IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received May 5, 2020, accepted May 10, 2020, date of publication May 19, 2020, date of current version June 2, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2995567

Cardiomegaly Detection on Chest Radiographs: Segmentation Versus Classification

ECEM SOGANCIOGLU, KEELIN MURPHY, ERDI CALLI, ERNST T. SCHOLTEN, STEVEN SCHALEKAMP, AND BRAM VAN GINNEKEN
Diagnostic Image Analysis Group, Department of Radiology and Nuclear Medicine, Radboud University Medical Center, 6525 Nijmegen, The Netherlands

Corresponding author: Ecem Sogancioglu (ecem.sogancioglu@radboudumc.nl)

This work was supported by the Dutch Technology Foundation STW, which formed the NWO Domain Applied and Engineering Sciences and partly funded by the Ministry of Economic Affairs (Perspectief Programme DLMedIA: Deep Learning for Medical Image Analysis) under Grant P15-26.

ABSTRACT In this study, we investigate the detection of cardiomegaly on frontal chest radiographs through two alternative deep-learning approaches - via anatomical segmentation and via image-level classification. We used the publicly available ChestX-ray14 dataset, and obtained heart and lung segmentation annotations for 778 chest radiographs for the development of the segmentation-based approach. The classification-based method was trained with 65k standard chest radiographs with image-level labels. For both approaches, the best models were found through hyperparameter searches where architectural, learning, and regularization related parameters were optimized systematically. The resulting models were tested on a set of 367 held-out images for which cardiomegaly annotations were hand-labeled by two independent expert radiologists. Sensitivity, specificity, positive predictive value, negative predictive value, and area under the receiver operating characteristic curve (AUC) were calculated. The performance of the segmentation-based system with an AUC of 0.977 is significantly better for classifying cardiomegaly than the classification-based model which achieved an AUC of 0.941. Only the segmentation-based model achieved comparable performance to an independent expert reader (AUC of 0.978). We conclude that the segmentation-based model requires 100 times fewer annotated chest radiographs to achieve a substantially better performance, while also producing more interpretable results.

INDEX TERMS Deep learning, chest radiographs, anatomy segmentations, cardiomegaly.

I. INTRODUCTION

Recent literature on the automatic interpretation of chest X-ray (CXR) images has been dominated by methods which learn to predict labels indicating the presence or absence of a specific abnormality in the CXR [2], [22], [30]. Such labels are frequently referred to as 'image-level' labels since they refer to the image as a whole and provide no more specific information, for example, regarding the location or severity of the abnormality. The popularity of this method of analysis is likely related to the recent release of numerous large public datasets, each of which provides multiple image-level labels for a variety of abnormalities [5], [22], [23], [38]. However, image-level labels may not be the optimal way to learn to recognise specific abnormalities. Since these labels provide no information on the shape or location of the abnormality, it is likely that a very large number of labelled samples will be needed to train a supervised-learning system. Furthermore, the trained system provides no insight or intuition into how it infers labels. Such a 'black-box' system is more difficult to trust and less likely to find acceptance in a clinical setting.

In this work, we investigate how a more intuitive and interpretable segmentation-based method to detect abnormality compares with the state of the art in deep-learning using image-level labels. The abnormality investigated in this case is cardiomegaly, one of the most frequently mentioned findings in radiology reports for chest radiography exams. Cardiomegaly refers to an enlargement of the heart and can be used as a marker for heart disease [14], [26]. Due to its wide availability, high cost-effectiveness, and low radiation dose, chest X-rays are often the first imaging study acquired and can be utilized as a fast screening tool for cardiomegaly. In order to detect this condition, radiologists examine the cardiac silhouette and calculate the cardiothoracic ratio (CTR),

The associate editor coordinating the review of this manuscript and approving it for publication was Orazio Gambino.

a commonly used radiographic index measured as the ratio of maximum horizontal cardiac diameter to the maximum horizontal thoracic diameter [15] (Figure 1). A CTR greater than 0.5 is the generally accepted threshold considered to indicate an enlarged cardiac silhouette, referred to as cardiomegaly.

A vast number of studies have addressed the cardiomegaly detection task along with other abnormalities in a multi-label classification scenario [2], [18], [30], [39], [40], predicting all available labels from the datasets used. Many of these works use the ChestX-ray14 dataset [38] which was released by the National Institutes of Health in 2017 with 112,120 CXRs, each labelled with binary labels for 14 different abnormalities. The labels are automatically extracted from the text analysis of radiology reports. These studies employed widely used state-of-the-art classification architectures, and applied slightly different augmentation and preprocessing techniques to tackle the classification problem. In particular, Baltruschat *et al.* [2] investigated the performance of different network architectures, namely ResNet-38, ResNet-50, and ResNet-101, for classification of 14 abnormalities on the ChestX-ray14 dataset [38]. They achieved a similar level of performance as other recently published studies [18], [40], but all these studies were limited due to their evaluation on the noisy held-out evaluation set where the labels were extracted from radiology reports using natural language processing [28]. In order to address this, Rajpurkar *et al.* [30] annotated a held-out evaluation set from ChestX-ray14 with the majority vote of 3 radiologists (not publicly available), and employed a 121-layer DenseNet architecture. The images were resized to 512 x 512 and normalized with the mean and standard deviation of images in the ImageNet training set before being fed into the network. They reported state-of-the-art results where the proposed algorithm achieved radiologist-level performance on 11 abnormalities in their held-out evaluation set, however, performed significantly worse than the radiologists for 3 abnormalities, one of which was cardiomegaly.

Some earlier works attempted to detect cardiomegaly through segmentation-based solutions via measuring CTR. Ginneken *et al.* [37] investigated the performance of three supervised segmentation methods for anatomical segmentations, namely active shape models, pixel classification, and active appearance models. They showed that both active shape models and active appearance models reached a mean absolute error of 0.012 for cardiothoracic ratio measurement on their 247 held-out set. Candemir *et al.* [7] proposed a graph-cut lung field segmentation method which was then adapted to localize the heart region using heart models in order to measure the CTR. They reported 0.77 sensitivity and 0.76 specificity for the detection of cardiomegaly on 500 held-out evaluation images. Similarly, Dallal *et al.* [13] proposed a method that employed the same lung segmentation method proposed by Candemir *et al.* [7] and using the Harris operator to detect the heart boundaries from the resulting lung field segmentation in order to measure the CTR. They reported a root mean squared error of 0.06 on their 103 held-out images. Recent work by Li *et al.* [25] used a deep learning system for heart and lung field segmentation and showed improved performance for detection of cardiomegaly achieving a sensitivity of 0.97 and specificity of 0.92 on their 500 held-out set.

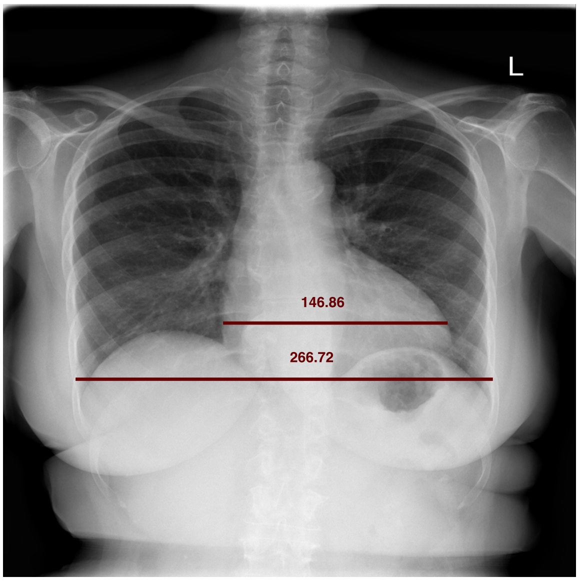


FIGURE 1. Measurement of the cardiothoracic ratio in chest radiographs. Maximum horizontal thoracic diameter = 266.72 (in mm), maximum horizontal cardiac diameter = 146.86 (in mm), CTR = 0.55 (146.86/266.72). CTR > 0.5 and therefore this is a case of cardiomegaly.

This study is the first to directly compare segmentation-based and classification-based solutions for cardiomegaly detection. We implement state-of-the-art deep learning methods for heart and lung segmentation, through which we calculate CTR directly, and also for image-level classification of cardiomegaly. Hyperparameter optimization is applied in all cases to ensure the best possible solution is obtained. We investigate the performance differences between the segmentation-based and classification-based systems for cardiomegaly detection, and the effect of varying the training-set size in each case.

II. DATA

The data used in this study was retrospectively obtained from the publicly available ChestX-ray14 dataset [38]. It is composed of 112,120 frontal view chest radiographs from 30,805 patients stored as 8-bit grayscale images with dimensions of 1024 × 1024. The dataset was automatically labeled from text reports, indicating the presence or absence of 14 different thoracic abnormalities including cardiomegaly.

Heart enlargement, i.e. cardiomegaly, cannot reliably be assessed on AP view chest radiographs since the distance between the X-ray source and the patient is non-standardized on AP view, which causes a variable magnification of the heart. Hence, we selected only posteroanterior (PA) studies. This resulted in 67,310 PA images of 28,868 patients, 44% male, 41% abnormal.

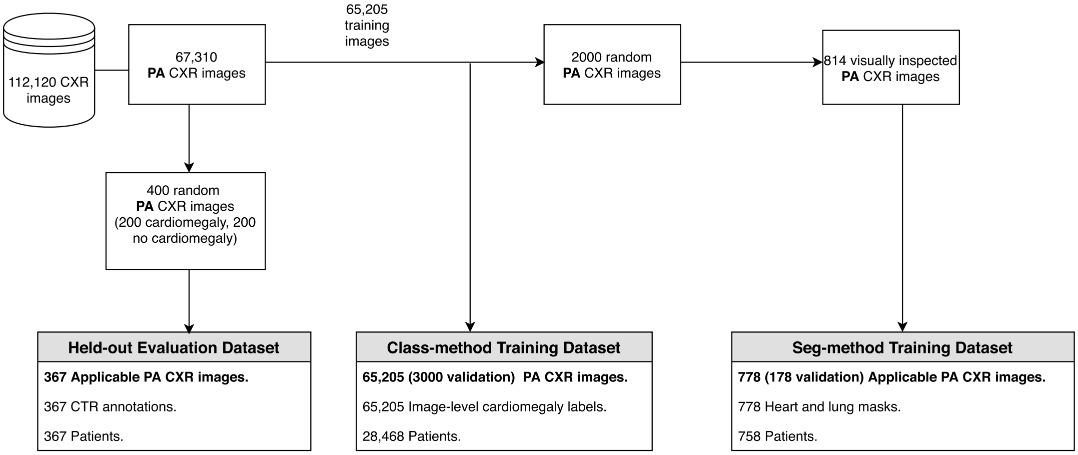

112,120 CXR images
67,310
PA CXR images
65,205
training
images
2000 random
PA CXR images
814 visually inspected
PA CXR images
400 random
PA CXR images
(200 cardiomegaly, 200
no cardiomegaly)
Held-out Evaluation Dataset
367 Applicable PA CXR images.
367 CTR annotations.
367 Patients.
Class-method Training Dataset
65,205 (3000 validation) PA CXR images.
65,205 Image-level cardiomegaly labels.
28,468 Patients.
Seg-method Training Dataset
778 (178 validation) Applicable PA CXR images.
778 Heart and lung masks.
758 Patients.

FIGURE 2. Flowchart of the data selection procedure. CXR = chest x-ray, PA = posteroanterior, CTR = cardiothoracic ratio, class-method = image-level cardiomegaly classification, seg-method = heart and lung segmentation. Images are from the publicly available ChestX-ray14 dataset.

A. HELD-OUT EVALUATION SET

For the final model evaluation, we created a class-balanced set of 400 images (Figure 2). Using the labels provided we randomly sampled 200 cases with cardiomegaly (200/1563) and 200 without cardiomegaly (200/65,747).

A chest radiologist with over 30 years of experience and another chest radiologist with over 5 years of experience independently annotated the maximal horizontal cardiac and thoracic diameters on all evaluation cases. Cases where radiologists could not reliably locate the heart borders were excluded from the study, leaving 367 cases. The annotations of the more experienced radiologist are used as the reference standard throughout this work, while the other radiologist is used as a second reader, for comparison with our automated methods.

B. TRAINING & VALIDATION SET

1) CLASSIFICATION-BASED METHOD

After the selection of only posteroanterior (PA) studies as seen in Figure 2, there was a total of 65,205 chest radiographs from 28,468 patients (excluding the patients in held-out evaluation set). This set was used as our training&validation set (3000 for validation), using the publicly available image-level cardiomegaly labels for training the classification-based method.

2) SEGMENTATION-BASED METHOD

To develop deep neural networks to segment the heart and lungs we first set out to obtain manual segmentations of heart and lung boundaries. In order to select challenging cases for annotation of heart and lung boundaries, we developed a standard U-net [32] architecture which segments the heart and lung area, trained on a separate publicly available dataset, namely JSRT [34]. The JSRT dataset consists of 247 images from scanned films with a resolution of 2048 × 2048 and 12-bit depth. The reference standard for the heart and lung boundaries of those images are provided with the SCR dataset [37]. Our deep learning system was trained on a randomly selected 200 cases (200/247) and the remaining 47 cases were used as the validation set. The images were scaled to a dimension of 256 x 256, and the network was trained with Adam optimizer with a learning rate of 10^{-5}.

Further, a set of 2000 radiographs was randomly selected from the 65,205 remaining images in the ChestX-ray14 dataset (Figure 2). The JSRT-trained system was tested on those cases and visual inspection was used to select 814 cases most of which the algorithm performed sub-optimally. Those 814 cases were presented to a medical student and a computer scientist (with experience analyzing chest radiographs) who were instructed to annotate the heart and lung areas. An experienced radiologist was consulted for difficult cases and cases where the heart boundaries could not be inferred were excluded. This resulted in 778 radiographs (178 for validation) with lung and heart area annotations to be used as the segmentation training & validation set.

III. METHODS

Two approaches for cardiomegaly detection are described in this section: firstly a classification approach based on image level labels (class-method) and secondly the segmentation-based approach (seg-method). For each approach hyperparameter optimization was run for 200 experiments. The final hyperparameters chosen were those that yielded the highest performance on the validation set.

A. CLASSIFICATION-BASED METHOD

To classify cardiomegaly using image-level labels we implemented three state-of-the-art classification architectures, ResNet18, ResNet50 [19], and DenseNet121 [20], which have achieved excellent performance in several computer vision and medical image analysis tasks. Particularly, they were previously shown to achieve high-performance levels on the ChestX-ray14 dataset with multi-label classification settings [2], [30]. Training and architecture related hyperparameters of the class-method were systematically optimized to ensure optimum performance.

All the network architectures were pretrained on ImageNet, and a fully connected layer (2 output units with SoftMax activations) was added after the global average pooling layer. The networks were trained with 65,205 frontal standard chest radiographs (3000 for validation) from ChestX-ray14 dataset, as in Figure 2, using categorical cross-entropy loss. Since there is a class imbalance problem in such a scenario (1156 images with cardiomegaly among 65k), we employed an over-sampling technique [4] by sampling the positive cardiomegaly cases until the dataset was balanced.

All images underwent per sample mean-standard deviation normalization. Data augmentation was applied to the training samples by means of inception-like preprocessing [6], [36]. This consists of applying a random rotation up to 7 degrees, random resizing with a scale in the range [0.7, 1], and random cropping a 4:3 or 3:4 part of the chest X-ray.

1) CLASS-METHOD HYPERPARAMETER OPTIMIZATION

Several aspects of the hyperparameters were optimized for the class-method for 200 experiments.

Due to the very long training time of the class-method (which can take from 2 hours to 23 hours for one experiment depending on the network architecture and other hyperparameters), the hyperopt library [3] was used for 50 experiments. In every experiment during the optimization using hyperopt, the model being optimized is trained from scratch with the candidate hyperparameters for a maximum number of epochs predefined for each model. The selection of the candidate hyperparameters are based on Bayesian optimization, i.e., the hyperparameters were selected based on a trade-off between the results of the previous iterations, the regions of unexplored hyperparameter space, and their underlying distribution.

Further, we also optimized the hyperparameters through grid search, which can be run in parallel unlike hyperopt, for an additional 150 experiments.

The hyperparameters range and the values selected after the optimization can be seen in Table 1. We used three commonly used architectures, DenseNet121, ResNet50, and ResNet18, as a hyperparameter value in order to optimize the network architecture for our problem settings. Due to memory constraints, we made sure that the batch size was set to 8 when the network architecture was DenseNet121 or ResNet50 with an input resolution of 512 otherwise to 16.

Based on the hyperparameter optimization results, after every 100 iterations, the validation loss was calculated on the whole validation set. If the validation loss did not decrease compared to the previous step, the learning rate was reduced by multiplying it with 0.2. The model which showed the least validation error was selected as our final model.

After the hyperparameter optimization, the best model found for the class-method was ResNet50 trained with the largest input resolution of 512. During the experiments, we observed that all the deep learning models were powerful and achieved a high level of performance and that the most crucial hyperparameters on performance were learning-related, i.e. learning rate.

The hyperparameter optimization procedure took around 23 days with hyperopt on a PC equipped with TitanX GPU, and 6 days for grid search optimization (run in parallel) for 150 experiments using several GPU, TitanX, GTX1080, GTX1080ti, GTXTitanx, and TitanV. The code was implemented in Tensorflow [1].

B. SEGMENTATION-BASED METHOD

The segmentation-based approach (seg-method) is designed to address the cardiomegaly detection task on chest radiographs, through segmentation of the heart and lungs and subsequent calculation of the cardiothoracic ratio (CTR). As illustrated in Figure 3, two different models were developed for heart and lung field segmentation respectively. After segmentation, the maximum horizontal cardiac and thoracic diameters were calculated and used to calculate CTR and hence the presence or absence of cardiomegaly based on the clinically used CTR threshold of 0.5.

For the development of heart and lung segmentation models, a U-net-like fully convolutional network architecture [32] was implemented and its training, regularization, and architecture-related hyperparameters were systematically optimized for the best model selection.

The U-net architecture [32] is a state-of-the-art segmentation network, which has achieved promising results on a variety of medical image segmentation tasks [9], [13]. It consists of contracting and expanding paths, where the contracting path is composed of convolution operations decreasing the spatial resolution and the expanding path consists of transposed convolutions increasing the resolution. Further, the details that were lost through downsampling operations are recovered through skip connections which pass feature maps from the contracting to the expanding path.

During training, each model was trained by optimizing the binary cross-entropy loss between the predicted masks and the reference standard (heart or lung masks), which is formulated as follows:

$$\mathrm{BCE} = -\frac{1}{N}\sum_{i=1}^{N} y_i \log \hat{y}_i + (1 - y_i)\log(1 - \hat{y}_i)$$

where N denotes the number of images, y_i represents the reference standard for the sample i, $\hat{y}_i$ represents the model prediction for the sample i.

All images underwent per sample mean-standard deviation normalization. Data augmentation with random rotation, vertical and horizontal shift, zooming, and brightness was applied to improve system robustness. The model was trained for a maximum of 300 epochs, terminating if there was no improvement in the validation set performance for

TABLE 1. Optimized hyperparameters for the class-method. The naming convention follows [10]. LR = learning rate. LR reduced factor indicates the factor to multiply learning rate with in case of no improvement is seen on the validation set performance during training.

	Hyperparameter	Range	Best class-model
Learning	Optimizer	[Adam, SGD, Adagrad, RMSprop]	RMSprop
	Learning rate	{0.00001, 0.1}	0.012
	Initializer	[Orthogonal,glorot_normal,he_normal,lecun_normal]	he_normal
	LR reduced factor	[0.2, 0.9]	0.2
Architecture	Model	[ResNet18, ResNet50, DenseNet121]	ResNet50
	Input resolution	[64, 128, 256, 512]	512

20 successive epochs. We selected the epoch with the best performance on the validation set.

1) SEG-METHOD HYPERPARAMETER OPTIMIZATION

Similar to the class-method, the hyperparameters of the seg-method were optimized using the hyperopt library [3] for 200 experiments.

The heart and lung segmentation models were optimized separately. A specific set of learning, architecture and regularization-related parameters [11], [16], [24], [35] were selected for the hyperparameter search as listed in Table 2 (with the naming convention as in [10]). The learning rate was the only continuous hyperparameter and was sampled from a log uniform distribution. The other hyperparameters were sampled from a discrete uniform distribution between the defined choices.

As a regularization hyperparameter, the selection of dropout (with a probability of 0.5) [35] before each convolution in the expanding path was introduced as a binary hyperparameter. We used batch normalization [21] after every convolution layer as it improved performance by enabling more efficient learning.

Due to the limitations of computational memory, some restrictions on the combinations of hyperparameter settings were required. While a large batch size helps to stabilize the training, the depth of the network and the number of convolution operations per layer increase the capacity of the network, and the receptive field and the higher resolution images allow the network to see more details within the image. However, not all these conditions can be satisfied at the same time due to memory constraints. Therefore, the selection of these hyperparameters was conditioned on each other: when the input resolution was 512, the batch size was chosen as 4, and when the depth of the network was larger than 4, the number of convolution operations per depth was limited to 2 and the number of initial feature maps limited to 32.

The best models found after the hyperparameter optimization for both heart and lung segmentation were U-net architecture with the highest depth 6 as in Table 2. During the experiments, we observed that a larger input resolution yielded better performance.

The hyperparameter optimization procedure took around 13 days for each of the lung and heart segmentation models on a PC equipped with TitanX GPU and with the code implemented in Keras [10] with Tensorflow backend [1].

IV. EXPERIMENTS

The seg-method and class-method performance were investigated for cardiomegaly classification. Further, since the seg-method additionally produces a clinically relevant measure, CTR, the performance of this system was also evaluated in terms of CTR accuracy.

A. CARDIOMEGALY CLASSIFICATION

We evaluate the performance of the two methods and of the second reader by calculating the area under the receiver operating characteristic curve (AUC). To construct ROC curves the reference standard CTR values were thresholded at 0.5 in order to obtain binary cardiomegaly labels. The sensitivity and specificity of each system and the reader performance is then computed at all possible operating points by applying various thresholds on the CTR output score (second-reader and seg-method) or SoftMax prediction for cardiomegaly (class-method) in order to produce an ROC curve.

It is important to note that the class-method was trained on a considerably larger dataset compared to the seg-method. This was done considering the different levels of annotation efforts between the two methods in order to have a fair comparison, and to investigate the performance of the class-method in its full potential. To validate our experimental design, we have also included the performance of the class-method when being trained with the same small dataset as the seg-method in our ROC analysis.

The kappa statistic [12] between the reference standard and the second reader and the models are calculated. Further, the sensitivity, specificity, positive predictive value, and negative predictive value and their 95% confidence intervals [17], [27] are reported, based on a fixed threshold of 0.5.

B. TRAINING SET SIZE ANALYSIS

In order to investigate the effect of the number of training images on the cardiomegaly classification performance, we constructed learning curves. We train both the seg-method and the class-method networks with varying numbers of training images and determine the effect of this on the method performance. The seg-method was trained with 50, 100, 200, 300, 400, 500, 600 images using 178 images in the validation set for each experiment and the class-method was trained with 2.5k, 5k, 10k, 20k, 40k, 62k images each using 3000 images

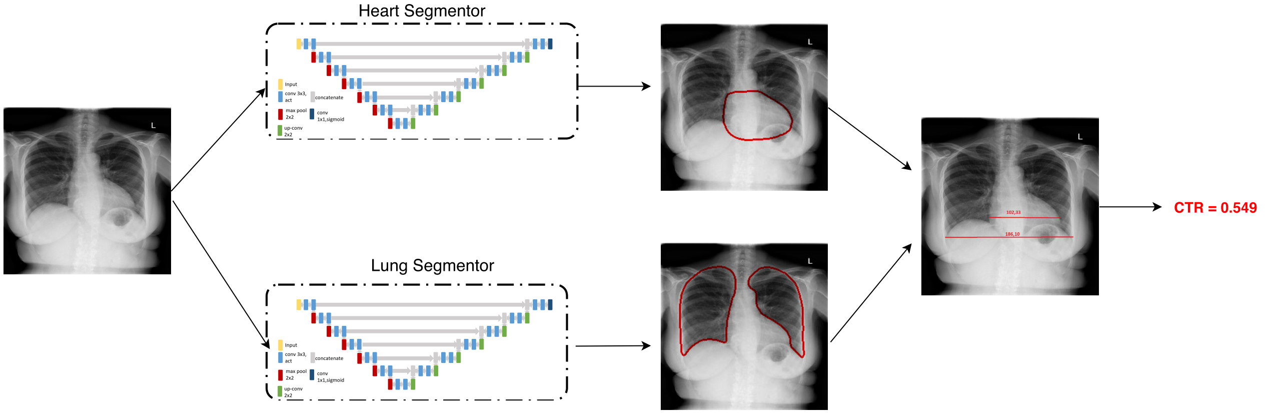


FIGURE 3. Illustration of the architecture pipeline for the seg-method. CTR = cardiothoracic ratio. Two different models are trained, for heart and lung field segmentation, respectively. CTR is derived from those predicted segmentation maps by determining the maximum horizontal thoracic and cardiac diameter and computing the ratio.

TABLE 2. Hyperparameter optimization for the seg-method. Regularization, learning and architecture related hyperparameters are optimized and ranges are demonstrated. The naming convention follows [10].

	Hyperparameter	Range	Best heart model	Best lung model
Regularization	Dropout	[True, False]	False	False
Learning	Batch size	[4,8,16]	4	4
	Optimizer	[Adam, SGD, Adagrad, RMSprop]	Adam	RMSprop
	Activation function	[ReLU, SELU, ELU]	SELU	ELU
	Learning rate	{0.00001, 0.01}	0.00018	0.00076
	Initializer	[Orthogonal, glorot_normal, he_normal, lecun_normal]	lecun_normal	he_normal
Architecture	Number of convolution per depth	[1, 2, 3]	2	1
	Depth of the network	[2, 3, 4, 5, 6]	6	6
	Input resolution	[64, 128, 256, 512]	512	512
	Number of initial feature maps	[32, 64]	32	32

as the validation set. We analyzed the results with the AUC score.

C. CTR ANALYSIS

1) HEART AND LUNG SEGMENTATION

Since the seg-method detects cardiomegaly through lung and heart segmentation, the segmentation performance of the final models, which were found through hyperparameter optimization, were evaluated on the full JSRT dataset (247 images). We used intersection over union (IOU), also known as Jaccard index, as a performance measure which is calculated as follows:

$$\mathrm{IOU} = \frac{\mid X \cap Y \mid}{\mid X \cup Y \mid}$$

where X represents the output of the network, and Y is the reference standard segmentation output. IOU quantifies the overlap between X and Y as the ratio between the number of pixels that are common between X and Y (cardinality of the intersection set) and the total number of pixels present across both of them (cardinality of the union set).

2) CTR CALCULATION

The performance of the seg-method was analyzed as a regression task in order to evaluate the performance in terms of CTR accuracy. Segmentation predictions can directly be used to calculate maximal horizontal cardiac and thoracic diameter, and used to calculate CTR as their respective ratio. The reference standard is created from the first radiologist CTR annotations to which the performance of the seg-method and the second reader can be compared.

The mean absolute error was used to evaluate the accuracy of CTR predictions with respect to the reference standard as follows:

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N} \mid \epsilon_t \mid,$$

where N denotes the number of images, and ϵ_t represents the difference between the predicted CTR and the reference standard CTR.

Moreover, CTR performance was also evaluated with Pearson correlation coefficient to summarize the strength of the linear relationship between the reference standard and the CTR predictions. The differences in CTR measurements, and the cardiac (in mm) and thoracic diameters (in mm) between the reference standard and the seg-method and the second reader were analyzed.

V. RESULTS

A. CARDIOMEGALY CLASSIFICATION

As shown in Figure 4, the class-method performed reasonably well, but with clearly much lower specificity at all sensitivity settings compared to the seg-method. The performance of the second reader and the seg-method are very similar to each other on this dataset with an AUC of 0.978 (95% confidence

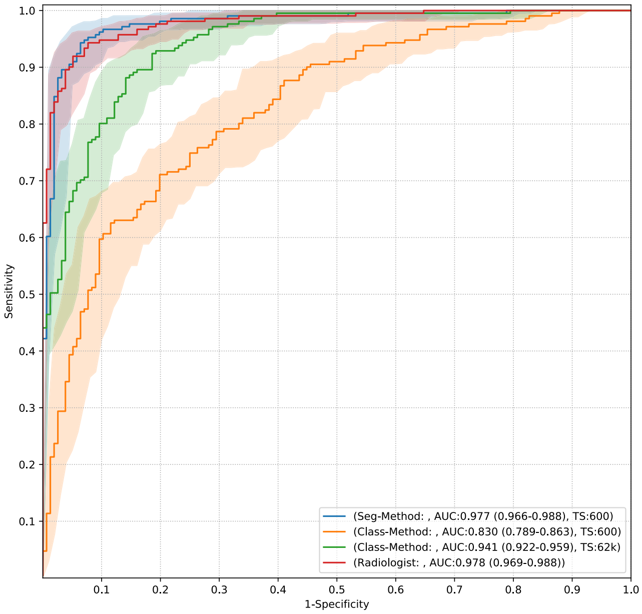


FIGURE 4. Receiver operating characteristic curves for detection of cardiomegaly in the held-out evaluation set (*N* = 367). Reference = Radiologist 1, TS: Number of training samples, The second reader (Radiologist 2). Shaded areas represent the 95% confidence intervals. The reference standard CTR values were thresholded at 0.5 in order to obtain binary cardiomegaly labels.

interval [CI]: 0.969, 0.988) and 0.977 (95% [CI]: 0.966, 0.988), respectively. In contrast, the class-method obtained an AUC of only 0.941 (95% confidence interval [CI]: 0.922, 0.959) when it was trained on a large dataset (62k). Further, the performance of the class-method decreased considerably achieving an AUC of 0.830 (95% confidence interval [CI]: 0.789, 0.863) when it was trained on the same small dataset as the seg-method (600).

The kappa statistic for cardiomegaly classification (at a threshold of 0.5) between the reference standard and the second reader was 0.856 while for the seg-method and class-method were 0.870 and 0.683, respectively. The sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) (at a fixed threshold of 0.5) on the held-out evaluation set are provided in Table 3. The seg-method and the second reader showed similar performance levels with the sensitivity of 0.97 and 0.91 and specificity of 0.90 and 0.95, respectively.

B. TRAINING SET SIZE ANALYSIS

The impact of the number of training images on the classification performance is illustrated in Figure 5a and 5b for both seg-method and the class-method. Figure 5a illustrates that seg-method benefits from an increased number of training images until the number of training images reaches 500. It seems that increasing this number further does not bring any performance gain.

The effect of the number of training images for the performance of the class-method appears to be more crucial compared to the seg-method in Figure 5b. The performance continues to increase substantially with the addition of more training data even after 40k training images.

Moreover, Figure 5a and 5b demonstrates that only 100 training images were sufficient for the seg-method to achieve a better performance than the class-method which was trained with 62k training images.

C. CTR ANALYSIS

1) HEART AND LUNG SEGMENTATION

The seg-method achieved 0.87 and 0.95 intersection over union (IOU) on the full JSRT dataset (247 images) for heart and lung segmentation, respectively.

2) CTR CALCULATION

The mean absolute error between both the seg-method and the second reader against the CTR reference standard was 0.0135 as seen in Table 3. The scatter plots of the reference standard CTR against the predicted CTR values of the model and the second reader are provided in Figure 6a and 6b, respectively. In line with our expectations, the misclassified cases for both the second reader and the seg-method are consistently those cases where the CTR is close to the threshold value of 0.5. Both the model and the second reader CTR predictions against the reference standard appear highly correlated, showing 0.960 and 0.965 Pearson correlation coefficient, respectively.

The histogram of the differences between the reference standard CTR values and the seg-method and the second reader are illustrated in Figure 6c and 6d, respectively. For both the seg-method and the second reader, the majority of the differences were less than 0.06. In particular, there were 7 cases out of 367 where the differences between both the seg-method and the second reader to the reference standard were higher than this value.

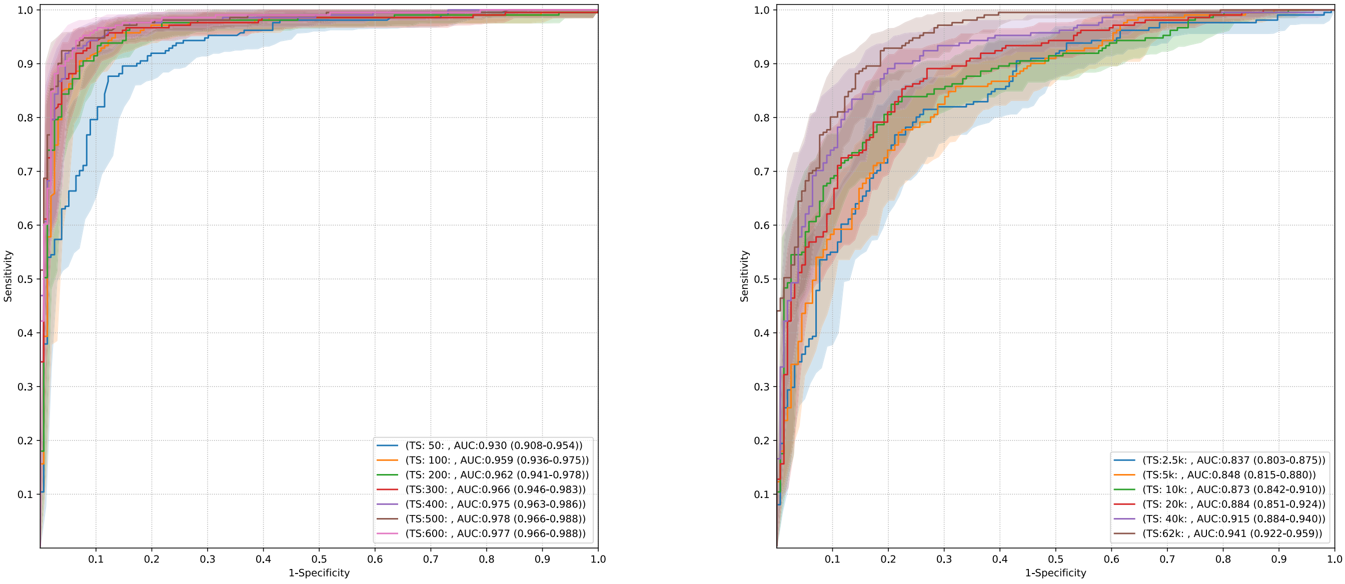
The range of differences between the reference standard maximal horizontal cardiac and thoracic diameters and the model and the second reader are shown in Figure 6e and 6f, respectively. The measurement differences for both the cardiac and thoracic diameters were in a similar range for the model and the second reader.

3) DIFFICULT CASE ANALYSIS

Example cases for the predictions of seg-method and class-method are shown in Figure 7. Misclassified cases where the reference standard is close to the CTR threshold of 0.5 are less interesting since these differences can be caused by inter-reader variability. Therefore we analyzed the misclassified cases where the reference standard was higher than 0.55 or lower than 0.45. There were no misclassified cases for both seg-method and class-method when the reference standard was lower than 0.45. However, class-method misclassified 8 cases where the reference standard was higher than 0.55 while the seg-method misclassified only one single case.

TABLE 3. Comparison of the seg-method with the class-method and the second reader. PPV = positive predictive value, NPV = negative predictive value, MAE = mean absolute error. The number between brackets denote 95% confidence intervals. MAE is calculated against the reference standard for CTR. Since the class-method produces a binary output, MAE can not be calculated. All other measures relate to binary classification of cardiomegaly status.

	MAE	Sensitivity	Specificity	PPV	NPV
Seg-Method	**0.0135**	0.97 [0.93, 0.99]	0.90 [0.84,0.94]	0.93 [0.88,0.96]	0.95 [0.90,0.98]
Class-Method		0.81 [0.75,0.86]	0.89 [0.83,0.93]	0.91 [0.86,0.94]	0.77 [0.70, 0.83]
Second Reader	**0.0135**	0.91 [0.87,0.95]	0.95 [0.90,0.98]	0.96 [0.92,0.98]	0.89 [0.83,0.93]



(a) Effect of training set size for the seg-method

(b) Effect of training set size for the class-method

FIGURE 5. The performance of the seg-method and the class-method for various training set sizes. TS = number of training samples. All curves are computed for the held-out evaluation set (*N* = 367). Shaded areas represent the 95% confidence intervals.

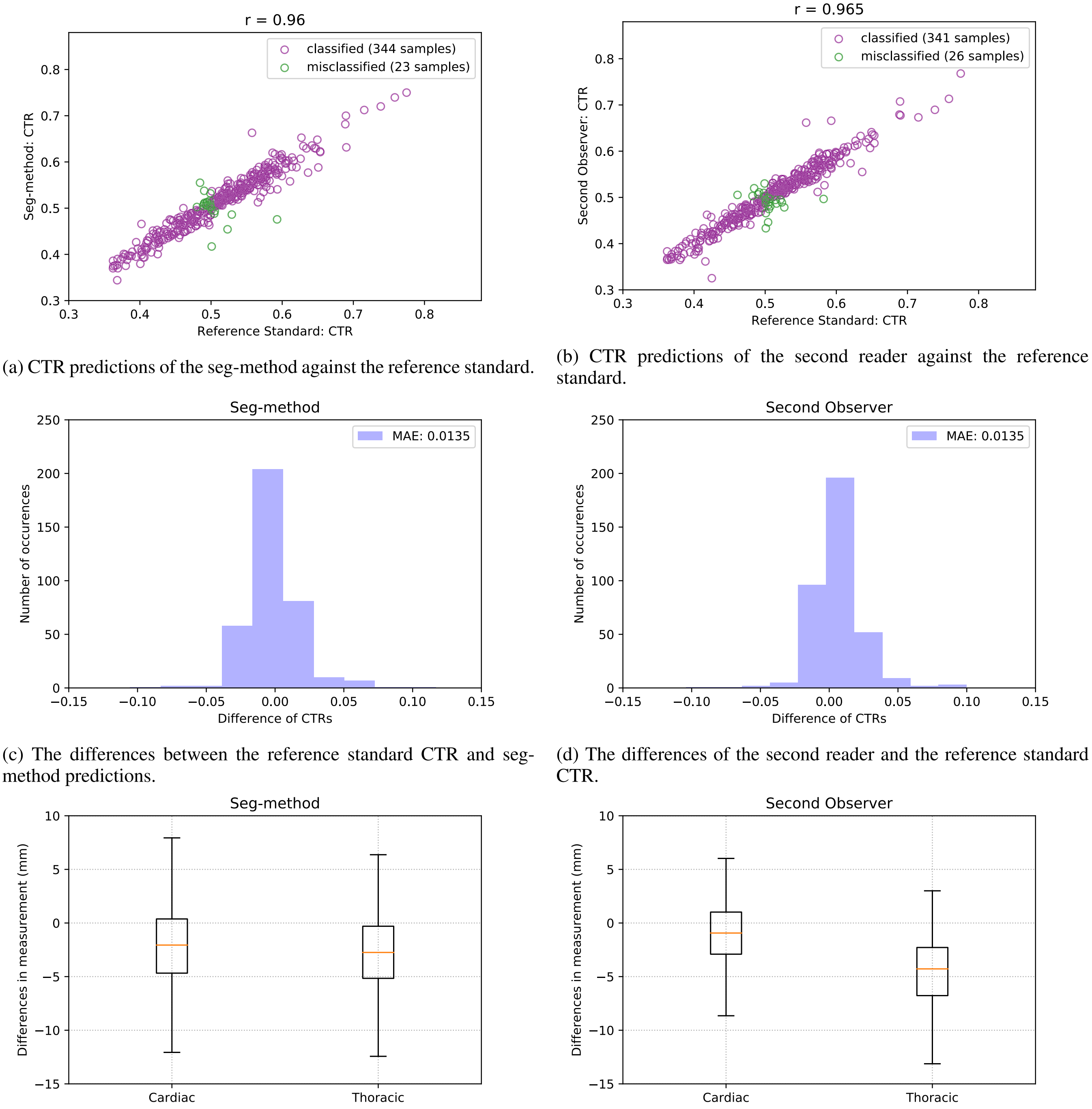
VI. DISCUSSION

In this work, it was demonstrated that a segmentation-based model trained on a modestly sized collection of chest radiographs (778 images) achieves an AUC of 0.977 for the detection of cardiomegaly, which is comparable to an independent second reader with an AUC of 0.978. The seg-method reached a high sensitivity and specificity on this task at 97% and 90%, respectively. In contrast, the class-method of image-level classification for cardiomegaly achieves a significantly lower performance with an AUC of 0.941 although it has been trained on 65,205 images. The performance achieved by the class-method is nonetheless representative of the state-of-the-art for classification-based solutions since several studies [2], [8], [22], [29], [30], [33] reported similar or lower cardiomegaly classification performance which were evaluated on a variety of datasets.

Experimental results demonstrated that the seg-method trained on only 100 annotated images can still outperform the class-method (Figure 5), trained on 65k images. This result highlights the difference between the methods in several aspects. First, it reveals that integrating domain knowledge from segmentations in subsequent image analysis may greatly reduce the volume of annotated training data required to achieve high performance. It additionally suggests that much improved accuracy can be obtained on these tasks, even with very limited training data. Finally, the seg-method opens the black-box solution of the class-method by producing the heart and lung segmentation and the diameters making up the CTR measure, rather than producing a single classification output. This is likely to be useful in clinical settings where the use of black-box algorithms is typically viewed as a high-risk solution.

It is notable that the class-method continued to improve in performance as additional training data was added. We hypothesize that with enough training samples it would eventually obtain a similar performance to the seg-method and the second reader. Further, the performance of class-method might be improved if the training labels did not contain any noise, although deep-learning systems have been shown to be robust to training label noise in recent studies [6], [31]. However the method would remain, nonetheless, inexplicable to clinicians.

Compared to the previous studies using segmentation-based solutions for cardiomegaly classification [7], [13],



(a) CTR predictions of the seg-method against the reference standard.

(b) CTR predictions of the second reader against the reference standard.

(c) The differences between the reference standard CTR and seg-method predictions.

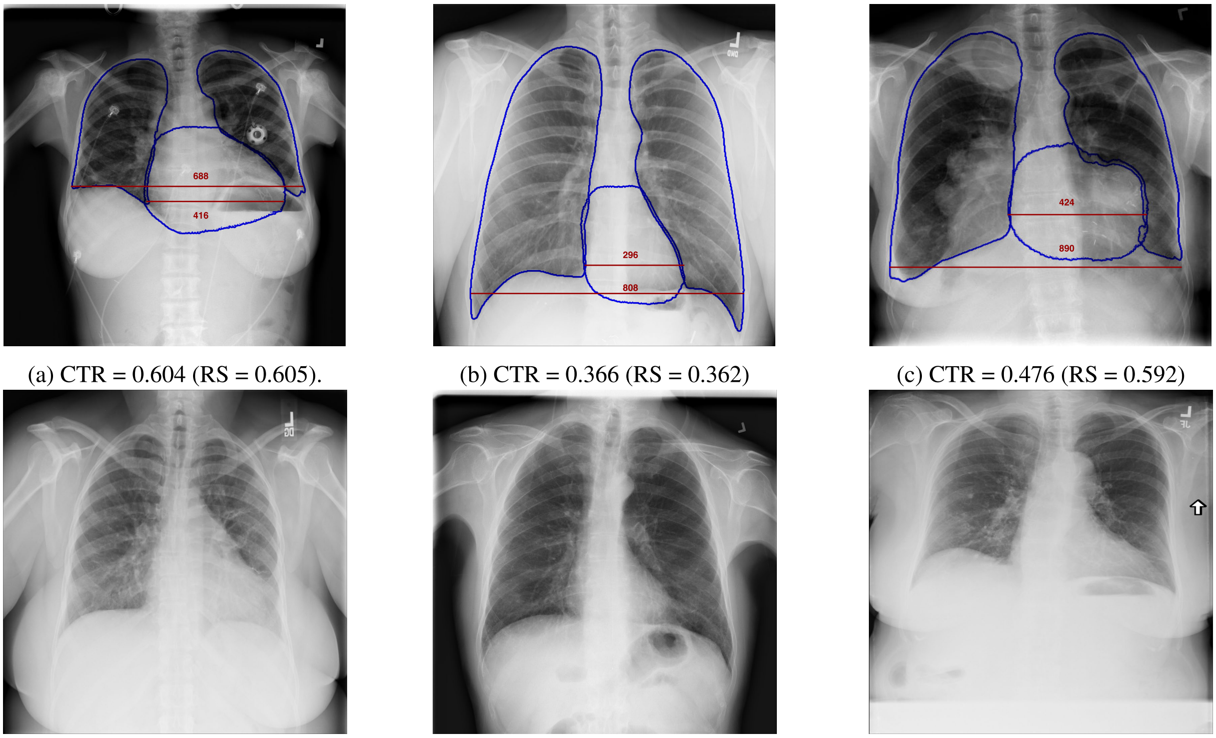
(d) The differences of the second reader and the reference standard CTR.

(e) The cardiac and thoracic diameter differences between the reference standard and the seg-method predictions.

(f) The cardiac and thoracic diameter differences between the reference standard and the second reader.

FIGURE 6. MAE = mean absolute error, CTR = cardiothoracic ratio. (a) and (b): The scatter plots of the reference standard CTR values against the CTR values of the seg-method and the second reader respectively. Correctly classified and misclassified samples are visualized in purple and green, respectively. (c) and (d): The histogram of the CTR differences between the reference standard and the seg-method and second reader respectively. (e) and (f): The box plot of the differences between the maximal horizontal cardiac and thoracic diameters between the reference standard and the seg-method and the second reader in mm.

our seg-method showed a substantially improved performance. Considering the fact that the heart and lung field segmentation performance is the key to the algorithm performance, it is clear that the improved performance of our seg-method relies heavily on our segmentation methodology. Unlike earlier studies, we employed a deep learning model, a state-of-the-art segmentation network [32], and systematically optimized its hyperparameters to segment the lung



(a) CTR = 0.604 (RS = 0.605).

(b) CTR = 0.366 (RS = 0.362)

(c) CTR = 0.476 (RS = 0.592)

(d) RS CTR = 0.610

(e) RS CTR = 0.425

(f) RS CTR = 0.569

FIGURE 7. Example cases of the model predictions. CTR = cardiothoracic ratio, RS = reference standard. (a)-(c): Three example cases of the seg-method predictions. Model prediction CTR (reference standard CTR). Cases a and b are correctly classified and case c is misclassified. (d)-(i): Example cases of the class-method predictions. d and e are the correctly classified cases, whereas f is an example of misclassification.

and heart field with optimal accuracy. This can be seen with the intersection over union (IOU) score reported for the heart and lung field segmentation in these studies. For instance, Candemir *et al.* [7] showed that they achieved IOU of 0.70 and 0.95 for the heart and lung segmentation, respectively whereas Dallal *et al.* [13] achieved IOU of 0.57 with their heart segmentation approach. However, our model achieved IOU of 0.87 and 0.95 on the JSRT dataset for heart and lung field segmentation, respectively, outperforming the results reported in [37].

This result suggests that there is a difference with a large margin in terms of heart segmentation performance between our proposed deep learning approach and the earlier studies. Recent work by Li *et al.* [25] which also used a deep learning segmentation model supports this result. In this work, the obtained CTR values are comparable with manual measurements although they required 5000 manually segmented scans for training, compared to just 778 in this work. Our work is the first to provide a direct comparison between segmentation-based and end-to-end image-level cardiomegaly classification demonstrating the advantages of the former, both in terms of clinical interpretation and performance. We also provide an online demo[1] where interested readers can test out our seg-method algorithm.

While it is clear that annotations of heart and lung boundaries are more time-consuming to obtain than image-level labels (which are often extracted using automatic methods from radiology reports), we believe that segmentation of anatomy is important not only for cardiomegaly detection, but also in the identification and quantification of many other abnormalities. Our manual segmentations took an average of 2 minutes per image (for both heart and lung boundaries) and we expect that our trained segmentation networks could now serve as guidance in many clinically interpretable abnormality detection systems. Future work will investigate the incorporation and importance of anatomical segmentation in other clinically relevant tasks.

This study has several limitations. First, all chest radiographs were retrieved from a single institution, which may affect the robustness of the system in evaluating images from other sources. Second, lateral view chest radiographs were not considered in our study although they might potentially be used, when in doubt, as complementary information to accept or reject cardiomegaly. Further, the cases for which the determination of CTR measurements was not possible (due to invisibility of anatomical boundaries) were manually excluded from our held-out evaluation set. In clinical practice, such images cannot be used for the determination of cardiomegaly. The automated rejection of such cases by the model would be a useful tool in clinical settings and might be a good future research direction.

[1] https://grand-challenge.org/algorithms/cxr-cardiomegaly-detection/

We conclude that we have implemented a segmentation-based cardiomegaly algorithm with performance comparable to a human reader, and with the advantages of improved accuracy and better interpretability compared to the image-level classification method. Future work will investigate extending the segmentation-based approach to other diagnostic tasks.

REFERENCES

[1] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, and X. Zheng, ''TensorFlow: A system for large-scale machine learning,'' in *Proc. USENIX Conf. Operating Syst. Design Implement.* Berkeley, CA, USA: USENIX Association, 2016, pp. 265–283.

[2] I. M. Baltruschat, H. Nickisch, M. Grass, T. Knopp, and A. Saalbach, ''Comparison of deep learning approaches for multi-label chest X-ray classification,'' *Sci. Rep.*, vol. 9, no. 1, pp. 1–10, Dec. 2019.

[3] J. Bergstra, B. Komer, C. Eliasmith, D. Yamins, and D. D. Cox, ''Hyperopt: A python library for model selection and hyperparameter optimization,'' *Comput. Sci. Discovery*, vol. 8, no. 1, 2015, Art. no. 014008.

[4] M. Buda, A. Maki, and M. A. Mazurowski, ''A systematic study of the class imbalance problem in convolutional neural networks,'' *Neural Netw.*, vol. 106, pp. 249–259, Oct. 2018.

[5] A. Bustos, A. Pertusa, J.-M. Salinas, and M. de la Iglesia-Vayá, ''PadChest: A large chest X-ray image dataset with multi-label annotated reports,'' 2019, *arXiv:1901.07441*. [Online]. Available: http://arxiv.org/abs/1901.07441

[6] E. Calli, E. Sogancioglu, E. T. Scholten, K. Murphy, and B. van Ginneken, ''Handling label noise through model confidence and uncertainty: Application to chest radiograph classification,'' in *Medical Imaging 2019: Computer-Aided Diagnosis*, vol. 10950. Bellingham, WA, USA: SPIE, 2019, Art. no. 1095016.

[7] S. Candemir, S. Jaeger, W. Lin, Z. Xue, S. K. Antani, and G. R. Thoma, ''Automatic heart localization and radiographic index computation in chest X-rays,'' in *Medical Imaging 2019: Computer-Aided Diagnosis*, vol. 9785. Bellingham, WA, USA: SPIE, 2016, Art. no. 978517.

[8] S. Candemir, S. Rajaraman, G. Thoma, and S. Antani, ''Deep learning for grading cardiomegaly severity in chest X-rays: An investigation,'' in *Proc. IEEE Life Sci. Conf. (LSC)*, Oct. 2018, pp. 109–113.

[9] G. Chlebus, A. Schenk, J. H. Moltz, B. van Ginneken, H. K. Hahn, and H. Meine, ''Automatic liver tumor segmentation in CT with fully convolutional neural networks and object-based postprocessing,'' *Sci. Rep.*, vol. 8, no. 1, pp. 1–7, Dec. 2018.

[10] F. Chollet. (2015). *Keras*. [Online]. Available: https://keras.io

[11] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, ''Fast and accurate deep network learning by exponential linear units (ELUs),'' 2015, *arXiv:1511.07289*. [Online]. Available: http://arxiv.org/abs/1511.07289

[12] J. Cohen, ''A coefficient of agreement for nominal scales,'' *Educ. Psychol. Meas.*, vol. 20, no. 1, pp. 37–46, Apr. 1960.

[13] A. H. Dallal, C. Agarwal, M. R. Arbabshirani, A. Patel, and G. Moore, ''Automatic estimation of heart boundaries and cardiothoracic ratio from chest X-ray images,'' in *Medical Imaging 2017: Computer-Aided Diagnosis*, vol. 10134. Bellingham, WA, USA: SPIE, Mar. 2017, Art. no. 101340K.

[14] D. Demner-Fushman, M. D. Kohli, M. B. Rosenman, S. E. Shooshan, L. Rodriguez, S. Antani, G. R. Thoma, and C. J. McDonald, ''Preparing a collection of radiology examinations for distribution and retrieval,'' *J. Amer. Med. Inform. Assoc.*, vol. 23, no. 2, pp. 304–310, Mar. 2016.

[15] K. Dimopoulos, G. Giannakoulas, I. Bendayan, E. Liodakis, R. Petraco, G.-P. Diller, M. F. Piepoli, L. Swan, M. Mullen, N. Best, P. A. Poole-Wilson, D. P. Francis, M. B. Rubens, and M. A. Gatzoulis, ''Cardiothoracic ratio from postero-anterior chest radiographs: A simple, reproducible and independent marker of disease severity and outcome in adults with congenital heart disease,'' *Int. J. Cardiol.*, vol. 166, no. 2, pp. 453–457, Jun. 2013.

[16] J. Duchi, E. Hazan, and Y. Singer, ''Adaptive subgradient methods for online learning and stochastic optimization,'' *J. Mach. Learn. Res.*, vol. 12, pp. 2121–2159, Feb. 2011.

[17] B. Efron, ''Nonparametric standard errors and confidence intervals,'' *Can. J. Statist.*, vol. 9, no. 2, pp. 139–158, 1981.

[18] S. Guendel, S. Grbic, B. Georgescu, K. Zhou, L. Ritschl, A. Meier, and D. Comaniciu, ''Learning to recognize abnormalities in chest X-rays with location-aware dense networks,'' 2018, *arXiv:1803.04565*. [Online]. Available: http://arxiv.org/abs/1803.04565

[19] K. He, X. Zhang, S. Ren, and J. Sun, ''Deep residual learning for image recognition,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[20] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, ''Densely connected convolutional networks,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269.

[21] S. Ioffe and C. Szegedy, ''Batch normalization: Accelerating deep network training by reducing internal covariate shift,'' 2015, *arXiv:1502.03167*. [Online]. Available: http://arxiv.org/abs/1502.03167

[22] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. L. Ball, K. S. Shpanskaya, J. Seekins, D. A. Mong, S. S. Halabi, J. K. Sandberg, R. Jones, D. B. Larson, C. P. Langlotz, B. N. Patel, M. P. Lungren, and A. Y. Ng, ''Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison,'' in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 590–597.

[23] A. E. W. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C.-Y. Deng, R. G. Mark, and S. Horng, ''MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports,'' *Sci. Data*, vol. 6, no. 1, pp. 1–8, Dec. 2019.

[24] G. Klambauer, T. Unterthiner, A. Mayr, and S. Hochreiter, ''Self-normalizing neural networks,'' in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 972–981.

[25] Z. Li, Z. Hou, C. Chen, Z. Hao, Y. An, S. Liang, and B. Lu, ''Automatic cardiothoracic ratio calculation with deep learning,'' *IEEE Access*, vol. 7, pp. 37749–37756, 2019.

[26] H. MacMahon, K. Doi, H.-P. Chan, M. L. Giger, S. Katsuragawa, and N. Nakamori, ''Computer-aided diagnosis in chest radiology,'' *J. Thoracic Imag.*, vol. 5, no. 1, pp. 67–76, 1990.

[27] R. G. Newcombe, ''Two-sided confidence intervals for the single proportion: Comparison of seven methods,'' *Statist. Med.*, vol. 17, no. 8, pp. 857–872, Apr. 1998.

[28] L. Oakden-Rayner, ''Exploring large-scale public medical image datasets,'' *Academic Radiol.*, vol. 27, no. 1, pp. 106–112, Jan. 2020.

[29] Q. Que, Z. Tang, R. Wang, Z. Zeng, J. Wang, M. Chua, T. S. Gee, X. Yang, and B. Veeravalli, ''CardioXNet: Automated detection for cardiomegaly based on deep learning,'' in *Proc. 40th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2018, pp. 612–615.

[30] P. Rajpurkar *et al.*, ''Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists,'' *PLOS Med.*, vol. 15, no. 11, Nov. 2018, Art. no. e1002686.

[31] D. Rolnick, A. Veit, S. Belongie, and N. Shavit, ''Deep learning is robust to massive label noise,'' 2017, *arXiv:1705.10694*. [Online]. Available: http://arxiv.org/abs/1705.10694

[32] O. Ronneberger, P. Fischer, and T. Brox, ''U-net: Convolutional networks for biomedical image segmentation,'' in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Intervent.*, in LNCS, vol. 9351, 2015, pp. 234–241.

[33] J. Rubin, D. Sanghavi, C. Zhao, K. Lee, A. Qadir, and M. Xu-Wilson, ''Large scale automated reading of frontal and lateral chest X-Rays using dual convolutional neural networks,'' 2018, *arXiv:1804.07839*. [Online]. Available: http://arxiv.org/abs/1804.07839

[34] J. Shiraishi, S. Katsuragawa, J. Ikezoe, T. Matsumoto, T. Kobayashi, K.-I. Komatsu, M. Matsui, H. Fujita, Y. Kodera, and K. Doi, ''Development of a digital image database for chest radiographs with and without a lung nodule: Receiver operating characteristic analysis of Radiologists' detection of pulmonary nodules,'' *Amer. J. Roentgenol.*, vol. 174, no. 1, pp. 71–74, Jan. 2000.

[35] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, ''Dropout: A simple way to prevent neural networks from overfitting,'' *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[36] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, ''Going deeper with convolutions,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[37] B. van Ginneken, M. B. Stegmann, and M. Loog, ''Segmentation of anatomical structures in chest radiographs using supervised methods: A comparative study on a public database,'' *Med. Image Anal.*, vol. 10, no. 1, pp. 19–40, Feb. 2006.

[38] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2097–2106.

[39] L. Yao, E. Poblenz, D. Dagunts, B. Covington, D. Bernard, and K. Lyman, "Learning to diagnose from scratch by exploiting dependencies among labels," 2017, *arXiv:1710.10501*. [Online]. Available: http://arxiv.org/abs/1710.10501

[40] L. Yao, J. Prosky, E. Poblenz, B. Covington, and K. Lyman, "Weakly supervised medical diagnosis and localization from multiple resolutions," 2018, *arXiv:1803.07703*. [Online]. Available: http://arxiv.org/abs/1803.07703

ECEM SOGANCIOGLU received the bachelor's degree in computer science with Hacettepe University, and the master's degree in computer science with the University of Freiburg with specialization in machine learning, in March 2017. She is currently pursuing the Ph.D. degree with the Diagnostic Image Analysis Group. Her research interest includes deep learning algorithms for chest x-rays.

KEELIN MURPHY received the Ph.D. degree in automated analysis of chest CT from the University Medical Center, Utrecht, in 2011. She worked for several years as a Postdoctoral Researcher at University College Cork, Ireland, on the analysis of neonatal MRI and EEG. She is currently with the Diagnostic Image Analysis Group, Radboud University Medical Center, Nijmegen. Her project incorporates analysis of x-ray and CT scans.

ERDI CALLI received the bachelor's degree in mathematical engineering from Isik University, in 2009, following his graduation, he spent six years working as a Software Engineer and the Manager, and the master's degree in artificial intelligence from Radboud University, in August 2017. Upon graduation, he started working at the Artificial Cognitive Systems lab as a Research Assistant, focusing on convolutional neural networks for mobile applications. He is currently pursuing the Ph.D. degree with the Diagnostic Image Analysis Group. His research interest includes deep learning algorithms for chest x-rays.

ERNST T. SCHOLTEN studied medicine at Vrije Universiteit (VU) Amsterdam, where he graduated, in 1974. He has been working as a General Radiologist with special interest in chest radiology and computed tomography, from September 1978 to August 2010. After his clinical career, he wrote several articles resulting in his thesis entitled subsolid nodules in lung cancer screening, which he defended in September 2014. As of 2004, he was involved in the NELSON trial on lung cancer screening as coordinating radiologist in Haarlem. Since February 2014, he has been a part-time Researcher with the Diagnostic Image Analysis Group.

STEVEN SCHALEKAMP received the medical degree from Vrije Universiteit Amsterdam, in 2011. Afterwards, he joined the Diagnostic Image Analysis Group and in 2015, he completed his Ph.D. entitled: advanced processing in chest radiography: impact on observer performance. In 2019, he finished his radiology residency at the Radboudumc. He is currently working as a Fellow Chest Radiology with the Meander Medical Center, Amersfoort. As a part-time Researcher, his focus is on validation and implementation of AI in radiology.

BRAM VAN GINNEKEN studied physics at the Eindhoven University of Technology and Utrecht University. He received the Ph.D. degree in computer-aided diagnosis in chest radiography from the Image Sciences Institute, in 2001. He is currently a Professor of medical image analysis with the Radboud University Medical Center, and the Chair of the Diagnostic Image Analysis Group. He also works for Fraunhofer MEVIS in Bremen, Germany. He is a Founder of Thirona, a company that develops software and provides services for medical image analysis. He has (co)authored over 200 publications in international journals. He is a member of the Fleischner Society and the Editorial Board of Medical Image Analysis. He pioneered the concept of challenges in medical image analysis.