SCIENTIFIC REPORTS

OPEN

Comparison of Deep Learning Approaches for Multi-Label Chest X-Ray Classification

Received: 14 August 2018
Accepted: 25 March 2019
Published online: 23 April 2019

Ivo M. Baltruschat[1,2], Hannes Nickisch[3], Michael Grass[3], Tobias Knopp[1,2] & Axel Saalbach[3]

The increased availability of labeled X-ray image archives (e.g. ChestX-ray14 dataset) has triggered a growing interest in deep learning techniques. To provide better insight into the different approaches, and their applications to chest X-ray classification, we investigate a powerful network architecture in detail: the ResNet-50. Building on prior work in this domain, we consider transfer learning with and without fine-tuning as well as the training of a dedicated X-ray network from scratch. To leverage the high spatial resolution of X-ray data, we also include an extended ResNet-50 architecture, and a network integrating non-image data (patient age, gender and acquisition type) in the classification process. In a concluding experiment, we also investigate multiple ResNet depths (i.e. ResNet-38 and ResNet-101). In a systematic evaluation, using 5-fold re-sampling and a multi-label loss function, we compare the performance of the different approaches for pathology classification by ROC statistics and analyze differences between the classifiers using rank correlation. Overall, we observe a considerable spread in the achieved performance and conclude that the X-ray-specific ResNet-38, integrating non-image data yields the best overall results. Furthermore, class activation maps are used to understand the classification process, and a detailed analysis of the impact of non-image features is provided.

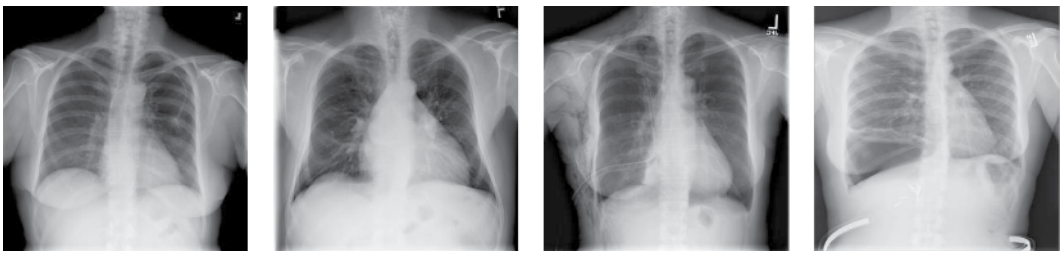
In the United Kingdom, the care quality commission recently reported that – over the preceding 12 months – a total of 23,000 chest X-rays (CXRs) were not formally reviewed by a radiologist or clinician at Queen Alexandra Hospital alone. Furthermore, three patients with lung cancer suffered significant harm because their CXRs had not been properly assessed[1]. The Queen Alexandra Hospital is probably not the only hospital having problems with providing expert readings for every CXR. Growing populations and increasing life expectancies are expected to drive an increase in demand for CXR readings.

In computer vision, deep learning has already shown its power for image classification with superhuman accuracy[2–5]. In addition, the medical image processing field is vividly exploring deep learning. However, one major problem in the medical domain is the availability of large datasets with reliable ground-truth annotation. Therefore, transfer learning approaches, as proposed by Bar *et al.*[6], were often considered to overcome such problems.

Two larger X-ray datasets have recently become available: The CXR dataset from Open-i[7] and the ChestX-ray14 dataset from the National Institutes of Health (NIH) Clinical Center[8]. Figure 1 illustrates four selected examples from ChestX-ray14. Due to its size, the ChestX-ray14 consisting of 112,120 frontal CXR images from 30,805 unique patients attracted considerable attention in the deep learning community. Triggered by the work of Wang *et al.*[8] using convolution neural networks (CNNs) from the computer vision domain, several research groups have begun to address the application of CNNs for CXR classification. In the work of Yao *et al.*[9], they presented a combination of a CNN and a recurrent neural network to exploit label dependencies. As a CNN backbone, they used a DenseNet[10] model which was adapted and trained entirely on X-ray data. Li *et al.*[11] presented a framework for pathology classification and localization using CNNs. More recently, Rajpurkar *et al.*[12] proposed transfer-learning with fine tuning, using a DenseNet-121[10], which raised the AUC results on ChestX-ray14 for multi-label classification even higher.

Unfortunately, a faithful comparison of approaches remains difficult. Most reported results were obtained with differing experimental setups. This includes (among others) the employed network architecture, loss

[1]Institute for Biomedical Imaging, Hamburg University of Technology, Hamburg, Germany. [2]Department of Biomedical Imaging, University Medical Center Hamburg-Eppendorf, Hamburg, Germany. [3]Philips Research, Hamburg, Germany. Correspondence and requests for materials should be addressed to I.M.B. (email: ivo-matteo.baltruschat@tuhh.de)



(a) "No Finding" **(b)** "Cardiomegaly" **(c)** "Pneumothorax" **(d)** "Pneumothorax"

Figure 1. Four examples of the ChestX-ray14 dataset. ChestX-ray14 consists of 112,120 frontal chest X-rays from 30,805 patients. All images are labeled with up to 14 pathologies or "No Finding". The dataset does not only include acute findings, as the pneumothorax in figure (**c**), but also treated patients with a drain as "pneumothorax" (**d**).

function and data augmentation. In addition, differing dataset splits were used and only Li *et al.*[11] reported 5-fold cross-validated results. In contrast to these results, our experiments (Sec. 3) demonstrate that performance of a network depends significantly on the selected split. To have a fair comparison, Wang *et al.*[8] released an official split later. Yao *et al.*[13] and Guendel *et al.*[14] reported results for this official split. While Guendel *et al.*[14] hold the state-of-the-art results in all fourteen classes with a location-aware DenseNet-121.

To provide better insights into the effects of distinct design decisions for deep learning, we perform a systematic evaluation using a 5-fold re-sampling scheme. We empirically analyze three major topics:

1. weight initialization, pre-training and transfer learning (Section 2.1)
2. network architectures such as ResNet-50 with large input size (Section 2.2)
3. non-image features such as age, gender, and view position (Section 2.3)

Prior work on ChestX-ray14 has been limited to the analysis of image data. In clinical practice however, radiologists employ a broad range of additional features during the diagnosis. To leverage the complete information of the dataset (i.e. age, gender, and view position), we propose in Section 2.3 a novel architecture integrating this information in addition to the learned image representation.

Methods

In the following, we cast pathology detection as a multi-label classification problem. All images $X = \{\vec{x_1}, \ldots, \vec{x_N}\}$, $\vec{x_i} \in \mathcal{X}$ are associated with a ground truth label $\vec{y_i}$, while we seek a classification function $\vec{f} : \mathcal{X} \rightarrow \mathcal{Y}$ that minimizes a specific loss function l using N training sample-label pairs $(\vec{x_i}, \vec{y_i})$, $i = 1 \ldots N$. Here, we encode the label for each image as a binary vector $\vec{y} \in \{0, 1\}^M = \mathcal{Y}$ (with M labels). We encode "No Finding" as an explicit additional label and hence have $M = 15$ labels. After an initial investigation of weighting loss functions such as positive/negative balancing[8] and class balancing, we noticed no significant difference and decided to employ the class-averaged binary cross entropy (BCE) as our objective:

$$\ell(\vec{y}, \vec{f}) = \frac{1}{M}\sum_{m=1}^{M} H[y_m, f_m], \quad \text{with} \quad H[y, f] = -y \log f - (1 - y)\log(1 - f). \tag{1}$$

Prior work on the ChestX-ray14 dataset concentrates primarily on ResNet-50 and DenseNet-121 architectures. Due to its outstanding performance in the computer vision domain[10], we focus in our experiments on the ResNet-50 architecture[15]. To adapt the network to the new task, we replace the last dense layer of the original architecture with a new dense layer matching the number of labels and add a sigmoid activation function for our multi-label problem (see Table 1).

Weight Initialization and Transfer Learning. We investigate two distinct initialization strategies for the ResNet-50. First, we follow the scheme described by He *et al.*[5], where the network parameters are initialized with random values and thus the model is trained from scratch. Second, we initialize the network with pre-trained weights, where knowledge is transferred from a different domain and task. Furthermore, we distinguish between *off-the-shelf* (OTS) and *fine-tuning* (FT) in the transfer-learning approach.

A major drawback in medical image processing with deep learning is the limited size of datasets compared to the computer vision domain. Hence, training a CNN from scratch is often not feasible. One solution is transfer-learning. Following the notation in the work of Pan *et al.*[16], a source domain $\mathcal{D}_s = \{\mathcal{X}_s, P_s(X_s)\}$ with task $\mathcal{T}_s = \{\mathcal{Y}_s, f_s(\cdot)\}$ and a target domain $\mathcal{D}_t = \{\mathcal{X}_t, P_t(X_t)\}$ with task $\mathcal{T}_t = \{\mathcal{Y}_t, f_t(\cdot)\}$ are given with $\mathcal{D}_s \neq \mathcal{D}_t$ and/or $\mathcal{T}_s \neq \mathcal{T}_t$. In transfer-learning, the knowledge gained in $\mathcal{D}_s$ and $\mathcal{T}_s$ is used to help learning a prediction function $f_t(\cdot)$ in $\mathcal{D}_t$.

Employing an off-the-shelf approach[17,18], the pre-trained network is used as a feature extractor, and only the weights of the last (classifier) layer are adapted. In fine-tuning, one chooses to re-train one or more layers with samples from the new domain. For both approaches, we use the weights of a ResNet-50 network trained on ImageNet as a starting point[19]. In our fine-tuning experiment, we retrained all conv-layers as shown in Table 1.

Layer name	Output size	Original 50-layer	Off-the-shelf	Fine-tuned
conv1	112 × 112	7 × 7, 64-d, stride 2	same	**fine-tuned**
pooling1	56 × 56	3 × 3, 64-d, max pool, stride 2	same	same
conv2_x	56 × 56	$\begin{bmatrix} 1 \times 1, 64\text{-d}, \text{stride1} \\ 3 \times 3, 64\text{-d}, \text{stride1} \\ 1 \times 1, 256\text{-d}, \text{stride1} \end{bmatrix} \times 3$	same	**fine-tuned**
conv3_0	28 × 28	$\begin{bmatrix} 1 \times 1, 128\text{-d}, \text{stride2} \\ 3 \times 3, 128\text{-d}, \text{stride1} \\ 1 \times 1, 512\text{-d}, \text{stride1} \end{bmatrix}$	same	**fine-tuned**
conv3_x	28 × 28	$\begin{bmatrix} 1 \times 1, 128\text{-d}, \text{stride1} \\ 3 \times 3, 128\text{-d}, \text{stride1} \\ 1 \times 1, 512\text{-d}, \text{stride1} \end{bmatrix} \times 3$	same	**fine-tuned**
conv4_0	14 × 14	$\begin{bmatrix} 1 \times 1, 256\text{-d}, \text{stride2} \\ 3 \times 3, 256\text{-d}, \text{stride1} \\ 1 \times 1, 1024\text{-d}, \text{stride1} \end{bmatrix}$	same	**fine-tuned**
conv4_x	14 × 14	$\begin{bmatrix} 1 \times 1, 256\text{-d}, \text{stride1} \\ 3 \times 3, 256\text{-d}, \text{stride1} \\ 1 \times 1, 1024\text{-d}, \text{stride1} \end{bmatrix} \times 5$	same	**fine-tuned**
conv5_0	7 × 7	$\begin{bmatrix} 1 \times 1, 512\text{-d}, \text{stride2} \\ 3 \times 3, 512\text{-d}, \text{stride1} \\ 1 \times 1, 2048\text{-d}, \text{stride1} \end{bmatrix}$	same	**fine-tuned**
conv5_x	7 × 7	$\begin{bmatrix} 1 \times 1, 512\text{-d}, \text{stride1} \\ 3 \times 3, 512\text{-d}, \text{stride1} \\ 1 \times 1, 2048\text{-d}, \text{stride1} \end{bmatrix} \times 2$	same	**fine-tuned**
pooling2	1 × 1	7 × 7, 2048-d, average pool, stride 1	same	same
dense	1 × 1	1000-d, dense-layer	**15-d, dense-layer**	
loss	1 × 1	1000-d, softmax	**15-d, sigmoid, BCE**	

Table 1. Architecture of the original, off-the-shelf, and fine-tuned ResNet-50. In our experiments, we use the ResNet-50 architecture and this table shows differences between the original architecture and ours (off-the-shelf and fine-tuned ResNet-50). If there is no difference to the original network, the word "same" is written in the table. The violet and bold text emphasizes, which parts of the network are changed for our application. All layers do employ automatic padding (i.e. depending on the kernel size) to keep spatial size the same. The conv3_0, conv4_0, and conv5_0 layers perform a down-sampling of the spatial size with a stride of 2.

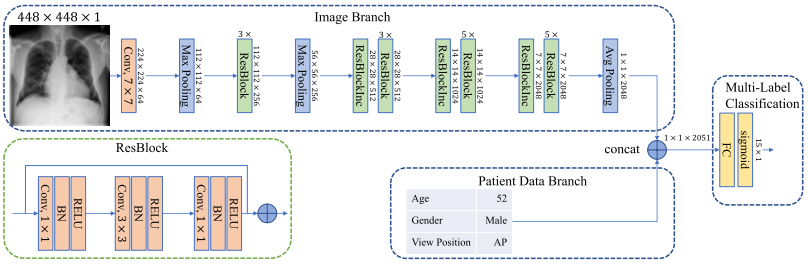


Figure 2. Patient-data adapted model architecture: ResNet-50-large-meta. Our architecture is based on the ResNet-50 model. Because of the enlarged input size, we added a max-polling layer after the first three ResBlocks. In addition, we fused image features and patient features at the end of our model to incorporate patient information.

Architectures. In addition to the original ResNet-50 architecture, we employ two variants: First, we reduce the number of input channels to one (the ResNet-50 is designed for the processing of RGB images from the ImageNet dataset), which should facilitate the training of an X-ray specific CNN. Second, we increase the input size by a factor of two (i.e. 448 × 448). To keep the model architectures similar, we only add a new max-pooling layer after the first bottleneck block. This max-pooling layer has the same parameters as the "pooling1" layer (i.e. 3 × 3 kernel, stride 2, and padding). In Fig. 2, our changes are illustrated at the image branch. A higher effective resolution could be beneficial for the detection of small structures, which could be indicative of a pathology (e.g. masses and nodules). In the following, we use the postfix "-1channel" and "-large" to refer to our model changes.

Finally, we investigate different model depths with the best performing setup. First, we implement a shallower ResNet-38 where we reduce the number of bottleneck blocks for conv2_x, conv3_x, and conv4_x down to two,

(a) Diseases			
Pathology	**True**	**False**	**Prevalence [%]**
Cardiomegaly	2,776	109,344	2.48
Emphysema	2,516	109,604	2.24
Edema	2,303	109,817	2.05
Hernia	227	111,893	0.20
Pneumothorax	5,302	106,818	4.73
Effusion	13,317	98,803	11.88
Mass	5,782	106,338	5.16
Fibrosis	1,686	110,434	1.50
Atelectasis	11,559	100,561	10.31
Consolidation	4,667	107,453	4.16
Pleural Thicken.	3,385	108,735	3.02
Nodule	6,331	105,789	5.65
Pneumonia	1,431	110,689	1.28
Infiltration	19,894	92,226	17.74

Table 2. Overview of label distributions in the ChestX-ray14 dataset.

(b) Meta-information			
	Female	**Male**	**Ratio**
Patient Gender	63,340	48,780	1.30
	PA	AP	Ratio
View Position	67,310	44,810	1.50

Table 3. Overview of label distributions in the ChestX-ray14 dataset.

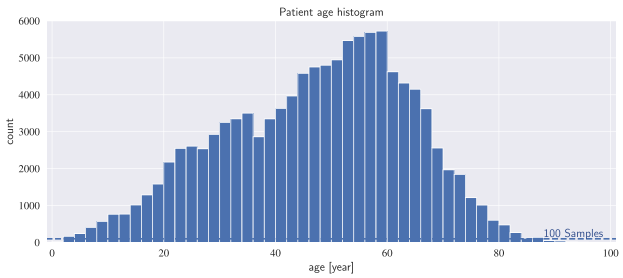


Figure 3. Distribution of patient age in the ChestX-ray14 dataset. Each bin covers a width of two years. The average patient age is 46.87 years with a standard deviation of 16.60 years.

two, and three, respectively. Secondly, we also test the ResNet-101 and increased the number of conv_3 blocks from 5 to 22 compare to the ResNet-50.

Non-Image Features. ChestX-ray14 contains information about the patient age, gender, and view position (i.e. if the X-ray image is acquired posterior-anterior (PA) or anterior-posterior (AP)). Radiologists use information beyond the image to conclude which pathologies are present or not. The view position changes the expected position of organs in the X-ray images (i.e. PA images are horizontally flipped compared to AP). In addition, organs (e.g. the heart) are magnified in an AP projection as the distance to the detector is increased.

As illustrated in Fig. 2, we concatenate the image feature vector (i.e. output of the last pooling layer with dimension 2024×1) with the new non-image feature vector (with dimension 3×1). Therefore, view position and gender is encoded as $\{0,1\}$ and the age is linearly scaled $[\ \min(X_{\text{pa}}),\ \max(X_{\text{pa}})] \mapsto [0,1]$, in order to avoid a bias towards features with a large range of values. In our experiments, we used "-meta" to refer our model architecture with non-image features.

ChestX-ray14 Dataset. To evaluate our approaches for multi-label pathology classification, the entire corpus of ChestX-ray14 (Fig. 1) is employed. In total, the dataset contains 112, 120 frontal chest X-rays from 30,805 patients. The dataset does not include the original DICOM images but Wang *et al.*[8] performed a simple preprocessing based on the encoded display settings while the pixel depth was reduced to 8-bit. In addition, each

	Without non-image features				With non-image features			
Pathology	OTS	FT	1channel	large	OTS	FT	1channel	large
Cardiomegaly	72.7 ± 1.8	88.5 ± 0.7	88.9 ± 0.5	89.7 ± 0.3	75.9 ± 1.4	88.4 ± 0.8	**90.2 ± 0.4**	89.8 ± 0.8
Emphysema	77.8 ± 2.1	89.2 ± 1.0	87.0 ± 0.8	88.3 ± 1.3	79.8 ± 1.9	**89.4 ± 1.2**	87.4 ± 1.3	89.1 ± 1.2
Edema	84.4 ± 0.6	**89.1 ± 0.4**	**89.1 ± 0.6**	88.8 ± 0.5	85.7 ± 0.5	**89.1 ± 0.7**	89.0 ± 0.6	88.9 ± 0.3
Hernia	78.8 ± 1.4	85.5 ± 3.8	88.1 ± 4.2	87.5 ± 4.5	81.9 ± 2.5	88.2 ± 3.2	89.3 ± 4.4	**89.6 ± 4.4**
Pneumothorax	77.3 ± 1.3	**87.0 ± 0.8**	85.7 ± 0.9	85.9 ± 0.9	79.1 ± 1.2	86.5 ± 0.6	85.4 ± 0.7	85.9 ± 1.1
Effusion	79.4 ± 0.4	87.1 ± 0.2	**87.6 ± 0.2**	**87.6 ± 0.2**	80.6 ± 0.4	87.2 ± 0.3	**87.6 ± 0.2**	87.3 ± 0.3
Mass	66.8 ± 0.6	82.2 ± 1.0	83.3 ± 0.6	**83.9 ± 0.9**	68.6 ± 0.6	82.2 ± 1.0	83.3 ± 0.7	83.2 ± 0.3
Fibrosis	72.0 ± 0.9	**80.0 ± 0.9**	79.9 ± 0.8	79.2 ± 1.6	73.9 ± 0.8	**80.0 ± 0.9**	79.6 ± 0.5	78.9 ± 0.5
Atelectasis	71.8 ± 0.6	**80.3 ± 0.7**	79.9 ± 0.4	79.2 ± 0.7	73.2 ± 0.7	80.1 ± 0.6	79.3 ± 0.6	79.1 ± 0.4
Consolidation	74.3 ± 0.3	79.5 ± 0.5	**80.6 ± 0.4**	80.0 ± 0.3	75.3 ± 0.3	79.6 ± 0.5	80.4 ± 0.5	80.0 ± 0.7
Pleural Thicken.	68.8 ± 1.0	**79.0 ± 0.7**	78.4 ± 0.9	78.0 ± 1.1	70.8 ± 1.1	78.6 ± 1.1	78.2 ± 1.3	77.1 ± 1.3
Nodule	65.0 ± 0.8	72.6 ± 0.9	73.3 ± 0.8	75.1 ± 1.3	66.5 ± 0.7	74.7 ± 0.6	74.0 ± 0.7	**75.8 ± 1.4**
Pneumonia	66.4 ± 2.7	74.4 ± 1.6	74.3 ± 1.5	75.3 ± 2.2	68.3 ± 2.3	73.3 ± 1.3	74.8 ± 1.5	**76.7 ± 1.5**
Infiltration	65.9 ± 0.2	69.9 ± 0.6	**70.2 ± 0.3**	**70.2 ± 0.5**	67.0 ± 0.4	**70.2 ± 0.2**	70.1 ± 0.5	70.0 ± 0.7
Average	73.0 ± 1.1	81.7 ± 1.0	81.9 ± 0.9	82.1 ± 1.2	74.8 ± 1.1	82.0 ± 0.9	82.0 ± 1.0	**82.2 ± 1.1**
No Findings	71.6 ± 0.3	76.9 ± 0.5	**77.3 ± 0.3**	77.1 ± 0.4	72.5 ± 0.3	76.8 ± 0.4	77.1 ± 0.4	77.1 ± 0.3

Table 4. AUC result overview for all our experiments. In this table, we present averaged results over all five splits and the calculated standard deviation (std) for each pathology. We divide our experiments into three categories. First, without and with non-image features. Second, transfer-learning with off-the-shelf (OTS) and fine-tuned (FT) models. Third, from scratch where "1channel" refers to same input size as in transfer-learning but changed number of channels. "large" means we changed the input dimensions to $448 \times 448 \times 1$. For better comparison, we present the average AUC and the standard deviation over all pathologies in the last row. Bold text emphasizes the overall highest AUC value. Values are scaled by 100 for convenience.

image was resized to 1024×1024 pixel without preserving the aspect ratio. In Table 2 and Fig. 3, we show the distribution of each class and the statistics for non-image information. The prevalence of individual pathologies is generally low and varies between 0.2% and 17.74% as shown in Table 2. While, the distribution of patient gender and view position is quite even with a ratio of 1.3 and 1.5, respectively (see Table 3). In Fig. 3, the histogram shows the distribution of patient age in ChestX-ray14. The average patient age is 46.87 years with a standard deviation of 16.60 years.

To determine if the provided non-image features contain information for a disease classification, we performed an initial experiment. We trained a very simple Multi-layer Perceptron (MLP) classifier only with the three non-image feature as input. The MLP classifier has a low average AUC of 0.61 but this still indicates that those non-image features could help to improve classification results when provided to our novel model architecture.

Experiments and Results

For an assessment of the generalization performance, we perform a 5 times re-sampling scheme[20]. Within each split, the data is divided into 70% training, 10% validation, and 20% testing. When working with deep learning, hyper-parameters, and tuning without a validation set and/or cross-validation can easily result in over-fitting. Since individual patients have multiple follow-up acquisitions, all data from a patient is assigned to a single subset only. This leads to a large patient number diversity (e.g. split two has 5,817 patients and 22,420 images whereas split 5 has 6,245 patients and the same number of images). We estimate the average validation loss over all re-samples to determine the best models. Finally, our results are calculated for each fold on the test set and averaged afterwards.

To have a fair comparison to other groups, we conduct an additional evaluation using the best performing architecture with different depth on the official split of Wang *et al.*[8] in Section 3.1.

Implementation. In all experiments, we use a fixed setup. To extend ChestX-ray14, we use the same geometric data augmentation as in the work of Szegedy *et al.*[3]. At training, we sample various sized patches of the image with sizes between 8% and 100% of the image area. The aspect ratio is distributed evenly between 3:4 and 4:3. In addition, we employ random rotations between $\pm 7°$ and horizontal flipping. For validation and testing, we rescale images to 256×256 and 480×480 pixels for small and large spatial size, respectively. Afterwards, we use the center crop as input image. As in the work of He *et al.*[5], dropout is not employed[21]. As optimizer, we use ADAM[22] with default parameters for $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The learning rate lr is set to $lr = 0.001$ and $lr = 0.01$ for transfer-learning and from scratch, respectively. While training, we reduce the learning rate by a factor of 2 when the validation loss does not improve. Due to model architecture variations, we use batch sizes of 16 and 8 for transfer-learning and from scratch with a large input size, respectively. The models are implemented in CNTK and trained on GTX 1080 GPUs yielding a processing time of around 10 ms per image.

		Without				With			
		OTS	FT	1channel	large	OTS	FT	1channel	large
Without	OTS	—	0.65	0.74	0.73	0.46	0.38	0.40	0.59
	FT	0.65	—	0.81	0.80	0.38	0.42	0.43	0.64
	1channel	0.74	0.81	—	0.93	0.41	0.43	0.47	0.71
	large	0.73	0.80	0.93	—	0.40	0.43	0.47	0.71
With	OTS	0.46	0.38	0.41	0.40	—	0.32	0.33	0.39
	FT	0.38	0.42	0.43	0.43	0.32	—	0.35	0.42
	1channel	0.40	0.43	0.47	0.47	0.33	0.35	—	0.45
	large	0.59	0.64	0.71	0.71	0.39	0.42	0.45	—

Table 5. Spearman's rank correlation coefficient is calculated between all model pairs and is averaged over all five splits. Our experiments are grouped into three categories. First, "Without" and "With" non-image features. Second, transfer-learning with off-the-shelf (OTS) and fine-tuned (FT) models. Third, from scratch where "1channel" refers to same input size as in transfer-learning but changed number of channels. "large" means we changed the input dimensions to $448 \times 448 \times 1$. We identify three clusters: all models under "With", models trained from scratch and "Without", and the "OTS" model.

Results. Table 4 summarizes the outcome of our evaluation. In total, we evaluate eight different experimental setups with varying weight initialization schemes and network architectures as well as with and without non-image features. We perform an ROC analysis using the area under the curve (AUC) for all pathologies, compare the classifier scores by Spearman's pairwise rank correlation coefficient, and employ the state-of-the-art method Gradient-weighted Class Activation Mapping (Grad-CAM)[23] to gain more insight into our CNNs. Grad-CAM is a method for visually assessing CNN model predictions. The method highlights important regions in the input image for a specific classification result by using the gradient of the final convolutional layer.

The results indicate a high variability of the outcome with respect to the selected dataset split. Especially for "Hernia", which is the class with the smallest number of positive samples, we observe a standard deviation of up to 0.05. As a result, an assessment of existing approaches and comparison of their performance is difficult, since prior work focuses mostly on a single (random) split.

With respect to the different initialization schemes, we observe already reasonable results for OTS networks that are optimized on natural images. Using fine-tuning, the results are improved considerably, from 0.730 to 0.819 AUC on average. A complete training of the ResNet-50-1channel using CXRs results in a rather comparable performance. Only the high-resolution variant of the ResNet-50-large outperforms the FT approach by 0.002 on average AUC. In particular, for smaller pathologies like nodules and masses an improvement is observed (i.e. 0.018 and 0.006 AUC increase, respectively), while for other pathologies a similar, or slightly lower performance is estimated.

Finally, all our experiments with non-image features slightly increase the AUC on average to its counterpart (i.e. without non-image feature). Our from scratch trained ResNet-50-large-meta yields the best overall performance with 0.822 average AUC.

To get a better insight why the non-image features only slightly increased the AUC for our fine-tuned and from scratch trained models, we investigated the capability of predict the non-image features based on the extracted image features. We used our from scratch trained model (i.e. ResNet-50-large) as a feature extractor and trained three models to predict the patient age, patient gender, and view position (VP) – i.e. ResNet-50-large-age, ResNet-50-large-gender, ResNet-50-large-VP. We employed the same training setup as in our experiments before. First, our ResNet-50-large-VP model can predict with a very high AUC of 0.9983 ± 0.0002 the correct VP (i.e. we encoded AP as true and PA as false). After choosing the optimal threshold based on Youden index, we calculated a sensitivity and specificity of 99.3% and 99.1%, respectively. Secondly, the ResNet-50-large-gender predicts the patient gender also very precisely with a high AUC of 0.9435 ± 0.0067. The sensitivity and specificity with 87.8% and 85.9% is also high. Finally, to evaluate the performance of the ResNet-50-large-age we report the mean absolute error (MAE) with standard deviation because age prediction is a regression task. The model achieved a mean absolute error of 9.13 ± 7.05 years. The results show that the image features already encode information about the non-image features. This might be the reason that our proposed model architecture with the non-image features at hand did not increased the performance by a large margin.

Furthermore, the similarity between the trained models in terms of their predictions was investigated. Therefore, Spearman's rank correlation coefficient was computed for the predictions of all model pairs, and averaged over the folds. The pairwise correlations coefficients for the models are given in Table 5. Based on the degree of correlation, three groups can be identified. First, we note that the "from scratch models" (i.e. "1channel" and "large") without non-image features have the highest correlation of 0.93 amongst each other, followed by the fine-tuned models with 0.81 and 0.80 for "1channel" and "large", respectively. Second, the OTS model surprisingly has higher correlation with the from scratch models than the fine-tuned model. Third, for models with non-image feature, no such correlation is observed and their value is between 0.32 to 0.47. This indicates that models which have been trained exclusively on X-ray data achieve not only the highest accuracy, but are furthermore most consistent.

While our proposed network architecture achieves high AUC values in all categories of the ChestX-ray14 dataset, the applicability of such a technology in a clinical environment depends considerably on the availability

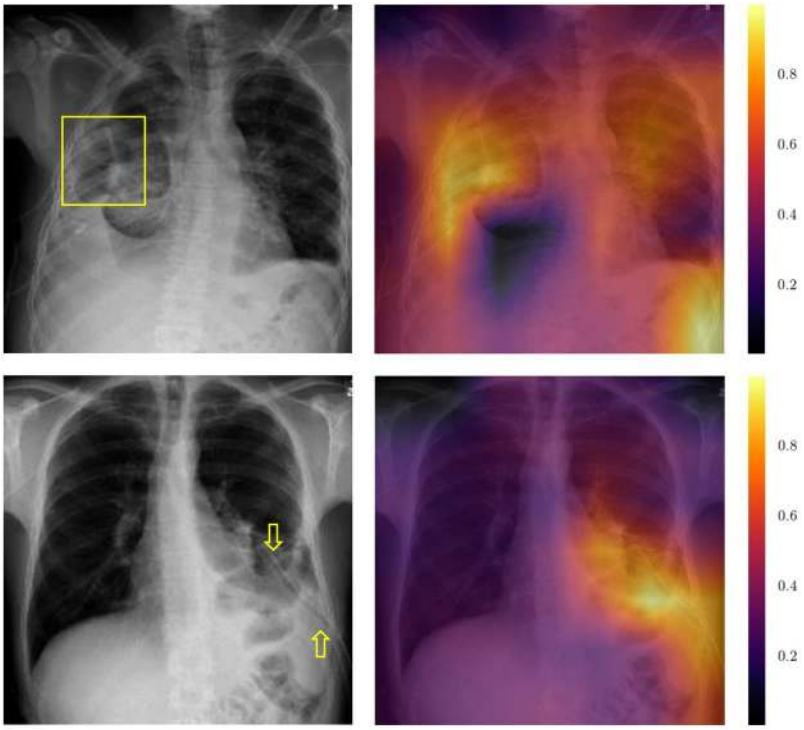


Figure 4. Grad-CAM result for two example images. In the first one, we marked the location of the pneumothorax with a yellow box. As shown in the Grad-CAM image next to it, the models highest activation for the prediction is within the correct area. The second row shows a negative example where the highest activation, which was responsible for the final predication "pneumothorax", is at the drain. This indicates that our trained CNN picked up drains as a main feature for "pneumothorax". We marked the drain with yellow arrows.

of data for model training and evaluation. In particular, for the NIH dataset the reported label noise[8] and the medical interpretation of the label are an important issue. As mention by Luke Oakden-Rayner[24], the class "pneumothorax" is often labeled for already treated cases (i.e. a drain is visible in the image which is used to tread the pneumothorax) in the ChestX-ray14 dataset. We employ Grad-CAM to get an insight, if our trained CNN picked up the drain as a main feature for "pneumothorax". Grad-CAM visualizes the areas which are most responsible for the final prediction as a heatmap. In Fig. 4, we show two examples of our test set where the highest activations are around the drain. This indicates that the network learned not only to detect an acute pneumothorax but also the presence of chest drains. Therefore, the utility of the ChestX-ray14 dataset for the development of clinical applications is still an open issue.

Comparison to other approaches. In our evaluation, we noticed a considerable spread of the results in terms of AUC values. Next to the employed data splits, this could be attributed to the (random) initialization of the models, and the stochastic nature of the optimization process.

When ChestX-ray14 was made publicly available, only images and no official dataset splitting was released. Hence, researcher started to train and test their proposed methods on their own dataset split. We noticed a large diversity in performance with different splits of our re-sampling. Therefore, a direct comparison to other groups might be miss leading in the sense of state-of-the-art results. For example, Rajpurkar *et al.*[12] reported state-of-the-art results for all 14 classes on their own split. In Fig. 5, we compare our best performing model architecture (i.e. ResNet-50-large-meta) of the re-sampling experiments to Rajpurkar *et al.* and other groups. For our model, we plot the minimum and maximum AUC over all re-samplings as error bars to illustrate the effect of random splitting. We achieve state-of-the-art results for "effusion" and "consolidation" when directly comparing our AUC (i.e. averaged over 5 times re-sampling) to former state-of-the-art results. Comparing the maximum AUC over all re-sampling splits results in state-of-the-art performance for "effusion", "pneumonia", "consolidation", "edema", and "hernia" and indicates that a fair comparison between groups without the same splitting might be non-conclusive.

Later, Wang *et al.*[8] released an official split of the ChestX-ray14 dataset. To have a fair comparison to other groups, we report results on this split for our best performing architecture with different depths – ResNet-38-large-meta, ResNet-50-large-meta, and ResNet-101-large-meta – in Table 6. First, we compare our results to Wang *et al.*[8] and Yao *et al.*[13] because Guendel *et al.*[14] used an additional dataset – PLCO dataset[25] – with 185,000 images. While the ResNet-101-large-meta already has a higher average AUC with 0.785 and in 12 out of 14 classes a higher individual AUC, the performance is compared to our ResNet-38-large-meta and ResNet-50-larg-meta lower. Reducing the number of layers increased the averaged AUC from 0.785 to 0.795 and 0.806 for ResNet50-large-meta and ResNet38-larg-meta, respectively. Hence, our results indicate that training a model

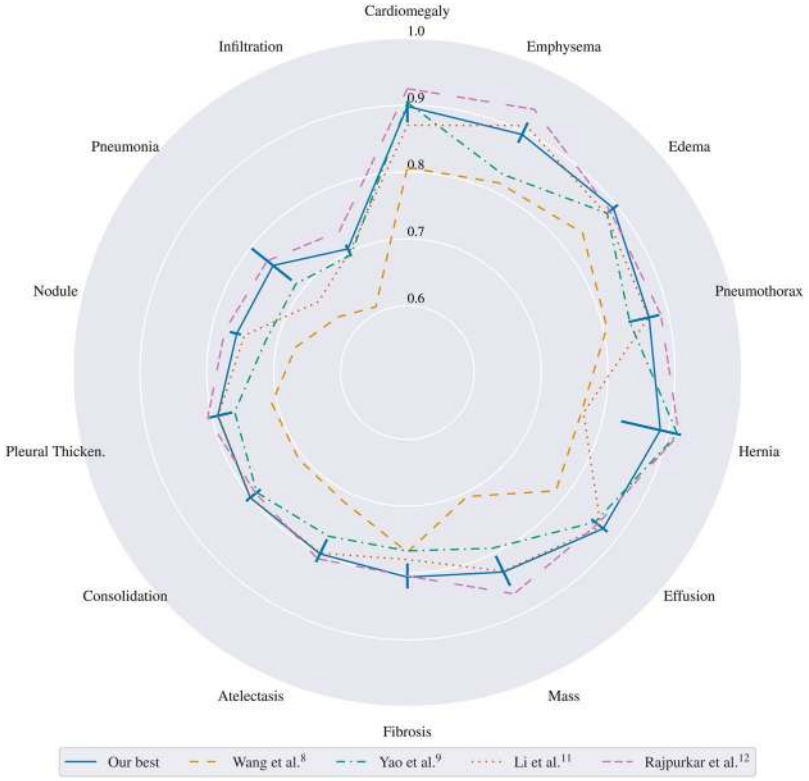


Figure 5. Comparison of our best model to other groups. We sort the pathologies with increasing average AUC over all groups. For our model, we report the minimum and maximum over all folds as error bar to illustrate the effect of splitting.

Pathology	Wang *et al.*[8]	Yao *et al.*[13]	Guendel *et al.*[14]	"-large-meta"		
				ResNet-38	ResNet-50	ResNet-101
Cardiomegaly	0.810	0.856	**0.883**	0.875	0.877	0.865
Emphysema	0.833	0.842	**0.895**	**0.895**	0.875	0.868
Edema	0.805	0.806	0.835	**0.846**	0.842	0.828
Hernia	0.872	0.775	0.896	**0.937**	0.916	0.855
Pneumothorax	0.799	0.805	**0.846**	0.840	0.819	0.839
Effusion	0.759	0.806	**0.828**	0.822	0.818	0.818
Mass	0.693	0.777	**0.821**	0.820	0.810	0.796
Fibrosis	0.786	0.743	**0.818**	0.816	0.800	0.778
Atelectasis	0.700	0.733	**0.767**	0.763	0.755	0.747
Consolidation	0.703	0.711	0.745	**0.749**	0.742	0.734
Pleural Thicken.	0.684	0.724	0.761	**0.763**	0.742	0.739
Nodule	0.669	0.724	**0.758**	0.747	0.736	0.738
Pneumonia	0.658	0.684	**0.731**	0.714	0.703	0.694
Infiltration	0.661	0.673	**0.709**	0.694	0.694	0.686
Average	0.745	0.761	**0.807**	0.806	0.795	0.785
No Findings	—	—	—	0.727	0.725	0.720

Table 6. AUC result overview for our experiments on the official split. In this table, we present results for our best performing architecture with different depth (i.e. ResNet38-large-meta, ResNet50-large-meta, ResNet101-large-meta) and compare them to other groups. Additionally we provide an average AUC over all pathologies in the last row. Bold text emphasizes the overall highest AUC value.

with less parameter on Chest-Xray14 is beneficial for the overall performance. Secondly, Guendel *et al.*[14] reported state-of-the-art results for the official split in all 14 classes with an averaged AUC of 0.807. While our ResNet-38-large-meta is trained with 185,000 images less, it still achieved state-of-the-art results for "Emphysema", "Edema", "Hernia", "Consolidation", and "Pleural Thicken." and a slight less average AUC of 0.806.

Discussion and Conclusion

We present a systematic evaluation of different approaches for CNN-based X-ray classification on ChestX-ray14. While satisfactory results are obtained with networks optimized on the ImageNet dataset, the best overall results can be reported for the model that is exclusively trained with CXRs and incorporates non-image data (i.e. view position, patient age, and gender).

Our optimized ResNet-38-large-meta architecture achieves state-of-the art results in five out of fourteen classes compared to Guendel *et al.*[14] (who had state-of-the-art results in all fourteen classes on the official split). For other classes even higher scores are reported in the literature (see e.g. Rajpurkar *et al.*[12]). However, a comparison of the different CNN methods with respect to their performance is inherently difficult, as most evaluations have been performed on individual (random) partitions of the datasets. We observed substantial variability in the results when different splits are considered. This becomes especially apparent for "Hernia", the class with the fewest samples in the dataset (see also Fig. 5).

While the obtained results suggest that the training of deep neural networks in the medical domain is a viable option as more and more public datasets become available, the practical use of deep learning in clinical practice is still an open issue. In particular, for the ChestX-ray14 datasets, the rather high label noise[8] of 10% makes an assessment of the true network performance difficult. Therefore, a clean test set without label noise is needed for clinical impact evaluation. As discussed by Oakden-Rayner[24], the quality of the (automatically generated) labels and their precise medical interpretation may be a limiting factor addition to the presence of treated findings. Our Grad-CAM results proves Oakden- Rayner's concerns about the "pneumothorax" label. In a clinical setting, i.e. for the detection of critical findings, the focus would be on the reliably identification of acute cases of pneumothorax, while a network trained on ChestX-ray14 would also respond to cases with a chest drain.

Future work will include investigation of other model architectures, new architectures for leveraging label dependencies and incorporating segmentation information.

Data Availability

The datasets analyzed during the current study are available in the ChestXray-NIHCC repository, https://nihcc.app.box.com/v/ChestXray-NIHCC.

References

1. Commission, C. Q. Queen Alexandra hospital quality report. Available at, https://www.cqc.org.uk/location/RHU03 (2017).
2. Krizhevsky, A., Sutskever, I. & Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, 1097–1105 (2012).
3. Szegedy, C. *et al*. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1–9, https://doi.org/10.1109/CVPR.2015.7298594 (2015).
4. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).
5. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778, https://doi.org/10.1109/CVPR.2016.90 (2016).
6. Bar, Y. *et al*. Chest pathology detection using deep learning with non-medical training. In *2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, 294–297 (Citeseer, 2015).
7. Demner-Fushman, D. *et al*. Preparing a collection of radiology examinations for distribution and retrieval. *J. Am. Med. Informatics Assoc.* **23**, 304–310 (2015).
8. Wang, X. *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)*, 3462–3471, 10.1109/CVPR.2017.369 (2017).
9. Yao, L. *et al*. Learning to diagnose from scratch by exploiting dependencies among labels. *arXiv preprint arXiv:1710.10501* (2017).
10. Huang, G., Liu, Z., v. d. Maaten, L. & Weinberger, K. Q. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2261–2269, https://doi.org/10.1109/CVPR.2017.243 (2017).
11. Li, Z. *et al*. Thoracic disease identification and localization with limited supervision. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8290–8299, https://doi.org/10.1109/CVPR.2018.00865 (2018).
12. Rajpurkar, P. *et al*. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225* (2017).
13. Yao, L., Prosky, J., Poblenz, E., Covington, B. & Lyman, K. Weakly supervised medical diagnosis and localization from multiple resolutions. *arXiv preprint arXiv:1803.07703* (2018).
14. Guendel, S. *et al*. Learning to recognize abnormalities in chest x-rays with location-aware dense networks. *arXiv preprint arXiv:1803.04565* (2018).
15. He, K., Zhang, X., Ren, S. & Sun, J. Identity mappings in deep residual networks. In *Computer Vision – ECCV 2016*, 630–645 (Springer International Publishing, 2016).
16. Pan, S. J. & Yang, Q. A survey on transfer learning. *IEEE Transactions on Knowl. Data Eng.* **22**, 1345–1359, https://doi.org/10.1109/TKDE.2009.191 (2010).
17. Yosinski, J., Clune, J., Bengio, Y. & Lipson, H. How transferable are features in deep neural networks? In *Advances in neural information processing systems*, 3320–3328 (2014).
18. Razavian, A. S., Azizpour, H., Sullivan, J. & Carlsson, S. Cnn features off-the-shelf: An astounding baseline for recognition. In *2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 512–519, https://doi.org/10.1109/CVPRW.2014.131 (2014).
19. Russakovsky, O. *et al*. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.* **115**, 211–252 (2015).
20. Molinaro, A. M., Simon, R. & Pfeiffer, R. M. Prediction error estimation: a comparison of resampling methods. *Bioinforma.* **21**, 3301–3307, https://doi.org/10.1093/bioinformatics/bti499 (2005).
21. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15**, 1929–1958 (2014).
22. Kingma, D. P. & Ba, J. L. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)* (2015).
23. Selvaraju, R. R. *et al*. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 618–626, https://doi.org/10.1109/ICCV.2017.74 (2017).

24. Oakden-Rayner, L. Exploring the ChestXray14 dataset: Problems. Available at, https://lukeoakdenrayner.wordpress.com/2017/12/18/ (2017).
25. Gohagan, J. K., Prorok, P. C., Hayes, R. B. & Kramer, B.-S. The prostate, lung, colorectal and ovarian (PLCO) cancer screening trial of the national cancer institute: history, organization, and status. *Control. clinical trials* **21**, 251S–272S (2000).

Acknowledgements
Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - Projektnummer 392323616 and the Hamburg University of Technology (TUHH) in the funding programme *Open Access Publishing*. We would like to thank NIH Clinical Center for their publicly released dataset (i.e. ChestX-ray14).

Author Contributions
I.M.B., A.S. and H.N. conceived the experiments, I.M.B. conducted the experiments, I.M.B., A.S. and H.N. analyzed the results. All authors reviewed the manuscript.

Additional Information
Competing Interests: M.G., H.N. and A.S. are employees of Philips Research, Hamburg, Germany. I.M.B. and T.K. declare no potential conflict of interest.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.