Abstract: Radiographs are a versatile diagnostic tool for the detection and assessment of pathologies, for treatment planning or for navigation and localization purposes in clinical interventions. However, their interpretation and assessment by radiologists can be tedious and error-prone. Thus, a wide variety of deep learning methods have been proposed to support radiologists interpreting radiographs.

Mostly, these approaches rely on convolutional neural networks (CNN) to extract features from images. Especially for the multi-label classification of pathologies on chest radiographs (CXR), CNNs have proven to be well suited. On the Contrary, Vision Transformers (ViTs) have not been applied to this task despite their high classification performance on generic images and interpretable local saliency maps which could add value to clinical interventions. ViTs do not rely on convolutions but on patch-based self-attention and in contrast to CNNs, no prior knowledge of local connectivity is present. While this leads to increased capacity, ViTs typically require an excessive amount of training data which represents a hurdle in the medical domain as high costs are associated with collecting large medical data sets.

In this work, we systematically compare the classification performance of ViTs and CNNs for different data set sizes and evaluate more data-efficient ViT variants (DeiTs). Our results show that while the performance between ViTs and CNNs is on par with a small benefit for ViTs, DeiTs outperform the former if a reasonably large data set is available for training.

Keywords: Deep Learning, Chest Radiograph, Vision Transformer, Convolutional Neural Network, CheXpert

1 Introduction

Chest radiographs (CXR) are commonly used for the identification, assessment and localization of pathologies. CXRs enable a cost- and time-effective examination with low radiation dose and allow clinicians to detect a wide range of diseases, plan treatments and localize specific anatomic structures. Therefore, CXRs are the most performed imaging study with an annually increasing number of examinations [2, 9, 10, 14]. A direct consequence of the increasing amount of CXR examinations is a significantly increased workload for radiologists. Therefore, radiologists need to assess a large amount of CXRs manually in their daily routine which can lead to an increased amount of human-errors [1, 2]. Thus, a well-integrated computer-assisted tool that could give cues to the radiologists on what pathology might be present and where to look, could accelerate clinical workflows and reduce the number of human errors. Furthermore, such systems could be especially helpful for inexperienced radiologists and help to prioritize assessments of CXRs [13]. Various computer-assisted tools, including feature engineering and later statistical models that learn from training data have been proposed in the past for this task. Finally, the publication of large-scale data sets such as CheXpert or MIMIC-CXR [6, 7] paved the way towards human-level classification performance on CXRs with deep-learning based CNNs [6, 8, 13]. Furthermore, CNNs are proposed for a wide variety of tasks such as classification, localization, segmentation or automated report generation and have emerged to be the de-facto standard for the processing of radiographs. However, a recent publication challenges CNNs and proposes Vision Transformers (ViT) that use multi-headed self-attention between image patches instead of convolutions to learn meaningful feature representations from images [4]. Originally, transformer networks have shown strong performance for modeling and interpreting sequence-data like sentences and outperform traditional recurrent neural networks in many sequence-related tasks [12]. Applying the core principles of Transformers to the image domain as it is done in ViTs has shown to outperform plain CNNs for large-scale databases of generic images such as ImageNet.

Beside the potential performance gains, ViTs share the appeal-
ing property of class-level local attention maps \cite{12}. These attention maps could be helpful not only for the classification task in CXRs but also for tasks where localization of anatomical structures is required. However, ViTs do not impose prior knowledge of the local connectivity of image pixels as it is the case with convolutions. Thus, ViTs require an excessive amount of training data and are often only applicable when pre-trained on large-scale data sets \cite{11}. This opens the question, whether ViTs can be leveraged for multi-class classification problems with CXRs. As large-scale data sets are crucial for pre-training ViTs, it is of interest if their performance improvement against CNNs can also hold in an image domain different from ImageNet with smaller, medical data sets available for fine-tuning. In this work, we leverage ViTs for the classification task of pathologies in CXRs and investigate the use of knowledge distillation for data efficiency. In summary, our contribution is three-fold:

- We investigate the use of ViTs for multi-label classification in CXRs and compare their performance to CNNs.
- We study if knowledge distillation with data-efficient Vision Transformers (DeiT) \cite{11} can improve the classification performance.
- We systematically compare the effect of varying training set sizes for CNNs, ViTs and DeiTs, respectively.

## 2 Methods

### 2.1 Data Set

We use the publicly available CheXpert Data set \cite{6}. The data set consists of 224316 CXRs of 65240 patients together with 14 labels, that are automatically generated from radiology reports. There are three types of auto-generated labels. The labels 0 and 1 indicate positive and negative labels, respectively. The third label -1, denotes an uncertain decision. In this work, we treat all uncertain samples as positive samples. Further, we focus on the classification of five different pathologies, namely Atelectasis, Cardiomegaly, Consolidation, Edema and Pleural Effusion. We split our data into a train, validation and test set. 20% of the data are used for evaluation ($D_{\text{train}}$). From the remaining data, we sample 5 data folds, each consisting of 80% training data ($D_{\text{train}}^i$) and 20% validation data ($D_{\text{val}}^i$), where $i$ indicates the fold. To simulate varying training set sizes we sample subsets including (10, 20, ..., 90)% of $D_{\text{train}}^i$ for all folds respectively.

We do not perform a specific pre-processing of the images but resize the images to a resolution of $224 \times 224$px. For data augmentation, we apply random augment \cite{3} and random erasing.

### 2.2 Deep Learning Models

For our experiments, we utilize DenseNets \cite{5} as baseline CNNs as they have proven to be a strong baseline for the classification task on CXRs \cite{13}. In general, CNNs utilize blocks of convolutions, together with a normalization, non-linear activation functions and pooling operations stacked on each other to map an input image to a feature vector. A linear layer maps the feature vector to the output vector which is compared with the class labels. DenseNets add specific skip connections between the convolutional blocks to allow training deep stacks of these blocks \cite{5}. We compare different versions of the baseline CNN, namely DenseNet-121 and DenseNet-201 where the main difference is the depth of the architecture and thus, the number of trainable parameters.

In contrast to CNNs, ViTs do not process image arrays by convolutions. Instead, the image $X \in \mathbb{R}^{H \times W \times C}$ is cropped into $N$ patches $x_p \in \mathbb{R}^{P \times (P \times C)}$, where $H, W$ is the dimension of the image, $C$ is the number of channels and $P$ the resolution of the cropped patches. The patches $x_p$ are flattened and mapped to a fixed dimension $D$ by a linear layer. Additionally, a class token is prepended to the mapping which is later used as input for a classification layer. Furthermore, A 1-dimensional position embedding is added to each patch embedding. The resulting sequence of image patches, class tokens and positional embeddings is used as input to the encoder of the ViT. The encoder includes multiple stacked transformer blocks. Each block consists of a multi-headed self-attention and a multilayer perceptron layer with a normalization layer and skip-connections in between.

We include different versions of ViTs in our experiments, namely ViT-Small (ViT-S) and ViT-Base (ViT-B). The differences between the versions are the number of encoder layers, the dimension of the embeddings $D$, the MLP configuration and the number of attention heads \cite{12}.

We further include data-efficient Vision Transformers (DeiT) \cite{11} to our study. DeITs share the same overall architecture as ViTs. In addition to the class token, a distillation token is added to the patch embedding. Similar to the class token, the distillation token interacts with the patch embeddings through self-attention in the encoder blocks and is processed by a classification layer to obtain an output vector. It is used in a knowledge distillation framework, where the Kullback-Leibler divergence between the output of a teacher network and the output of the distillation token is added to the loss function together with the loss between the class token and the ground truth. The authors of \cite{11} speculate that by this, the inductive bias of CNNs can be distilled to ViTs, which makes DeITs more data-efficient compared to plain ViTs.

We include the pre-trained versions DeiT-S and DeiT-B in our studies and investigate two use-cases of DeITs. First, we use...
DenseNet and ViT. Comparing DeiT-B and DeiT-B-Dist, similars show superior classification performance compared to significantly fewer parameters. Considering DeiT, both variants are on par with the DenseNet baselines. Notably, DenseNet-Dist, a higher performance gain can be observed compared to S-Dist, a higher performance gain can be observed compared to DeiT-B-Dist, a higher performance gain can be observed compared to

As shown in Table 1, it can be observed that ViT models are on par with the DenseNet baselines. Notably, DenseNet-121 shows competitive performance to ViT-B while requiring significantly fewer parameters. Considering DeiT, both variants show superior classification performance compared to DenseNet and ViT. Comparing Deit-B and Deit-B-Dist, similar classification performance can be observed.

Figure 1 shows that for all models the data set size has a crucial impact on the classification performance. For DeiT-B-Dist, a higher performance gain can be observed compared to DenseNet-201 and ViT-B especially when training with larger training sets. Overall, it can be observed that even for small data set sizes, the transformer-based models show similar performance compared to DenseNets.

To visualize the pixel-wise attention of the networks, saliency maps are provided in Figure 2. For DenseNets, a Grad-Cam approach is used to visualize the attention. For transformer-based models, the self-attention weights are visualized. It is noticeable that both networks attend to meaningful regions in the CXR. While the visualization of the attention map weights of transformers leads to local attention maps, Grad-CAM-based saliency maps rather highlight coarse regions.

### 3 Results

We report the Area under Receiver Operator Curve (AUROC) and the F1 score to evaluate the classification performance. Both metrics are calculated as weighted averages over the five different pathology classes, where each class is weighted by the number of true instances for each label. We report the average performance of the 5-Fold cross-validation together with the standard deviation.

As shown in Table 1, it can be observed that ViT models are on par with the DenseNet baselines. Notably, DenseNet-121 shows competitive performance to ViT-B while requiring significantly fewer parameters. Considering DeiT, both variants show superior classification performance compared to DenseNet and ViT. Comparing Deit-B and Deit-B-Dist, similar classification performance can be observed.

Figure 1 shows that for all models the data set size has a crucial impact on the classification performance. For DeiT-B-Dist, a higher performance gain can be observed compared to DenseNet-201 and ViT-B especially when training with larger training sets. Overall, it can be observed that even for small data set sizes, the transformer-based models show similar performance compared to DenseNets.

To visualize the pixel-wise attention of the networks, saliency maps are provided in Figure 2. For DenseNets, a Grad-Cam approach is used to visualize the attention. For transformer-based models, the self-attention weights are visualized. It is noticeable that both networks attend to meaningful regions in the CXR. While the visualization of the attention map weights of transformers leads to local attention maps, Grad-CAM-based saliency maps rather highlight coarse regions.

### 4 Discussion and Conclusion

Recently, ViTs show performance gains over classical CNNs on generic images from benchmark data sets such as the ImageNet data set. Furthermore, they add appealing properties like directly accessible and local attention maps. However, due to the missing inductive bias and the exceeding number of trainable parameters, training ViTs requires large-scale data sets.

In this work, we investigate if we can utilize ViT models for multi-label classification on CXR images and compare their performance to a baseline CNN. We investigate the effect of different data set sizes and explore if knowledge distillation can make the training more data-efficient.
Our results indicate, that the amount of available training data might not be sufficient to reveal the true power of ViT models. We assume that for ViTs, increasing performance will occur at even larger data sets that are not included in this study. In contrast to that, the more data-efficient DeiT model shows increasing performance already for smaller training sets. While we can conclude that the distillation process makes the training more data-efficient, it is hard to verify if the data efficiency is achieved by mimicking the inductive bias of the teacher CNN [11]. Furthermore, even though the required amount of labelled training data is reduced, still, large data sets are required to achieve performance improvements over CNNs with transformer-based networks. However, regularizing the training by knowledge distillation shows to be beneficial and can help to efficiently train transformer-based models. Besides the improved performance of the transformer-based models, they show local and dense saliency patterns. This observation indicates that the attention maps of transformers can be helpful for the localization of lung diseases from CXRs and have the potential to guide treatment planning. Overall, we show that self-attention-based ViT models can be valuable alternatives for multi-label pathology classification, especially in combination with knowledge distillation. Our results motivate the research on combinations of CNNs that enforce local connectivity priors and highly expressive ViTs with global attention. This could be a promising direction, especially for the application of ViTs in the medical domain, where annotated data sets are typically small.

Author Statement

Research funding: This work was partially funded by Grant Number KK5208101KS0.

Conflict of interest: Authors state no conflict of interest.