# Differentiation Between Anteroposterior and Posteroanterior Chest X-Ray View Position With Convolutional Neural Networks

## Unterscheidung der Anterior-posterior und Posterior-Anterior Röntgen-Thoraxaufnahme mit Convolutional Neural Networks

**Authors**
**René Hosch, Lennard Kroll, Felix Nensa, Sven Koitka**

**Affiliation**
Institute of Diagnostic and Interventional Radiology and Neuroradiology, University Hospital Essen, Germany

**Key words**
radiography, deep learning, anteroposterior, posteroanterior, error correction, X-ray

**received** 02.02.2020
**accepted** 11.05.2020
**published online** 02.07.2020

**Bibliography**
Fortschr Röntgenstr 2021; 193: 168–176
**DOI** 10.1055/a-1183-5227
**ISSN** 1438-9029

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

**Correspondence**
René Hosch
Institute of Diagnostic and Interventional Radiology and Neuroradiology, University Hospital Essen, Hufelandstraße 55, 45122 Essen, Germany
Tel.: ++ 49 201/723-7882
rene.hosch@uk-essen.de

### ZUSAMMENFASSUNG

**Ziel** Detektion der Röntgen-Thorax-Aufnahmeposition anhand von Convolutional Neural Networks zur Verbesserung und Bereinigung von Metainformationen innerhalb der Dateninfrastruktur eines Krankenhauses.

**Material und Methoden** Innerhalb dieser Studie wurde ein Convolutional Neural Network entwickelt, das automatisch die verwendete Anterior-posterior- bzw. Posterior-anterior-Aufnahmeprojektion einer Röntgen-Thoraxaufnahme erkennt. Es wurden 2 unterschiedliche Netzwerkarchitekturen (VGG Variante und ResNet-34) auf Basis von Daten der RSNA (26 684 Röntgenaufnahmen, Klassenverteilung: 46 % AP, 54 % PA) trainiert und anschließend auf einem zusammengestellten hauseigenen Datensatz (Verwendung von manuellen Labeln) aus dem Datenbestand des Universitätsklinikums Essen (4507 Röntgenaufnahmen, Klassenverteilung: 55 % PA, 45 % AP) getestet. Für eine bessere Nachvollziehbarkeit der getätigten Vorhersagen der Modelle wurde zudem für jede Vorhersage eine Grad-CAM generiert. Die Resultate der Modelle wurden anhand der Accuracy, der Area under the Curve (AUC) und dem F1-Score berechnet auf Basis des Abgleichs der manuellen Label. Abschließend wurde zudem die Genauigkeit der Modellvorhersagen und der DICOM-Label anhand des Vergleichs mit den manuellen Labeln berechnet.

**Ergebnisse** Die zusammengefassten Modelle erreichten Accuracy- und F1-Score-Werte von mehr als 95 %. Alle Modelle erreichten eine AUC von über 0,99. Die generierten Grad-CAMSs zeigen, dass die Modelle relevante anatomische Referenzpunkte für ihre Vorhersage nutzen, die auch ein Radiologe für eine Unterscheidung heranziehen würde. Zudem zeigen die antrainierten Modelle die Fähigkeit zur Generalisierung, da diese auch falsch gekennzeichnete Röntgenbilder richtig einordnen können, was durch den Vergleich der manuellen Label mit den jeweiligen Modellvorhersagen und den DICOM-Labeln ersichtlich wurde.

**Schlussfolgerung** Die Resultate zeigen, dass falsch eingetragene Metainformationen innerhalb der radiologischen Bildgebung effektiv durch den Einsatz von Deep Learning korrigiert und somit die Datenqualität sowohl für die klinische Anwendung als auch für die Forschung erhöht werden können.

**Kernaussagen:**
- Die trainierten Modelle erzielen akkurate Vorhersagen auf externen Validierungsdaten.
- Die Netzwerke treffen ihre Vorhersagen basierend auf anatomischen Strukturen und Referenzpunkten, die mit dem menschlichen Fachwissen übereinstimmen.
- Die finalen Modelle konnten Label-Fehler in dem Testdatensatz finden.

### ABSTRACT

**Purpose** Detection and validation of the chest X-ray view position with use of convolutional neural networks to improve meta-information for data cleaning within a hospital data infrastructure.

**Material and Methods** Within this paper we developed a convolutional neural network which automatically detects the anteroposterior and posteroanterior view position of a chest radiograph. We trained two different network architec-

tures (VGG variant and ResNet-34) with data published by the RSNA (26 684 radiographs, class distribution 46 % AP, 54 % PA) and validated these on a self-compiled dataset with data from the University Hospital Essen (4507, radiographs, class distribution 55 % PA, 45 % AP) labeled by a human reader. For visualization and better understanding of the network predictions, a Grad-CAM was generated for each network decision. The network results were evaluated based on the accuracy, the area under the curve (AUC), and the F1-score against the human reader labels. Also a final performance comparison between model predictions and DICOM labels was performed.

**Results** The ensemble models reached accuracy and F1-scores greater than 95 %. The AUC reaches more than 0.99 for the ensemble models. The Grad-CAMs provide insight as to which anatomical structures contributed to a decision by the networks which are comparable with the ones a radiologist would use. Furthermore, the trained models were able to generalize over mislabeled examples, which was found by comparing the human reader labels to the predicted labels as well as the DICOM labels.

**Conclusion** The results show that certain incorrectly entered meta-information of radiological images can be effectively corrected by deep learning in order to increase data quality in clinical application as well as in research.

**Key Points:**

- The predictions for both view positions are accurate with respect to external validation data.
- The networks based their decisions on anatomical structures and key points that were in-line with prior knowledge and human understanding.
- Final models were able to detect labeling errors within the test dataset.

**Citation Format**

- Hosch R, Kroll L, Nensa F et al. Differentiation Between Anteroposterior and Posteroanterior Chest X-Ray View Position With Convolutional Neural Networks. Fortschr Röntgenstr 2021; 193: 168–176

## Introduction

The usage and importance of deep learning applications within the radiological workflow is increasing. A majority of scientific research is based on the aim of the automatic detection of diseases on CT, MR or X-ray images via deep learning algorithms [1–3]. To implement those kinds of algorithms researchers have to rely on valid information including not only the image data itself but also important metadata that are needed for the training and decision process [4]. Metadata such as the view position of X-ray images can play a key role for image interpretation by radiologists or a diagnostic algorithm within an automated diagnostic pipeline [5]. In the case of X-ray images, more than 1 billion radiological imaging procedures are performed worldwide per year. One of the most frequently performed examinations is chest X-ray [1, 6]. In general, we distinguish between the posteroanterior (PA) and anteroposterior (AP) view position as shown in ▶ **Fig. 1**.

The correct distinction between these two positions is significant because the view position can be decisive for image interpretation [5]. For example, for patients with cardiomegaly or pneumothorax, the PA position delivers more relevant information than the AP position because of less geometric magnification of anatomical structures such as the heart due to increased distance to the detector [5]. This shows which impact meta information like the view position can have within the radiological workflow. Comparing the importance of correct meta information with the potential room for error within the work routine shows that incorrect metadata can lead to billing errors, poor quality of research data or worse, e. g., incorrect diagnoses and treatments [7].

The goal of the present study was to design and train a convolutional neural network (CNN) to derive the correct view position of chest X-rays from the imaging data itself and thus be able to correct erroneously entered metadata.

## Materials and Methods

### Ethics Statement

This study was in compliance with the guidelines of the Institutional Review Board of the University Hospital Essen – Approval Number: 19-8916-BO. Due to the retrospective nature of the study, written informed consent was waived by the Institutional Review Board. The data were completely anonymized before being included in the study.

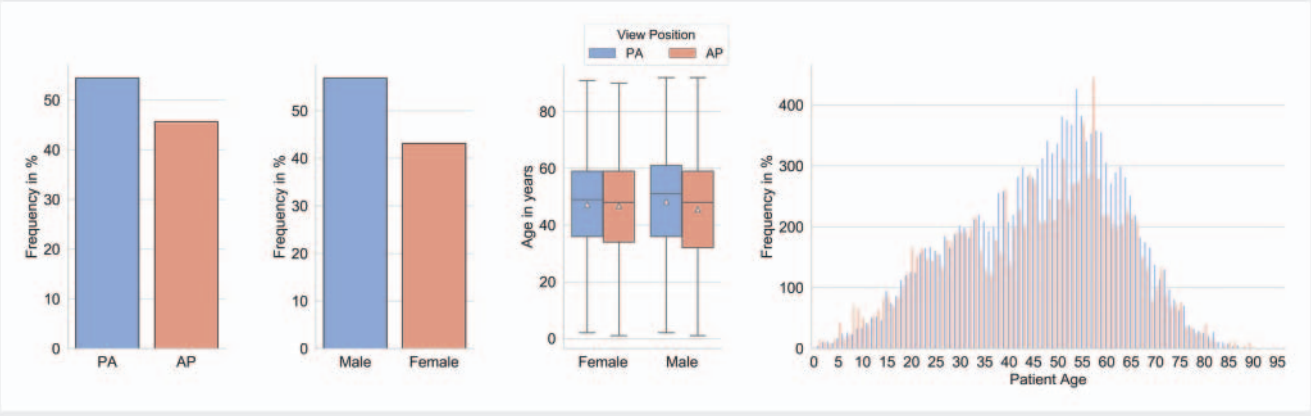
### Data

Within this study we used two different datasets for network training and testing. For the training process we used the "Pneumonia Detection Challenge" data published by the Radiological Society of North America (RSNA) [7]. The dataset contains 26 684 X-ray images in the size of 1024x1024 pixels including 46 % in the AP class and 54 % in the PA class. All images are grayscale and have a value quantization of 8 bit. Some of these images include digital markers to indicate which X-ray position was used. A statistical description of the RSNA dataset in regard to patient age, gender, and view position is visualized in ▶ **Fig. 2**.

In addition to the presented training dataset above a self-compiled dataset was used to test the performance of the generated models on independent data. This dataset is based on the picture archiving and communication system (PACS) archive of the University Hospital Essen (in-house data). The data was compiled using the procedure codes "KTH" and "KTHL", which represents the German in-house equivalent for PA and AP chest X-ray images. From both view types 3000 of the newest X-rays within the PACS were selected. Further selection criteria were that the necessary digital imaging and communications in medicine (DICOM) tags like the procedure code, series description, photometric interpre-

▶ **Fig. 1** Visualization of exemplary radiographs from both view positions. The first row contains PA and the second AP radiographs.

▶ **Abb. 1** Visualisierung von beispielhaften Röntgen-Thoraxaufnahmen beider Aufnahmepositionen. Die erste Reihe beinhaltet PA- und die zweite AP-Aufnahmen.




▶ **Fig. 2** Shows the distribution of the view position, the patients' gender distribution and the patients' age distribution based on the view position in the RSNA data.

▶ **Abb. 2** Zeigt die Verteilung der Aufnahmeposition, des Geschlechts und die Altersverteilung basierend auf der Aufnahmeposition für die RSNA-Daten.

tation, bits stored, patient age and patient sex contained valid information. Within the PA class we filtered all lateral X-rays based on the view position code (LL) or the study description. In total, this leads to a collected dataset of 4507 X-rays including 45 % in the AP and 55 % in the PA class. Similarly to ▶ **Fig. 2**, the same statistics were computed for the in-house dataset, which are visualized in ▶ **Fig. 3**.

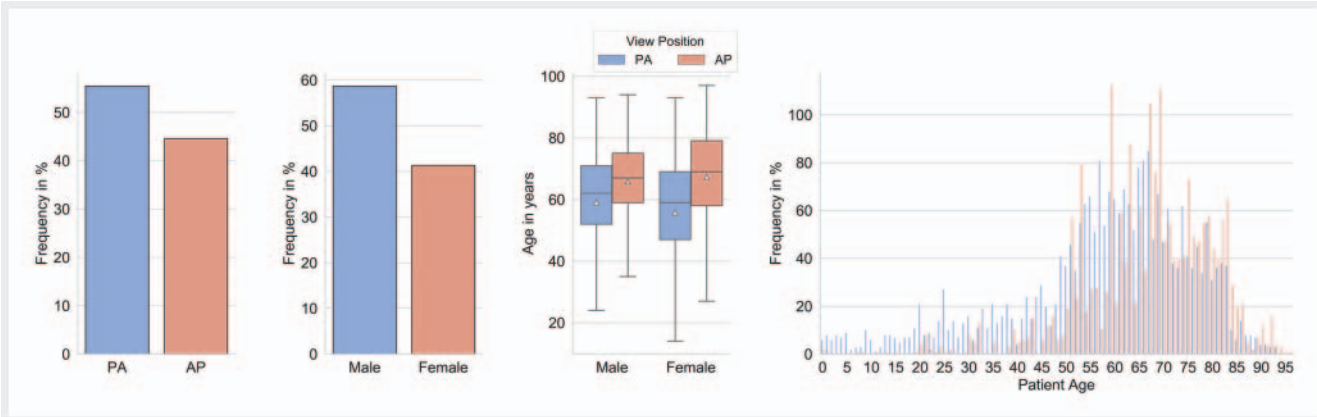
Furthermore, all images within the test dataset are in grayscale and contain a value quantization between 12 and 16 bit.

## Methodology

All X-rays were classified according to the given procedure code within the DICOM header which represent the following ground truth labels: Class 1 for PA and class 2 for AP. Since this task is defined as an image classification problem, suitable architectures with promising results regarding the image classification domain such as the VGG and the ResNet have been applied [8, 9]. The VGG-like architecture was modified in order to ensure a competitive receptive field in comparison to the utilized ResNet-34, as visualized in ▶ **Fig. 4**.

Both architectures use a repeating sequence of convolutional, instance normalization [10] and ReLU [11] layers in their building blocks (see ▶ **Fig. 4** bottom) [8, 9]. After a defined number of convolutional blocks, the VGG uses max-pooling layers for the further feature selection. Within our implementation of the VGG, the competitive receptive field was ensured through an additional max-pooling layer followed by three convolutional layers (512). At the same time we removed the fully connected layers and replaced them with a global average pooling layer followed by a 1x1 convolution. For the ResNet implementation only the batch normalization layers were replaced by instance normalization layers. Otherwise it starts with a 7 × 7 convolution followed by instance normalization and a max pooling layer to reduce the spatial dimensions of the input image [8, 9]. Subsequently, the data flows through a repeating number of residual blocks which perform the identity mapping (purple blocks in ▶ **Fig. 4** bottom) and a down-sampling through a selected stride of 2 (orange blocks in ▶ **Fig. 4** bottom).

The general difference between these architectures is the way the information flows through the network and thus the subse-




▶ **Fig. 3** Shows the distribution of the view position, the patients' gender distribution and the patients' age distribution based on the view position within the in-house data.

▶ **Abb. 3** Zeigt die Verteilung der Aufnahmeposition, des Geschlechtes und die Altersverteilung basierend auf der Aufnahmeposition für die hauseigenen Daten.

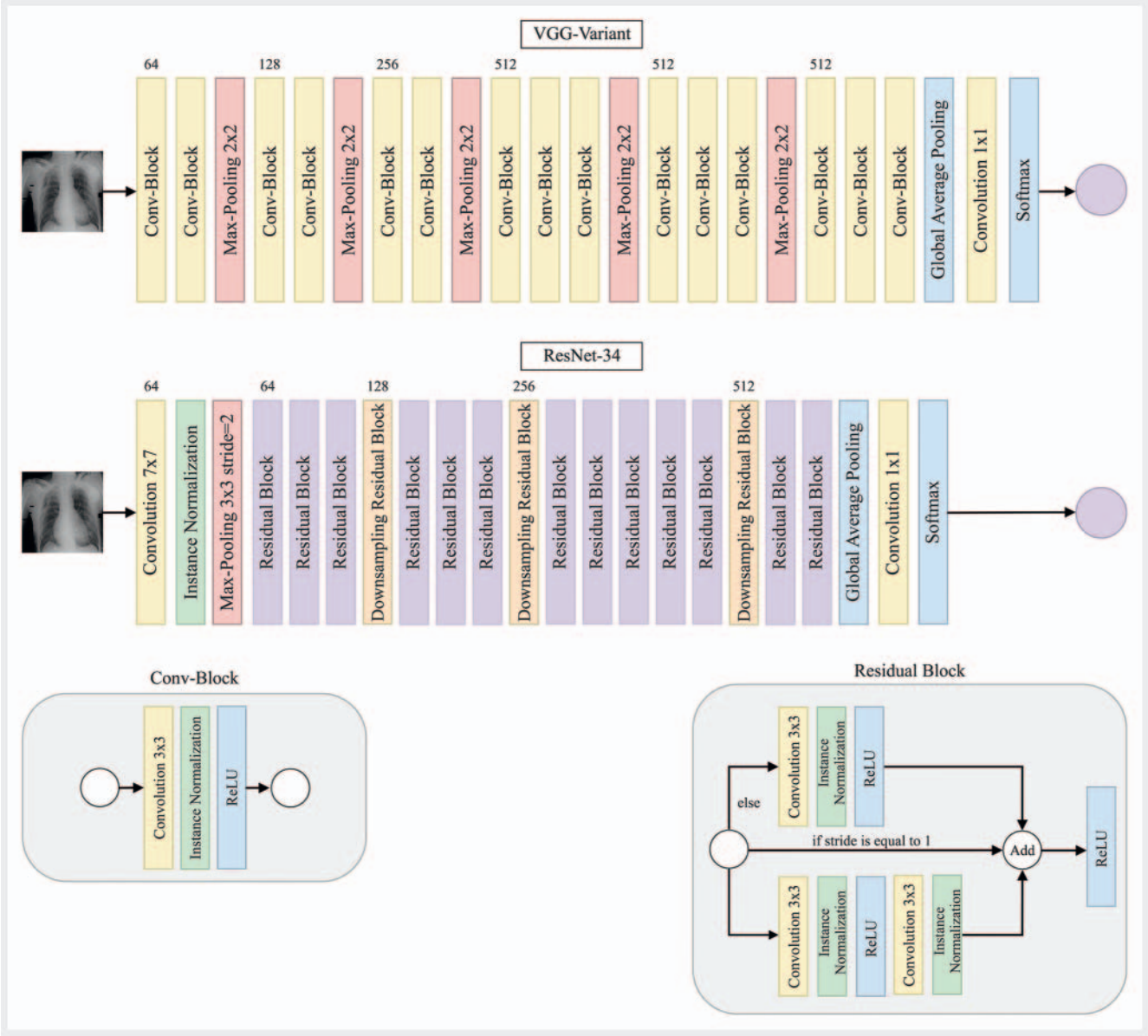
quent error feedback. The VGG follows linear information flow which harbors the danger of 'vanishing gradient' meaning that the gradients can become so small that they cause stagnation in the network's optimization process [9]. To counteract this problem, the skip connections were introduced within the ResNet architecture. These make it possible to merge information for a single block – on the one hand from the output of a residual block and on the other hand from the input of the previous block. These connections, unlike the VGG, allow a different kind of error tracing since they allow propagation of the error through the network using less layers [8].

In this study, the global average pooling layers were applied in order to reduce overfitting and at the same time to enable Grad-CAM visualization [8, 9, 12, 13]. The Grad-CAM uses the gradient information of the last convolutional layer to visualize the relevant regions for a given classification. This helps to create a better understanding as to which anatomic regions on a given chest radiograph are decisive for the algorithm's classification [13]. The implemented visualization additionally allows comparison between the algorithm's and the physician's region of interest, giving interesting insight into the differences and similarities between human and algorithmic assessment of chest X-ray images [5, 13]. The complete processing of an image, from preprocessing to the visualization of the final prediction, is shown in ▶ **Fig. 5**.

Three image sizes were chosen for training: 128 × 128, 256 × 256 and 512 × 512. With the different image sizes, we want to validate if more image information leads to more accurate predictions and this results in three pre-processed datasets. Within these datasets the images were analyzed for the presence of digital markers. Since these are represented by the maximum possible value, it was possible to detect and through dilation merge the relevant pixels into rectangular regions. The affected regions were extracted and replaced by black patches covering the identified pixels (see ▶ **Fig. 5** left). This image preprocessing prevents the networks from using the markers as a possible feature for prediction. The removal enables the opportunity to use the trained networks on radiographs with or without markers. Next, the cleaned images were normalized to the value range –1 to 1 as network inputs. Additionally, the images were randomly augmented by zooming (25 % in and out), horizontal flipping and cropping (87.5 % of the image size) to ensure that the networks generalize better by using slightly modified images and virtually increase the dataset size (see ▶ **Fig. 5** middle) [14]. Subsequently, the pre-processed images were fed into the network which resulted in two outputs. On the one hand a probability vector, which indicates the class for a given radiograph and on the other hand a Grad-CAM which can be used as an overlay for the input image to visualize the relevant regions for the classification (see ▶ **Fig. 5** output).

The following hyperparameters were used for both networks. A batch size of 16 was chosen and all models were trained for 30 epochs. For the optimization the Adam-optimizer was used with the default parameters (lr = 0.001, beta1 = 0.9, beta2 = 0.999, epsilon = 1e-07) [15]. In addition, a learning rate decay (every 10 epochs) was applied. As a loss function, the cross-entropy on softmax activations was used [16]. For regularization purposes, we used dropout with a probability of 0.5 between the global average pooling layer and the final 1x1 convolution. All models were trained using five-fold cross-validation with each fold containing 5336 to 5337 images, which results in 1335 optimization steps per epoch. All stand-alone models were then combined into an ensemble model merging the individual predictions into a single prediction based on the averaged probabilities from the softmax activation. In addition, a human reader labeled the in-house data by hand which enables us to compare the network results to those automatically derived as well as the human reader labels for further evaluation. In addition, the in-house dataset size was reduced by 13 since the human reader excluded those X-rays due to bad quality.

For the evaluation the following metrics have been chosen. First, the following four values were calculated: true positive (TP), true negative (TN), false positive (FP) and false negative




▶ **Fig. 4** Structural representation of the network architectures (top = VGG variant, bottom = ResNet-34) and their building blocks.

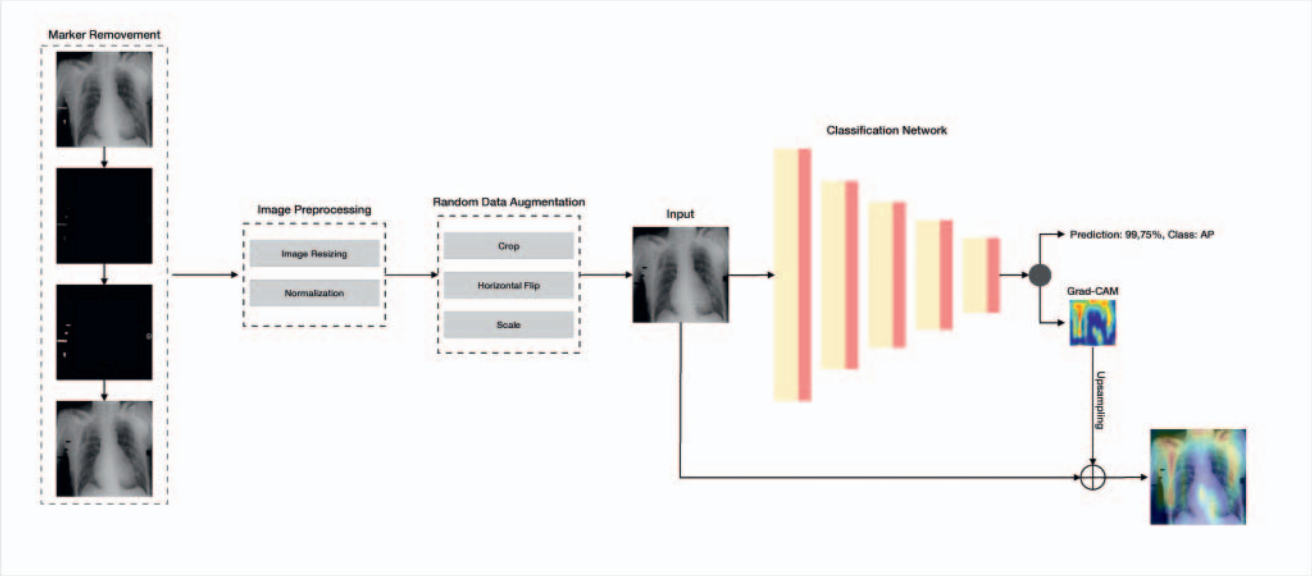
▶ **Abb. 4** Strukturelle Übersicht der verwendeten Netzwerkarchitekturen (oben = VGG-Variante, unten = ResNet) und der dazugehörigen Bausteine.

(FN). These represent the extent to which a model has learned the ability to classify a given sample into the correct class (TP and TN) or into the false class (FP and FN) [17]. These values build the foundation to calculate the accuracy and F1-score which represent the performance of a trained classifier [17]. Additionally, the area under the curve (AUC) was used to evaluate the likelihood that a given example is classified in the correct class which means that a higher AUC indicates a better classifier regardless of the prediction threshold [17].

## Results

▶ **Table 1** visualizes the averaged cross validation scores from the RSNA validation splits as well as the test scores from the in-house dataset for all stand-alone models including the standard deviation.

The results of the single models show that all models reach near perfect results on the validation splits of the RSNA data. In comparison, the results of the compiled in-house data show that all models drop about 3–4 % in terms of accuracy and the F1-score. In addition, the models tend to have more problems with the classification of an AP than with a PA example which is indicated through the higher number of FN. Besides the single model scores, ▶ **Table 2** visualizes the results of the ensemble models for each model configuration.




▶ **Fig. 5** Visualization of the pre-processing pipeline for the image data and the network prediction.

▶ **Abb. 5** Visualisierung der verwendeten Vorverarbeitungschritte für die Bilddaten und die resultierende Netzwerkvorhersage.

The presented results in ▶ **Table 2** show that all model configurations could improve their performance in all given metrics and scores through the ensemble approach. The accuracy and F1-score improved about 1 % and also the number of FP and FN could be reduced. Based on the ensemble performance in ▶ **Table 2**, we compared the DICOM labels of the in-house dataset with the ones from the human reader. This comparison results in 175 detected divergent labels between the in-house DICOM and the human reader labels. We then used the ResNet (512 × 512) model outputs as well as the DICOM labels as a prediction and hold them against the human reader labels to see which approach delivers better results (see ▶ **Table 3**).

The results in ▶ **Table 3** indicate that the model and the human reader have more common decisions on the samples within the in-house dataset than the human reader and the DICOM labels resulting in a slightly (about 1 % to 1.5 %) better performance using the model predictions. Besides the numeric evaluation and interpretation, ▶ **Fig. 6** shows the Grad-CAM visualizations for both classes including an averaged heatmap for each view position and model within the ensemble.

The provided Grad-CAMs indicate that the network used important anatomical reference points such as the scapula, the heart, the neck or ribs for the differentiation of the view positions. Based on the visualized examples, we generated an averaged heatmap for both view positions to provide a complete overview of which anatomical parts were often used by the network when declaring a decision. Those averaged Grad-CAMs show that the networks learned and used heterogeneous features for their decisions.

## Discussion

The goal of the present study was to design and train a convolutional neural network (CNN) to derive the correct view position of chest X-rays from the imaging data itself and thus be able to correct erroneously entered metadata. The results for the F1-score show that all networks are capable of a generalized distinction between both view positions. In addition, the networks not only learned important features, but also used those reference points on which radiologists would base their decisions, like the scapula, heart, ribs, collarbone and neck. However, it should be mentioned that there were slight differences between the trained models in terms of performance. No model configuration could reproduce the cross-validation scores from the RSNA dataset. Furthermore, in the in-house dataset all scores drop between 2 % and 3 % in the ensemble models in comparison to the cross-validation scores. After comparing the human reader labels against the model predictions as well as the DICOM labels, it can be stated that the models reach higher agreement with the human reader than the DICOM labels which is expressed through fewer labeling errors.

In general, our study shows that deep learning can be an option for automatic monitoring and, if necessary, correction of incorrectly entered metadata in the radiological workflow. In this way, deep learning can be used to prevent accounting errors, poor quality research data or even incorrect diagnoses and treatments.

In relation to this study, Rubin et al. showed that the view position is decisive for deep learning-based disease detection on X-ray images [18]. Parallel to our work, Kim et al. [19] published a study on this topic. The experimental setup and the results are similar to our study. Comparing the results, it becomes clear that both studies achieve approximately the same accura-

▶ **Table 1** Results of the single models with mean and standard deviation for the cross validation splits from the RSNA (top) and the test results on the in-house dataset (bottom).

▶ **Tab. 1** Resultate der einzelnen Modelle mit Mittelwert und Standardabweichung für die Kreuz-Validierungs-Splits der RSNA-Daten (oben) und der Testergebnisse auf den hauseigenen Daten (unten).

| Data | Network | Image Size | TP | FP | TN | FN | Accuracy (%) | $F_1$-Score (%) | AUC |
|---|---|---|---|---|---|---|---|---|---|
| RSNA (CV) | VGG | 128 | 2413 ± 39 | 16 ± 5 | 2885 ± 32 | 21 ± 3 | 99.3 ± 0.2 | 99.2 ± 0.2 | 0.9972 ± 0.0003 |
| | | 256 | 2412 ± 42 | 13 ± 3 | 2888 ± 40 | 22 ± 5 | 99.3 ± 0.0 | 99.3 ± 0.1 | 0.9982 ± 0.0006 |
| | | 512 | 2416 ± 39 | 14 ± 3 | 2888 ± 36 | 18 ± 3 | 99.4 ± 0.1 | 99.3 ± 0.1 | 0.9981 ± 0.0005 |
| | ResNet | 128 | 2414 ± 35 | 18 ± 4 | 2884 ± 34 | 20 ± 8 | 99.3 ± 0.1 | 99.2 ± 0.1 | 0.9980 ± 0.0009 |
| | | 256 | 2414 ± 39 | 11 ± 2 | 2890 ± 38 | 20 ± 4 | 99.4 ± 0.0 | 99.3 ± 0.0 | 0.9982 ± 0.0006 |
| | | 512 | 2416 ± 34 | 14 ± 5 | 2888 ± 33 | 18 ± 7 | 99.4 ± 0.1 | 99.3 ± 0.1 | 0.9982 ± 0.0008 |
| in-house | VGG | 128 | 2041 ± 24 | 26 ± 3 | 2289 ± 3 | 238 ± 24 | 96.4 ± 0.5 | 96.1 ± 0.7 | 0.9925 ± 0.0018 |
| | | 256 | 2031 ± 31 | 27 ± 5 | 2288 ± 5 | 148 ± 31 | 96.1 ± 0.6 | 95.9 ± 0.7 | 0.9925 ± 0.0007 |
| | | 512 | 1996 ± 46 | 24 ± 4 | 2291 ± 4 | 183 ± 46 | 95.4 ± 0.9 | 95.1 ± 1.1 | 0.9931 ± 0.0010 |
| | ResNet | 128 | 1956 ± 86 | 21 ± 4 | 2294 ± 4 | 223 ± 86 | 94.6 ± 1.8 | 94.1 ± 2.0 | 0.9924 ± 0.0007 |
| | | 256 | 2006 ± 65 | 20 ± 5 | 2295 ± 5 | 173 ± 65 | 95.7 ± 1.4 | 95.4 ± 1.5 | 0.9942 ± 0.0004 |
| | | 512 | 2026 ± 59 | 21 ± 7 | 2294 ± 7 | 159 ± 59 | 96.1 ± 1.2 | 95.8 ± 1.3 | 0.9938 ± 0.0005 |

▶ **Table 2** Test results of all ensemble mode configurations on the in-house dataset.

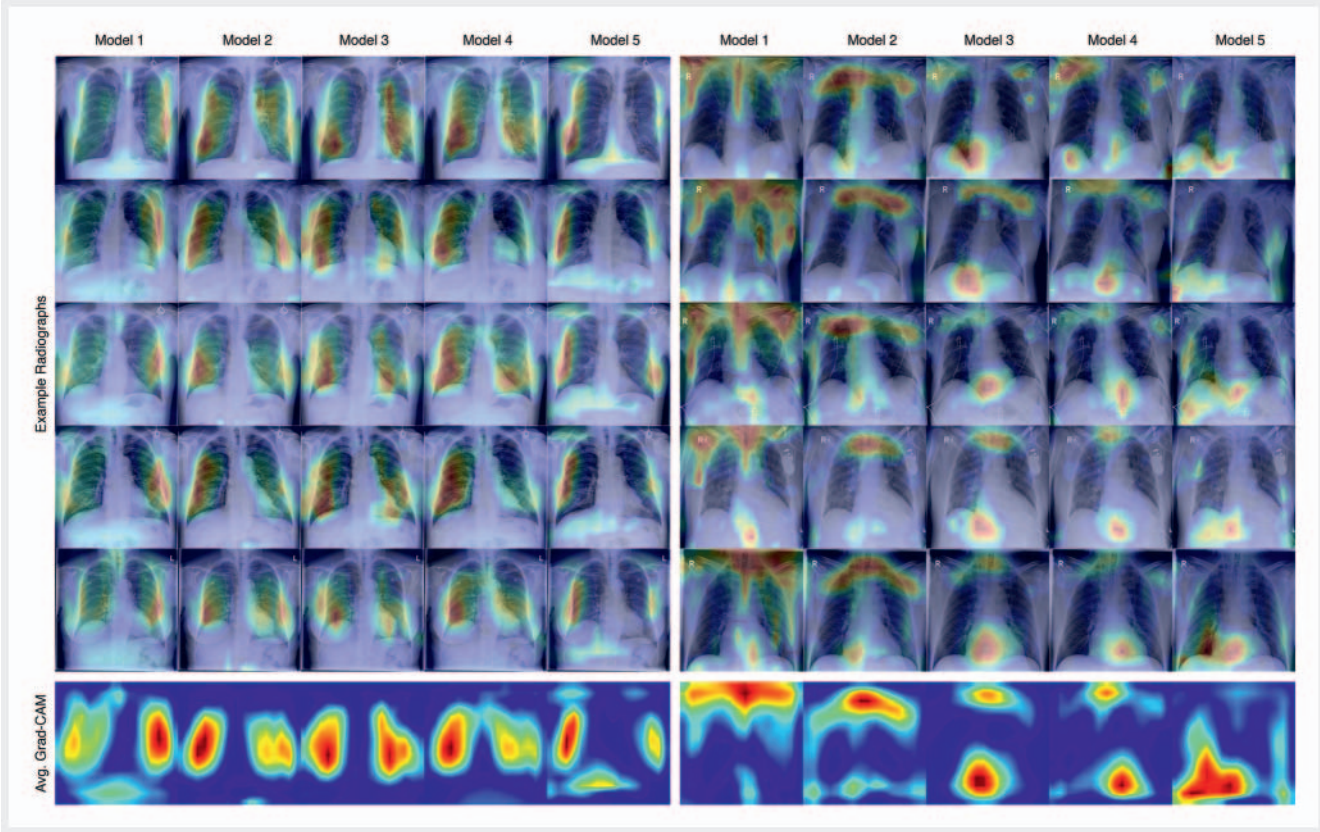
▶ **Tab. 2** Testergebnisse aller Ensemble-Modelle auf den hauseigenen Daten.

| Network | Image Size | TP | FP | TN | FN | Accuracy (%) | $F_1$-Score | AUC |
|---|---|---|---|---|---|---|---|---|
| VGG | 128 | 2070 | 23 | 2292 | 109 | 97.1 | 96.9 | 0.9949 |
| | 256 | 2062 | 25 | 2290 | 117 | 96.8 | 96.7 | 0.9936 |
| | 512 | 2040 | 21 | 2294 | 139 | 94.4 | 96.2 | 0.9947 |
| ResNet | 128 | 2015 | 20 | 2295 | 164 | 95.9 | 95.6 | 0.9937 |
| | 256 | 2036 | 18 | 2297 | 143 | 96.4 | 96.2 | 0.9954 |
| | 512 | 2062 | 20 | 2295 | 117 | 97.0 | 96.8 | 0.9945 |

▶ **Table 3** Comparison of the 512 × 512 ensemble ResNet results with the in-house procedure code and human reader labels.

▶ **Tab. 3** Vergleich der ResNet (512 × 512) -Ensemble-Ergebnisse mit dem hauseigenen Procedure Code und den menschlichen Labeln.

| | TP | FP | TN | FN | Accuracy (%) | $F_1$-score |
|---|---|---|---|---|---|---|
| Human Reader vs. Procedure Codes | 2005 | 1 | 2314 | 174 | 96.1 | 95.8 |
| Human Reader vs. ResNet (512 × 512) | 2062 | 20 | 2295 | 117 | 96.9 | 96.7 |




▶ **Fig. 6** Visualization of the generated Grad-CAMs for each model (columns) from the ResNet (512 × 512) for the PA (left) and AP (right) view position. The Grad-CAMs indicate that the networks use different reference points for the distinction between both view positions like the scapula, the heart, ribs, and the neck. The bottom row shows the averaged Grad-CAMs for each model within the ensemble.

▶ **Abb. 6** Visualisierung der generierten Grad-CAMs für jedes Modell (Spalte) innerhalb des ResNet (512 × 512) -Ensembles für die PA- (links) und AP (rechts) -Aufnahmeposition. Die Grad-CAMs zeigen, dass das Netzwerk unterschiedliche Referenzpunkte für die Unterscheidung der Aufnahmepositionen wie die Skapula, das Herz, die Rippen oder den Nacken nutzt. Die letzte Reihe zeigt die durchschnittlichen Grad-CAMs für jedes Modell innerhalb des Ensembles.

cies within the training data. This is not surprising since the RSNA dataset used in this study is based on the NIH dataset [7]. Also both studies reach high accuracy and AUC rates on self-compiled test data. One main difference is that the CNNs created in our work are not only validated on labeled images, but also through manual examination from a human reader. Those results show that the trained networks are capable of detecting labeling errors within the data storage with high accuracy. More differences are that we used a 4 times larger external test dataset and evaluated not only single models but also ensemble models. Also within our preprocessing pipeline, all digital markers within a radiograph were removed, hence the potential usage as a feature. All models were trained from scratch without using pretrained models from the natural imaging domain. This enables the usage of grayscale image inputs instead of artificially created RGB images. Overall, both studies show and prove the potential of deep learning for the validation of meta information within the clinical routine [19].

In addition to the results and related studies, the limitations of our study must also be considered. First, the use of an external dataset for the training of the networks can be regarded as a limitation. For further studies it would be useful to compile a training dataset completely from in-house data in order to better control the data quality itself and also the accessible meta information. In addition, it can be stated that the training and test quantity was sufficient, but for further and better generalization an increase in both should be considered. Another limitation of our study is due to the fact that only the PA/AP view was considered. From a clinical point of view, the PA/AP view can be complemented by the lateral view position. Based on the training dataset, this distinction was not possible and it would be a useful addition to distinguish not only between AP and PA but also between PA and lateral view to provide full support for all view positions in the clinical routine.

In summary, our study shows that it is possible to extract the AP/PA view position of a chest X-ray from the image data using deep learning and thus correct incorrectly entered metadata.

## CLINICAL RELEVANCE

- It is known that a certain percentage of manually entered meta information from radiological examinations can be incorrect.
- The manual monitoring and, if necessary, correction of such metadata would be very time-consuming and thus not practicable.
- An automatic correction of such metadata by deep learning-based software would be a cost-effective way to reduce billing errors, poor quality of research data or even wrong diagnoses and treatments.

## Conflict of Interest

The authors declare that they have no conflict of interest.

## References

[1] Rajpurkar P, Irvin J, Zhu K et al. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. CoRR 2017; abs/1711.05225. https://stanfordmlgroup.github.io/projects/chexnet/

[2] Liu C, Cao Y, Alcantara M et al. TX-CNN: Detecting tuberculosis in chest X-ray images using convolutional neural network. In: 2017 IEEE International Conference on Image Processing (ICIP) 2017: 2314–2318

[3] 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

[4] Liu X, Faes L, Kale AU et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. Lancet Digit Health 2019; 1: e271–e297

[5] Raoof S, Feigin D, Sung A et al. Interpretation of plain chest roentgenogram. Chest 2012; 141: 545–558

[6] Bruno MA, Walker EA, Abujudeh HH. Understanding and confronting our mistakes: the epidemiology of error in radiology and strategies for error reduction. Radiographics 2015; 35: 1668–1676

[7] 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: Proceedings of the IEEE conference on computer vision and pattern recognition 2017: 2097–2106

[8] He K, Zhang X, Ren S et al. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition 2016: 770–778

[9] Simonyan K, Zisserman A. Very Deep Convolutional Networks for Large-Scale Image Recognition. In: Bengio Y, LeCun Y, (Hrsg.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015. Conference Track Proceedings; 2015

[10] Ulyanov D, Vedaldi A, Lempitsky V. Improved Texture Networks: Maximizing Quality and Diversity in Feed-Forward Stylization and Texture Synthesis. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2017: 4105–4113

[11] He K, Zhang X, Ren S et al. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE international conference on computer vision 2015: 1026–1034

[12] Li K, Wu Z, Peng KC et al. Tell me where to look: Guided attention inference network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2018: 9215–9223

[13] Selvaraju RR, Cogswell M, Das A et al. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In: 2017 IEEE International Conference on Computer Vision (ICCV) 2017: 618–626

[14] Mikolajczyk A, Grochowski M. Data augmentation for improving deep learning in image classification problem. In: 2018 international interdisciplinary PhD workshop (IIPhDW) IEEE. 2018: 117–122

[15] Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. In: Bengio Y, LeCun Y, (Hrsg.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015. Conference Track Proceedings; 2015

[16] Zhang Z, Sabuncu M. Generalized cross entropy loss for training deep neural networks with noisy labels. In: Advances in neural information processing systems 2018: 8778–8788

[17] Powers D. Ailab. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness & correlation. J Mach Learn Technol 2011; 2: 2229–3981

[18] Rubin J, Sanghavi D, Zhao C et al. Lage Scale Automated Reading of Frontal and Lateral Chest X-Rays using Dual Convolutional Neural Networks. In: Conference on Machine Intelligence in Medical Imaging (CMIMI) San Francisco, California, 2018

[19] Kim TK, Paul HY, Wei J et al. Deep Learning Method for Automated Classification of Anteroposterior and Posteroanterior Chest Radiographs. J Digit Imaging 2019: 1–6