate of 0.003, and batch size of 32. We selected the checkpoint for the final model based on the validation area under the receiver operating characteristic curve (AUC). All images were resized to 224 × 224 pixels and normalized using the ImageNet mean and standard deviation. Data augmentation involved ten-crop, i. e., taking five crops of the regular and flipped image. The complete pipeline for the deep learning-based experimental setup is illustrated in ▶ **Fig. 3**.

We assessed the effect of the labeling method on pneumothorax classification performance on DS 2 by comparing fine-tuning using radiologists' annotations [26], rule-based [26] or deep learning-based extracted labels, with a DenseNet-121 fine-tuned on the chest X-ray 14 data set [24] (CheXnet [25]).

Statistical Evaluation

For all three experimental settings on DS 1, we measured mean F1 scores for the three tasks of mention extraction, negation detection, and uncertainty detection by comparing model predictions with manually annotated test reports.

Label extraction performance on DS 2 was measured using sensitivity and specificity. To simplify the comparison with DS 1, we applied the same metrics. We measured pneumothorax classification performance by analyzing receiver operating characteristics (ROC) and AUC. As our research involves numerous comparisons and is purely explorative, we abstained from reporting P-values and instead presented 95 % confidence intervals, which were calculated using 10000-fold resampling via non-parametric bootstrap methodology at the level of the image or report. Due to space limitations, 95 % confidence intervals for the F1 scores were not included.

All statistical analyses were performed using Python version 3.8.10, NumPy version 1.24.2, and Scikit-Learn version 1.2.2.

Due to the retrospective nature of the study, written informed consent was waived.

Results

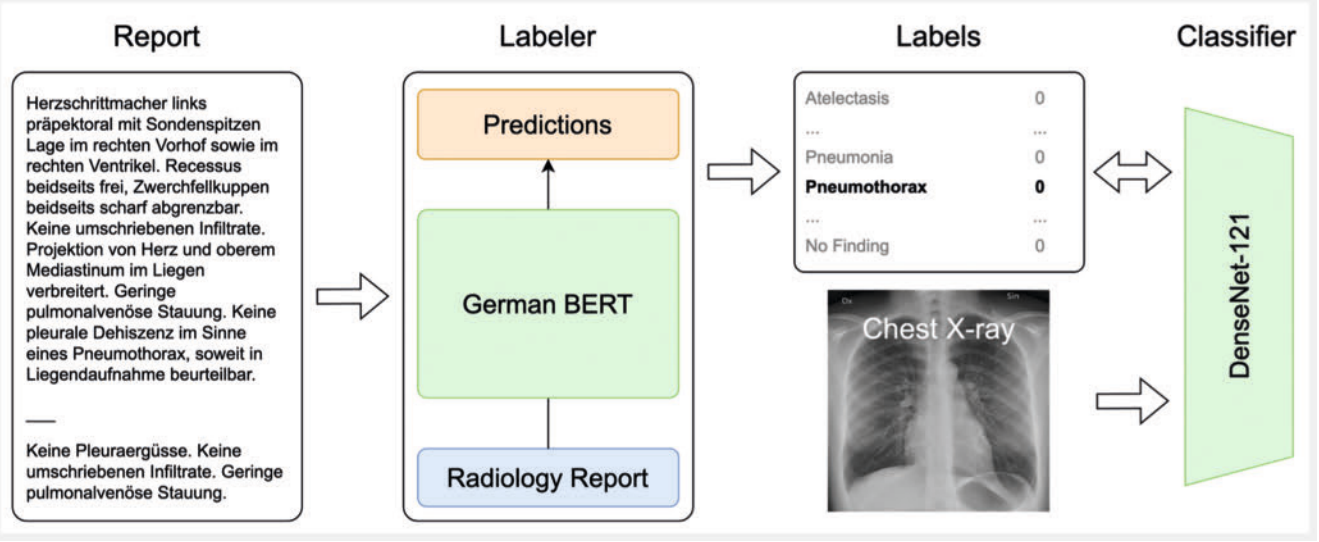
Label Extraction (DS 1)

The label extraction results are reported in ▶ **Table 3** and ▶ **Table 4**. Results marked as N/A did not have enough samples to calculate the corresponding F1 score. Overall, pre-training using weak labels followed by training on manually annotated data performed best across all tasks.

The results obtained when trained solely on DS 1 (supervised approach) are reported in ▶ **Table 4**. The model achieved a mean mention extraction F1 score of 0.846 [CI: 0.797–0.880], a nega-

▶ **Table 3** Comparison of mean test-F1 scores for mention extraction, negation detection, and uncertainty detection on data set 1 with corresponding 95 % confidence intervals. Weakly supervised and hybrid models were (pre-) trained on data set 0. Hybrid and supervised models were trained on data set 1.

Data Set 1	Mention Extraction	Negation Detection	Uncertainty Detection
Supervised	0.846 [0.797–0.880]	0.829 [0.754–0.880]	0.575 [0.444–0.677]
Weakly Supervised	0.905 [0.876–0.927]	0.818 [0.742–0.857]	0.534 [0.408–0.611]
Hybrid	**0.938** [0.908–0.959]	**0.891** [0.838–0.924]	**0.624** [0.492–0.716]



▶ **Fig. 3** Pneumothorax classification model trained with automatically extracted annotations.

tion detection F1 score of 0.829 [CI: 0.754–0.880], and a mean uncertainty detection F1 score of 0.575 [CI: 0.444–0.677].

When trained only with reports labeled by the rule-based labeler (weakly supervised approach), the model achieved a mean F1 score of 0.905 [CI: 0.876–0.927] for mention extraction, 0.818 [CI: 0.742–0.857] for negation detection, and 0.534 [CI: 0.408–0.611] for uncertainty detection (▶ **Table 3**). Note that although the model was trained on DS 0, the reported test results were measured on DS 1.

The effect of pre-training with automatically labeled reports first and then fine-tuning on varying amounts of manually annotated data (hybrid approach) is reported in ▶ **Table 3**. The model achieved a mean mention extraction F1 score of 0.938 [CI: 0.908–0.959], negation detection F1 score of 0.891 [CI: 0.838–0.924], and uncertainty detection F1 score of 0.624 [CI: 0.492–0.716].

Rule-based and deep learning-based label extraction results for all three evaluation tasks are compared in ▶ **Table 4**. The deep learning-based labeler was pre-trained with labels extracted by the rule-based labeler (DS 0) and fine-tuned on the manually annotated training data (DS 1) (hybrid approach). Across all three tasks, the deep learning model performed substantially better. For mention extraction, our proposed labeler had a mean F1 score of 0.938 [CI: 0.908–0.959] compared to the score of 0.844 [CI: 0.823–0.922] for the rule-based labeler. For negation and uncertainty detection, the improvement of using a deep learning-based labeler compared to a rule-based model was 0.891 [CI: 0.838–0.924] vs. 0.821 [CI: 0.747–0.858] mean F1 score for negation detection, and 0.624 [CI: 0.492–0.716] vs. 0.518 [CI: 0.395–0.594] mean F1 score for uncertainty detection.

To simplify comparison of labeling results on DS 1 with the labeling results on DS 2, we additionally measured sensitivity and specificity by considering uncertain labels as positive and blank labels as negative. The results are reported in ▶ **Table 5**. On average, the deep learning-based labeler achieved a sensitivity of 0.787 [CI: 0.746–0.886] compared to the rule-based approach with 0.782 [CI: 0.741–0.878] and a higher specificity with 0.934 [CI: 0.905–0.956] vs. 0.904 [CI: 0.875–0.929].

Pneumothorax Label Extraction (DS 2)

The comparison of the rule-based and the deep learning-based labeler for pneumothorax annotation on DS 2 is presented in ▶ **Table 5**. The rule-based labeler had a higher sensitivity compared to the deep learning-based model with 0.997 [CI: 0.994–0.999] vs. 0.972 [CI: 0.963–0.979]. In contrast, the deep learning-based labeler had a higher specificity with 0.995 [CI: 0.993–0.997] vs. 0.991 [CI: 0.988–0.994].

▶ **Table 4** Rule-based (RB) and deep learning-based (DL) label extraction F1 scores for the three evaluation tasks: mention extraction, negation detection, and uncertainty detection for each finding with corresponding 95 % confidence intervals. Labels were extracted from DS 1 and compared to manual annotations. N/A results could not be calculated due to insufficient data. Higher values are highlighted in bold.

Data Set 1	Mention Extraction		Negation Detection		Uncertainty Detection	
Findings	**RB**	**DL**	**RB**	**DL**	**RB**	**DL**
Atelectasis	**0.982** [0.936–1.000]	0.963 [0.900–1.000]	**1.000** [1.000–1.000]	N/A	**0.769** [0.545–0.923]	0.700 [0.421–0.889]
Cardiomegaly	0.660 [0.547–0.757]	**0.955** [0.913–0.986]	0.649 [0.444–0.810]	**0.898** [0.791–0.978]	0.571 [0.341–0.744]	**0.809** [0.667–0.917]
Consolidation	0.950 [0.909–0.980]	**0.979** [0.952–1.000]	0.738 [0.600–0.846]	**0.909** [0.833–0.968]	0.400 [0.111–0.643]	0.400 [0.000–0.750]
Edema	**0.992** [0.975–1.000]	0.984 [0.959–1.000]	0.939 [0.871–0.987]	**0.970** [0.921–1.000]	**0.600** [0.000–0.909]	N/A
Enlarged Cardiomediastinum	0.817 [0.732–0.885]	**0.932** [0.883–0.971]	**0.800** [0.650–0.913]	0.776 [0.622–0.889]	0.500 [0.261–0.696]	**0.821** [0.667–0.933]
Fracture	0.900 [0.784–0.979]	0.900 [0.778–0.980]	0.545 [0.000–0.857]	**0.857** [0.571–1.000]	N/A	N/A
Lung Lesion	0.857 [0.710–0.968]	**0.938** [0.828–1.000]	0.889 [0.500–1.000]	0.889 [0.500–1.000]	0.182 [0.000–0.500]	**0.714** [0.364–0.933]
Lung Opacity	0.935 [0.889–0.973]	**0.951** [0.912–0.980]	0.679 [0.512–0.812]	**0.848** [0.742–0.932]	0.316 [0.000–0.571]	N/A
No Finding	**0.025** [0.025–0.096]	N/A	–	–	–	–
Pleural Effusion	**0.980** [0.952–1.000]	0.973 [0.944–0.994]	0.954 [0.893–1.000]	**0.970** [0.919–1.000]	0.588 [0.250–0.833]	**0.750** [0.429–0.947]
Pleural Other	**0.842** [0.600–1.000]	0.706 [0.364–0.923]	N/A	N/A	**0.500** [0.000–0.857]	0.333 [0.000–0.800]
Pneumonia	0.921 [0.867–0.965]	**0.964** [0.927–0.993]	0.889 [0.806–0.953]	**0.966** [0.920–1.000]	0.600 [0.353–0.786]	**0.688** [0.467–0.857]
Pneumothorax	0.987 [0.966–1.000]	**0.994** [0.980–1.000]	**0.964** [0.927–0.993]	0.957 [0.919–0.986]	**0.667** [0.000–1.000]	0.400 [0.000–1.000]
Support Devices	**0.971** [0.933–1.000]	0.962 [0.920–0.991]	**0.800** [0.545–0.960]	0.762 [0.500–0.938]	N/A	N/A
Mean	0.844 [0.823–0.922]	**0.938** [0.908–0.959]	0.821 [0.747–0.858]	**0.891** [0.838–0.924]	0.518 [0.395–0.594]	**0.624** [0.492–0.716]

Pneumothorax Classification (DS 2)

To assess the performance of our proposed label extraction algorithm, we trained a pneumothorax classifier on chest radiographs with labels generated by different methods. The classification ROC curves and AUC values with corresponding 95 % confidence intervals are shown in ▶ **Fig. 4**.

The baseline CheXnet model trained on the chest X-ray 14 data set achieved the lowest performance with an AUC of 0.720 [CI: 0.687–0.882], followed by the model trained on DS 2 with labels extracted from the rule-based model with an AUC of 0.858 [CI: 0.832–0.882]. When trained on labels created by radiologists inspecting both image and report, the model achieved an AUC score of 0.934 [CI: 0.918–0.949]. The highest AUC values were obtained when trained with labels extracted by our proposed deep learning-based model with 0.939 AUC [CI: 0.925–0.952].

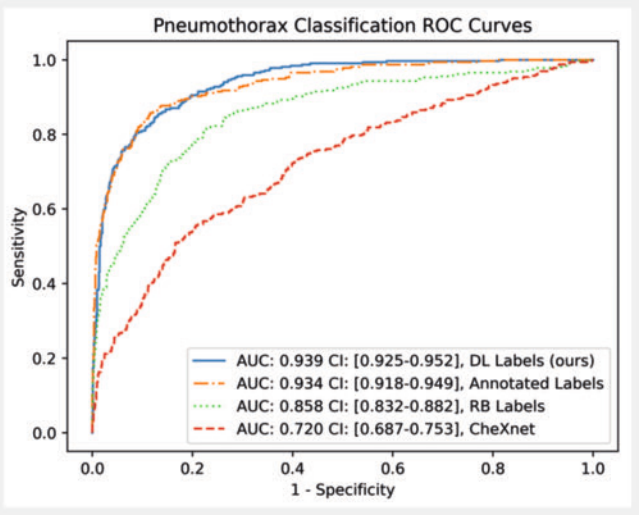
Discussion

In this study, we proposed a deep learning-based chest radiology report label extraction model. The best performing model was pre-trained on reports labeled by a rule-based labeler and fine-tuned on only a thousand manually labeled reports. On average, it outperformed the rule-based model in all three tasks (▶ **Table 4**). These results suggest that the improvements of employing deep learning-based compared to rule-based label extraction of CheXpert labels transfer from English to German radiology reports [21].

The pneumothorax chest X-ray classification results provide further evidence of the improvements of our proposed deep learning-based labeler compared to the rule-based labeler. Not only did the AUC increase from 0.858 to 0.939, but it also surpassed the model trained on the DS 2 labels that were annotated by radiologists based on inspecting the image and report. These

▶ **Table 5** Sensitivity and specificity of the extracted labels compared to the reference annotations on DS 1 and DS 2 with corresponding 95 % confidence intervals. To create binary labels, uncertain labels/annotations were considered positive, blank negative. The deep learning model was first pre-trained on weak labels and then fine-tuned on the manually annotated training data. RB = rule-based labeler, DL = deep learning-based labeler (ours). Higher values are highlighted in bold.

	Data Set 1				Data Set 2			
	Sensitivity		Specificity		Sensitivity		Specificity	
Findings	**RB**	**DL**	**RB**	**DL**	**RB**	**DL**	**RB**	**DL**
Atelectasis	**0.960** [0.867–1.000]	0.920 [0.800–1.000]	**1.000** [1.000–1.000]	0.981 [0.939–1.000]				
Cardiomegaly	0.561 [0.405–0.714]	**0.927** [0.838–1.000]	**0.892** [0.781–0.976]	0.838 [0.710–0.946]				
Consolidation	**0.966** [0.886–1.000]	0.897 [0.769–1.000]	0.735 [0.605–0.852]	**0.857** [0.750–0.945]				
Edema	**0.966** [0.886–1.000]	**0.966** [0.885–1.000]	0.918 [0.833–0.981]	**0.959** [0.896–1.000]				
Enlarged Cardio-mediastinum	0.659 [0.512–0.800]	**0.854** [0.737–0.953]	**0.784** [0.645–0.909]	0.730 [0.579–0.867]				
Fracture	**0.909** [0.700–1.000]	**0.909** [0.700–1.000]	0.940 [0.877–0.986]	**0.985** [0.952–1.000]				
Lung Lesion	**0.900** [0.667–1.000]	**0.900** [0.667–1.000]	0.956 [0.900–1.000]	**1.000** [1.000–1.000]				
Lung Opacity	**0.971** [0.903–1.000]	0.882 [0.765–0.974]	0.636 [0.489–0.773]	**0.818** [0.698–0.927]				
No Finding	0.000 [0.000–0.000]	0.000 [0.000–0.000]	0.922 [0.857–0.974]	**1.000** [1.000–1.000]				
Pleural Effusion	**0.925** [0.833–1.000]	**0.925** [0.833–1.000]	**0.974** [0.914–1.000]	**0.974** [0.914–1.000]				
Pleural Other	**0.750** [0.500–1.000]	0.583 [0.286–0.875]	**1.000** [1.000–1.000]	**1.000** [1.000–1.000]				
Pneumonia	0.857 [0.696–1.000]	**0.952** [0.842–1.000]	**0.982** [0.943–1.000]	**0.982** [0.943–1.000]				
Pneumothorax	**0.600** [0.273–0.900]	0.400 [0.100–0.727]	0.971 [0.925–1.000]	**0.985** [0.954–1.000]	**0.997** [0.994, 0.999]	0.972 [0.963–0.979]	0.991 [0.988, 0.994]	**0.995** [0.993–0.997]
Support Devices	**0.932** [0.850–1.000]	0.909 [0.818–0.979]	0.941 [0.850–1.000]	**0.971** [0.903–1.000]				
Mean	0.782 [0.741–0.878]	**0.787** [0.746–0.886]	0.904 [0.875–0.929]	**0.934** [0.905–0.956]				



▶ **Fig. 4** Pneumothorax classification receiver operating characteristic (ROC) curves and areas under the ROC curve (AUC) with corresponding 95 % confidence intervals. All models, except the CheXnet baseline, were trained on DS 2 with manual expert annotations (Annotated Labels), extracted with a rule-based (RB Labels), or our proposed deep learning-based labeler (DL Labels). The CheXnet model was trained on the chest X-ray 14 data set.

results suggest that training with labels extracted from free-text reports by the deep learning-based labeler is an alternative to time-consuming manual labeling.

The differences between pneumothorax sensitivity in DS 1 and DS 2 can be explained by the respective data collection and annotation process. DS 2 samples were specifically selected for clear cases or the absence of pneumothorax. Uncertain cases were removed in the data collection process. In contrast, the data for DS 1 was not filtered and uncertain cases were kept. Given similar specificity and strong chest X-ray classification results, we interpret the sensitivity differences due to the different data collection processes. The low pneumothorax uncertainty detection F1 score further supports this interpretation.

Similar to Nowak et al. [11], the deep learning-based model outperformed the rule-based model on German reports. Apart from using different data sets and labels, a direct comparison is not conclusive, as their model was trained differently, and they considered both uncertain and negative mentions as negative labels. Our rule-based labeler served as a strong baseline (▶ **Table 5**). Consequently, pre-training with such weak supervision improved the performance compared to only training on manually annotated data alone. For example, the mean mention extraction F1 score improved from 84 % to 94 % when using all data. Furthermore, our model was trained on approximately 1000 manually labeled reports, compared to the total of 14580 used for development by Nowak et al. [11]. They showed that increasing the amount of manually annotated training data improved mean F1 scores from 70.9 % to 95.5 % when increasing training data from 500 to 14580 samples. However, annotating all 14580 samples took 197 hours. Based on their results, we assume that increasing the number of manually annotated samples could further improve our model.

Our study has several limitations. First, due to the limited number of available manually annotated reports, most data were used for training. To compensate for this, we tested the model on a larger pneumothorax data set (DS 2). A future study with more manually annotated data could both improve model performance and reduce the variance of test scores. Another limitation is that the labels of DS 1 were created by a single radiologist, possibly introducing label biases or errors made due to annotation fatigue.

In conclusion, we demonstrated a considerable improvement in German radiology report labeling using our proposed deep learning-based labeler. Our results provide evidence of the benefits of employing a deep learning-based model, even in scenarios with sparse data, and the use of the rule-based labeler as a tool for weak supervision.

Clinical Relevance

One of the main motivations of employing deep learning models in clinical decision support systems is to reduce the effects of the worldwide shortage of radiologists. However, the data to train or test such models must be annotated by radiologists. Our presented labeler drastically reduces the required amount of manually annotated reports and performed equivalently compared to the pneumothorax classification model trained with labels created by radiologists.

Funding Information

Bundesministerium für Gesundheit (2520DAT920) | http://dx.doi.org/10.13039/501100003107

Conflict of Interest

The authors declare that they have no conflict of interest.

References

[1] Rosenkrantz AB, Hughes DR, Duszak R Jr. The US radiologist workforce: an analysis of temporal and geographic variation by using large national datasets. Radiology 2016; 279: 175–184. doi:10.1148/radiol.2015150921

[2] Rimmer A. Radiologist shortage leaves patient care at risk, warns royal college. BMJ: British Medical Journal (Online) 2017; 359. doi:10.1136/bmj.j4683

[3] Bastawrous S, Carney B. Improving patient safety: avoiding unread imaging exams in the national VA enterprise electronic health record. Journal of digital imaging 2017; 30: 309–313. doi:10.1007/s10278-016-9937-2

[4] Rosman DA, Nshizirungu JJ, Rudakemwa E et al. Imaging in the land of 1000 hills: Rwanda radiology country report. Journal of Global Radiology 2015; 1: 5. doi:10.7191/jgr.2015.1004

[5] Bundesärztekammer. Accessed December 07, 2023 at: https://www.bundesaerztekammer.de/baek/ueber-uns/aerztestatistik/2022

[6] Saba L, Biswas M, Kuppili V et al. The present and future of deep learning in radiology. European Journal of Radiology 2019; 114: 14–24. doi:10.1016/j.ejrad.2019.02.038

[7] Syed A, Zoga A. Artificial Intelligence in Radiology: Current Technology and Future Directions. Semin Musculoskelet Radiol 2018; 22: 540–545. doi:10.1055/s-0038-1673383

[8] Dosovitskiy A, Beyer L, Kolesnikov A et al. An Image is Worth 16 × 16 Words: Transformers for Image Recognition at Scale. In: International Conference on Learning Representations 2020

[9] Wollek A, Graf R, Čečatka S et al. Attention-based Saliency Maps Improve Interpretability of Pneumothorax Classification. Radiology: Artificial Intelligence 2023; 5: e220187. doi:10.1148/ryai.220187

[10] Filice RW, Stein A, Wu CC et al. Crowdsourcing pneumothorax annotations using machine learning annotations on the NIH chest X-ray dataset. Journal of digital imaging 2020; 33: 490–496. doi:10.1007/s10278-019-00299-9

[11] Nowak S, Biesner D, Layer YC et al. Transformer-based structuring of free-text radiology report databases. Eur Radiol 2023. doi:10.1007/s00330-023-09526-y

[12] Oakden-Rayner L. Exploring large scale public medical image datasets. arXiv preprint arXiv:190712720 2019. doi:10.48550/arXiv.1907.12720

[13] Mikolov T, Chen K, Corrado G et al. Efficient Estimation of Word Representations in Vector Space. arXiv:13013781 [cs] 2013. doi:10.48550/arXiv.1301.3781

[14] Bojanowski P, Grave E, Joulin A et al. Enriching Word Vectors with Subword Information. Transactions of the Association for Computational Linguistics 2017; 5: 135–146. doi:10.1162/tacl_a_00051

[15] Devlin J, Chang M-W, Lee K et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:181004805 [cs] 2018. doi:10.48550/arXiv.1810.04805

[16] Radford A, Narasimhan K, Salimans T et al. Improving language understanding by generative pre-training; OpenAI blog; 2018

[17] Vaswani A, Shazeer N, Parmar N et al. Attention is all you need. In: Advances in neural information processing systems 2017: 5998–6008

[18] Schweter S, Akbik A. Flert: Document-level features for named entity recognition. arXiv preprint arXiv:201106993 2020. doi:10.48550/arXiv.2011.06993

[19] Howard J, Ruder S. Universal Language Model Fine-tuning for Text Classification. arXiv preprint arXiv:180106146 2018. doi:10.48550/arXiv.1801.06146

[20] Smit A, Jain S, Rajpurkar P et al. CheXbert: combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. arXiv preprint arXiv:200409167 2020. doi:10.48550/arXiv.2004.09167

[21] Irvin J, Rajpurkar P, Ko M et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: Proceedings of the AAAI Conference on Artificial Intelligence 2019: 590–597

[22] Loshchilov I, Hutter F. Decoupled Weight Decay Regularization. In: International Conference on Learning Representations 2018

[23] Huang G, Liu Z, Van Der Maaten L et al. Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition 2017: 4700–4708

[24] Wang X, Peng Y, Lu L et al. ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2017: 3462–3471

[25] Rajpurkar P, Irvin J, Zhu K et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. arXiv preprint arXiv:171105225 2017. doi:10.48550/arXiv.1711.05225

[26] Wollek A, Hyska S, Sedlmeyr T et al. German CheXpert Chest X-ray Radiology Report Labeler. Fortschr Röntgenstr 2024 doi:10.1055/a-2234-8268