




OPEN

Lung nodule detection in chest X-rays using synthetic ground-truth data comparing CNN-based diagnosis to human performance

Manuel Schultheiss^{1,2}, Philipp Schmette¹, Jannis Bodden², Juliane Aichele², Christina Müller-Leisse², Felix G. Gassert², Florian T. Gassert², Joshua F. Gawlitza², Felix C. Hofmann², Daniel Sasse², Claudio E. von Schacky², Sebastian Ziegelmayr², Fabio De Marco¹, Bernhard Renger², Marcus R. Makowski², Franz Pfeiffer^{1,2} & Daniela Pfeiffer²

We present a method to generate synthetic thorax radiographs with realistic nodules from CT scans, and a perfect ground truth knowledge. We evaluated the detection performance of nine radiologists and two convolutional neural networks in a reader study. Nodules were artificially inserted into the lung of a CT volume and synthetic radiographs were obtained by forward-projecting the volume. Hence, our framework allowed for a detailed evaluation of CAD systems' and radiologists' performance due to the availability of accurate ground-truth labels for nodules from synthetic data. Radiographs for network training (U-Net and RetinaNet) were generated from 855 CT scans of a public dataset. For the reader study, 201 radiographs were generated from 21 nodule-free CT scans with altering nodule positions, sizes and nodule counts of inserted nodules. Average true positive detections by nine radiologists were 248.8 nodules, 51.7 false positive predicted nodules and 121.2 false negative predicted nodules. The best performing CAD system achieved 268 true positives, 66 false positives and 102 false negatives. Corresponding weighted alternative free response operating characteristic figure-of-merits (wAFROC FOM) for the radiologists range from 0.54 to 0.87 compared to a value of 0.81 (CI 0.75–0.87) for the best performing CNN. The CNN did not perform significantly better against the combined average of the 9 readers ($p = 0.49$). Paramediastinal nodules accounted for most false positive and false negative detections by readers, which can be explained by the presence of more tissue in this area.

With accounting for over 1.7 million deaths in 2018, lung cancer is one of the most common causes of cancer death worldwide¹. Regular screening using chest x-ray (CXR) or low dose computed tomography (LDCT) is under investigation, with the latter being more effective, but also more expensive^{2–4}. While standard CXR screening has only shown to improve early detection but not a decrease in mortality², computer aided diagnosis (CAD) systems could increase sensitivity and therefore improve its benefit as a screening method³.

While the applied dose for CXR is significantly lower than for CT (typically 0.1 mSv for a posteroanterior and lateral CXR study and 1.5 mSv for low dose CT)^{5,6}, the detection of nodules in chest CXR is more challenging than for CT. Lung metastases often originate from extra-thoracic malignancies (ETM), with the lungs being a frequent site of metastatic growth: for patients, who died of an ETM, incidences of pulmonary metastases are reported to be greater than 19%^{7–9}.

Hence, it is of interest to identify positions in the lung, where radiologists have problems detecting nodules correctly in order to improve training.

Additionally, radiologists can be assisted by CAD systems for CXR diagnosis^{10,11} and it may further improve sensitivity for CXR based lung cancer screening³. Here, with the rise of computing power, deep-learning based CAD systems gained interest recently: for automatic x-ray image classification several approaches have been

¹Chair of Biomedical Physics, Department of Physics and Munich School of BioEngineering, Technical University of Munich, 85748 Garching, Germany. ²Department of Diagnostic and Interventional Radiology, School of Medicine and Klinikum rechts der Isar, Technical University of Munich, 81675 Munich, Germany. ✉email: manuel.schultheiss@tum.de

	TP	FP	FN
RetinaNet	268	66	102
U-Net	256	279	114
Reader 1	244	5	126
Reader 2	278	15	92
Reader 3	207	29	163
Reader 4	185	9	185
Reader 5	201	9	169
Reader 6	294	35	76
Reader 7	273	52	97
Reader 8	281	276	89
Reader 9	276	35	94

Table 1. True positives (TP), false positives (FP) and true negatives (FN) for RetinaNet, U-Net and the readers. For RetinaNet a nodule with a confidence score greater than 0.5 was counted as positive.

published^{12–15}, which may assist radiologists in clinical practice. U-Net like architectures were successfully employed for segmentation tasks^{16,17} and RetinaNet based detector for object detection tasks in radiographs^{18–21}. While U-Net based implementations yield a segmentation as output, RetinaNet detectors output bounding boxes which most likely contain the object of interest (in our case a nodule). Additional scores for each box reflect the certainty of the network for a detection here. Both approaches require either pixel-level or bounding-box annotations for training and evaluation of deep-learning systems.

Such annotations are cost expensive, as annotations usually have to be carried out by an expert in the respective area. In our case, for CXR nodule detection, a radiologist needs to mark suspected lesions by hand in order to make the data applicable for CNN training and evaluation. Therefore, an option to run pre-clinical trials is the use of synthetic data. Here, virtual clinical trials (VCTs) play an important role in the testing of imaging systems²². In these trials, usually the body anatomy and physics of the image acquisition system are simulated. Such systems have been developed for a wide range of modalities, such as CT²³ or mammography²⁴. However, the use of computational phantoms does never completely resemble a real human anatomy.

For a lung nodule detection task, it is possible to combine a real anatomy by the use of real CT scans with synthetic (or virtual) nodules: Yu et al. developed a simulation framework for nodule detection in CT scans where a virtual nodule was inserted into a real scan and nodule detection performance was evaluated by 4 observers²⁵. For lung nodule detection in radiographs our simulation approach is very similar, but additionally generates a radiograph from the CT scan: we place nodules in random positions within the lung of a CT scan and forward project the CT scan. Hence, it is possible to generate a lot of different radiographs for each available CT scan by altering the nodule positions. Contrary to manually annotated data, it is possible to retrieve the exact contours of every inserted tumor (e.g. the groundtruth), which is beneficial for training CNNs with box annotations or pixel-level annotations.

In this study, the usability of simulated CXRs for lung nodule detection performance evaluation and CNN training is demonstrated. We train multiple CNNs with synthetically generated data and evaluate the performance against nine radiologists. It is shown that on synthetic data, CNNs are able to reach a performance similar to radiologists. The simulation framework further allows to examine the areas of false negative detections, e.g. areas where radiologists had problems identifying tumors.

Results

The performance of nine radiologists and two CNN algorithms was evaluated for the nodule detection task. Example detections are shown in Fig. 1. False negative detections are shown in Fig. 2. Absolute true positive, false positive and false negative numbers are reported in Table 1.

Here, U-Net yielded a high rate of true positives, but also the most false positives. To retrieve a combined score of false positives and true positives, weighted alternative free response operating characteristics (wAFROC) FOMs were calculated and presented in Table 2.

The FOM score of the RetinaNet network was higher than that of four readers and lower than that of five readers. The FOM score of U-Net was lower than that of eight readers. Corresponding graphs are shown for FROC (Fig. 3A) and wAFROC (Fig. 3B) metrics. The RetinaNet CNN did not perform better against the average of all readers (0.78 average reader FOM, $p = 0.49$). Combining U-Net and RetinaNet, by counting the bounding boxes of RetinaNet as positive when a U-Net segmentation was found within the bounding-box area, the FOM score decreases slightly to 0.78. True positives with respect to nodule size are shown in Fig. 4.

Methods

Radiographs for network training and validation (U-Net and RetinaNet) were generated from 855 CT scans of a public dataset. For the reader study, radiographs were generated from nodule-free CT scans with altering nodule positions, sizes and nodule counts per radiograph. Nodules were segmented from another CT scan, augmented and inserted into each of the CT scans at a randomly selected position within the lung. Next, a forward projection was performed in order to generate a realistic, synthetic radiograph. By changing nodule position, size and

	wAFROC				
	FOM	CI Lower	CI Upper	StdErr	wLLF
RetinaNet	0.81	0.75	0.87	0.028	0.71
U-Net	0.58	0.47	0.68	0.052	0.41
Reader 1	0.82	0.79	0.86	0.017	0.71
Reader 2	0.87	0.83	0.90	0.017	0.79
Reader 3	0.74	0.68	0.79	0.029	0.57
Reader 4	0.74	0.70	0.78	0.022	0.59
Reader 5	0.78	0.75	0.81	0.015	0.65
Reader 6	0.87	0.83	0.91	0.020	0.79
Reader 7	0.83	0.79	0.88	0.022	0.74
Reader 8	0.54	0.44	0.63	0.048	0.31
Reader 9	0.84	0.79	0.88	0.023	0.75

Table 2. Figure of merits (FOM), 95% confidence intervals (CI) and standard error (StdErr) for wAFROC metrics. The weighted lesion localization fraction (wLLF score) was retrieved at an x-axis 0.2 operating point for all readers and CNNs on the wAFROC curve.

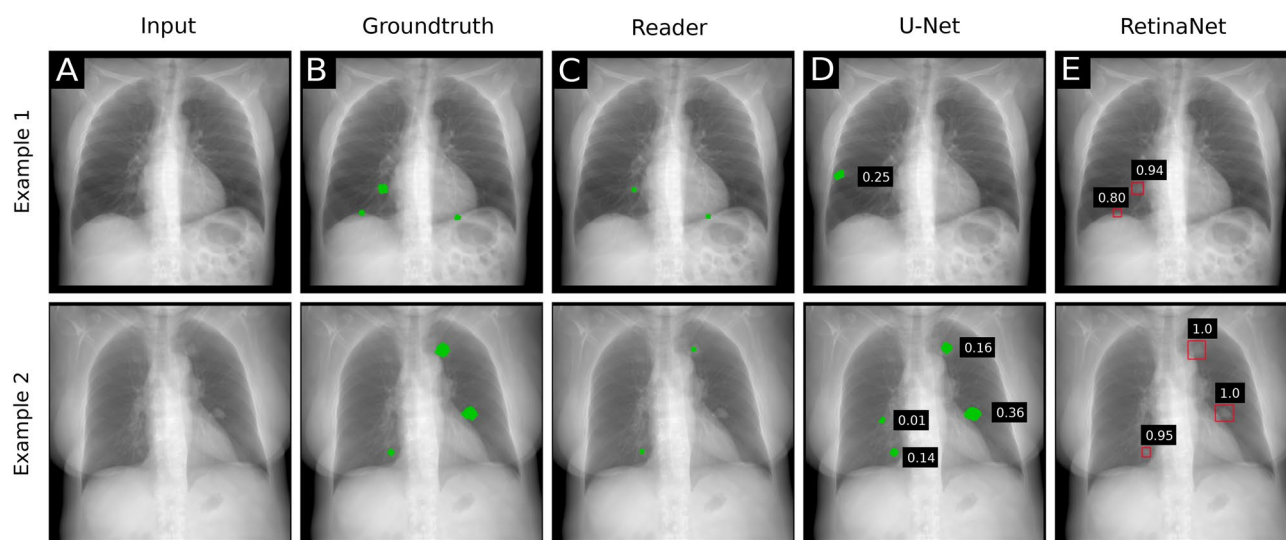


Figure 1. Synthetic radiographs, ground truth masks and results of reader and computer-based detection. (A) Synthetic input radiograph as shown to the reader and evaluated by the CNNs (B) corresponding ground-truth radiograph with nodules marked green (C) center position of nodules marked by a reader (D) U-Net prediction (E) RetinaNet bounding-box predictions with scores.

count, this technique allows the generation of multiple, different radiographs out of a single CT scan. Overall workflow is illustrated in Fig. 5.

Preprocessing. To simulate a x-ray image, HU values are converted back to their respective absorption values. The absorption value for a voxel μ_x , can be calculated from a HU voxel value HU_x according to:

$$\mu_x = \mu_{water} + (\mu_{water} - \mu_{air}) \cdot \frac{HU_x}{S},$$

whereat the scale factor S is vendor specific and usually 1000 or 1024 and $\mu_{water}=2.059 \times 10^{-1} \text{ cm}^{-1}$. A first step is to remove the patient table from the CT scan, as the table does not appear in radiographs. This is done by using a combination of thresholding and a connected component algorithm, which removes the second largest object (table). To segment the lung, a tissue mask around the lung is extracted by thresholding. Afterwards the lung-area is identified using a hole-filling algorithm.

Nodule insertion. During training, in each CT scan between 1 and 6 nodules are inserted. Each nodule is chosen from 19 segmented nodules with an equal probability. The nodule is normalized to values between 0 and 1 and multiplied with the absorption value of soft tissue. Furthermore the nodule is randomly augmented during the training process: Here, the nodule is rotated on the coronal plane by a value chosen from a uniform

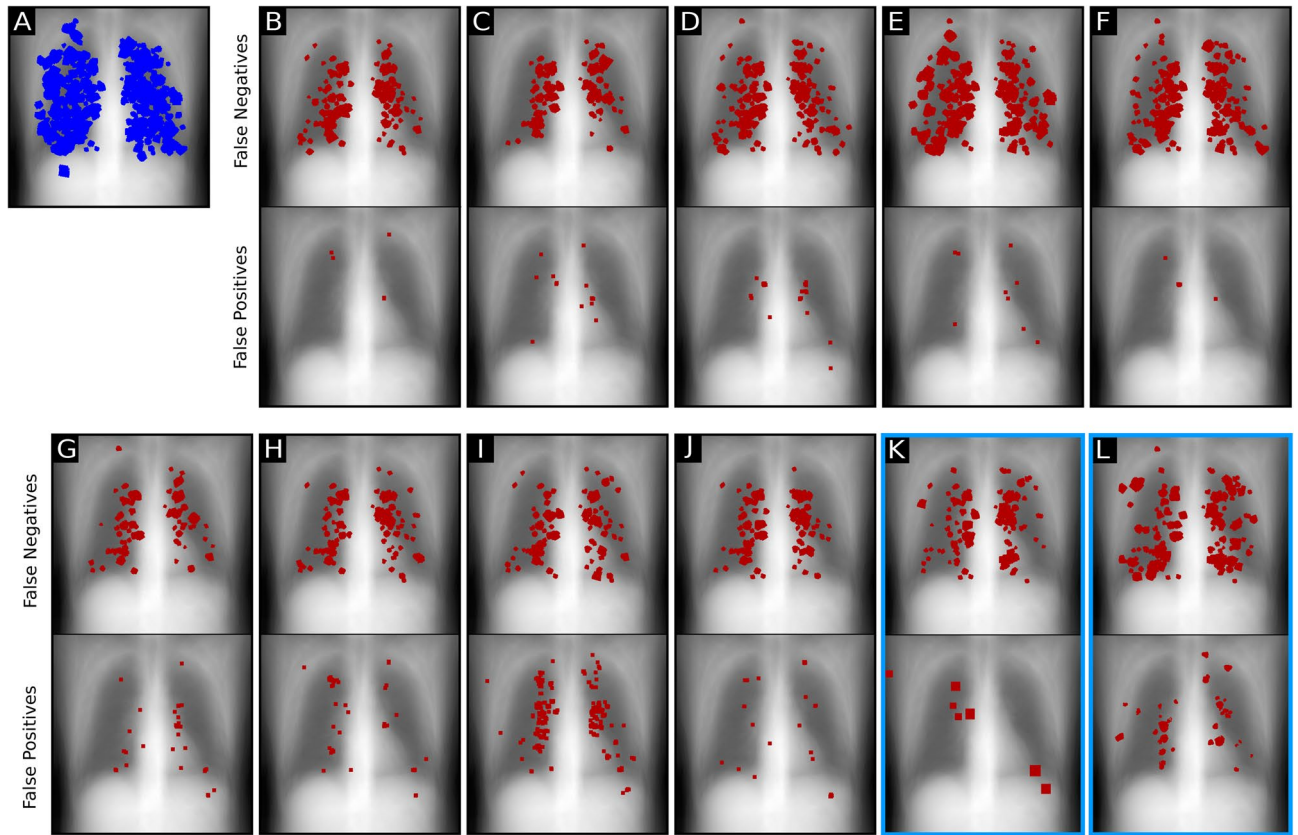


Figure 2. Localization of false negative and false positive predictions in the reader study. Backgrounds were determined by averaging over all reader study radiographs. (A) All inserted nodules of different sizes in all radiographs marked blue. (B–J) False negative and false positive predictions by reader. (K) Location of false negative predictions of RetinaNet and false positive predictions of RetinaNet. (L) False negative predictions of U-Net and false positive predictions of U-Net.

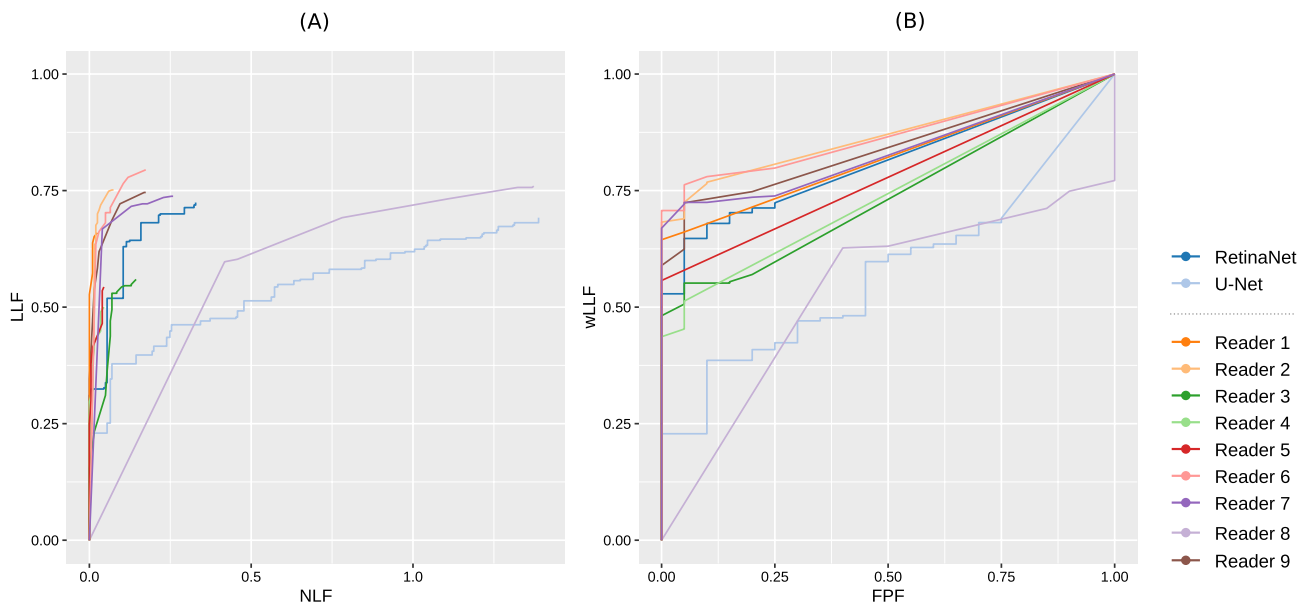


Figure 3. Comparison of CNN and reader based diagnostic performance. (A) FROC plot with lesion localization fraction (LLF) plotted against non lesion fraction (NLF) (B) wAFROC plot with weighted LLF on the ordinate. The plot was generated using Rjafroc²⁶.

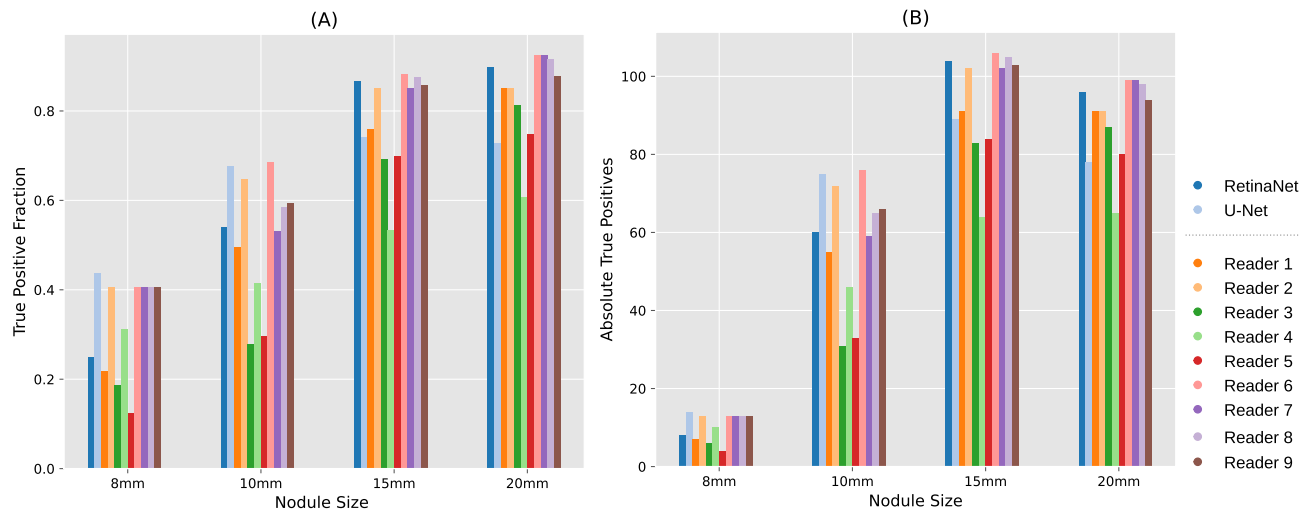


Figure 4. Comparison of detection performance by nodule size (A) Relative true positive fraction (B) absolute number of true positives for RetinaNet CNN, U-Net CNN and readers R1–R9. The plot was generated using Matplotlib²⁷.

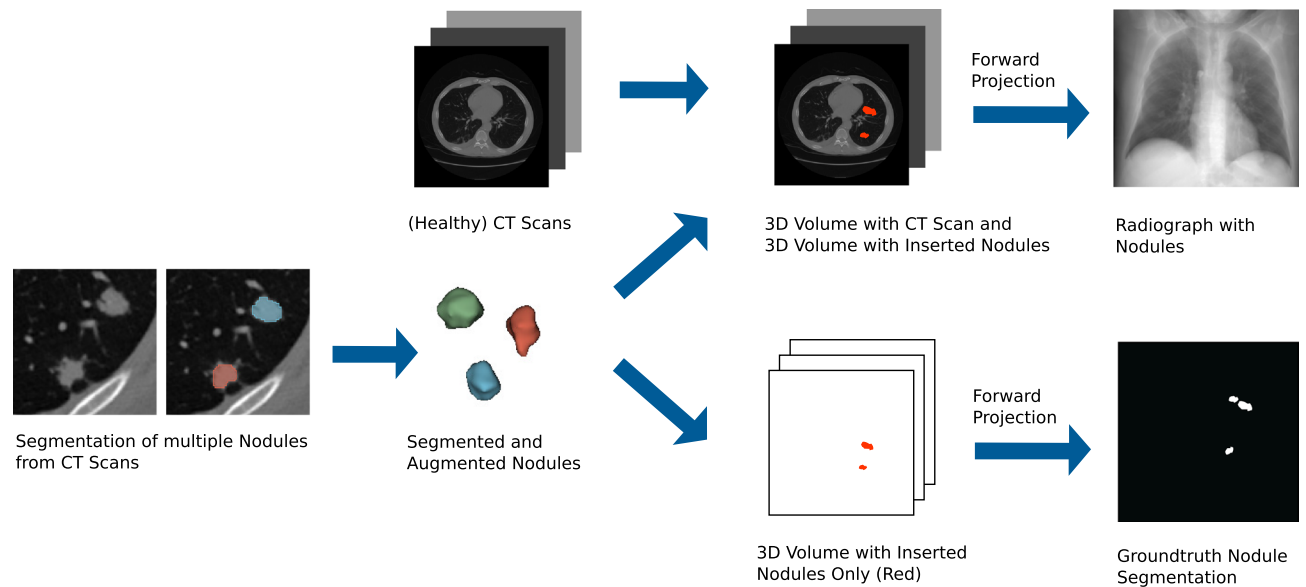


Figure 5. Workflow for generating synthetic radiographs containing tumour nodules with perfect ground truth knowledge. Based on natural shapes, various sizes of tumors are generated and subsequently inserted into clean CT scans and different locations. The 3D CT data set is then forward projected to generate p.a. thorax radiographs. In parallel, the tumors only are forward projected to obtain perfect ground-truth masks. These ground-truth masks is later used to compare the radiologist's findings with the expected findings.

distribution between 0 and 360 degrees. Furthermore, the nodule is randomly scaled to values between 8 and 20 mm along all three axes. The nodule-free volume and the volume with the nodules are forward projected and summed up in order to obtain a simulated diseased radiograph, whereby absorption coefficients are weighted by their voxel size. Also the nodules without the surrounding CT volume are forward projected in order to obtain the groundtruth. Both the diseased radiograph and the groundtruth are resized to 512×512 pixels for training.

Dataset description. Data access was approved by the institutional ethics committee at Klinikum Rechts der Isar (Ethikvotum 87/18 S) and the data was anonymized. The ethics committee has waived the need for informed consent. All research was performed in accordance with relevant guidelines and regulations.

CNN training was performed with 855 CT scans from the LUNA16²⁸ dataset, whereat 80% of data was chosen for training and 20% of data for validation. Nodules to be inserted in the CT scans were segmented from 5 CT scans. Segmented lung nodules were metastasis from various malignant tumors. Metastases origin tumor was

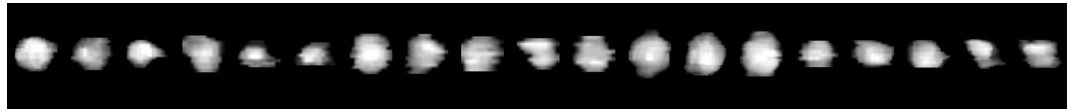


Figure 6. Projected segmentations of 19 nodules, which were artificially inserted into the synthetic radiographs.

carcinoma of the breast in two patients (9 metastases total), colorectal cancer in two patients (5 metastases total) and melanoma in one patient (5 metastases total). Projected segmentations are shown in Fig. 6.

For the reader study, a dataset of 21 CT scans was collected from our institution's picture archiving and communication system (PACS). These scans were checked to be unsuspecting (nodule-free) by one radiologist (JB, 3 years of experience).

Reader study. For the reader study, 201 radiographs were generated from 21 nodule-free CT scans with altering nodule positions, sizes and nodule counts per radiograph. Diameters for inserted nodules in the reader study were 8 mm, 10 mm, 15 mm and 20 mm. Of all radiographs, 20 radiographs contained no nodule, 53 radiographs contained 1 nodule, 67 radiographs contained 2 nodules and 61 radiographs contained 3 nodules. Within the nodule-present cases, corresponding fractions are 53/181 cases with one nodule, 67/181 cases with 2 nodules and 61/181 cases with three nodules.

Of 370 inserted nodules, 32 had a size of 8 mm, 111 had a size of 10 mm, 120 had a size of 15 mm, and 107 had a size of 20 mm.

Reader experience was one month for one radiologist, nine month for one radiologist, at least one year for two radiologists, at least two years for two radiologists and at least three years for three radiologists. Readers were given the task of marking lung tumors and indicating confidence for each tumor on a scale from 1 to 100. Only posterior-anterior radiographs were used in the reader study. In order to simulate a clinical setting, each radiologist was given a time constraint of 20 seconds per radiograph.

The reader study dataset was also the test set for CNN evaluation in order to compare CNN performance to reader performance. It was ensured, no CT scans of the test set or the reader study were part of the training or validation set.

CNN architectures and network training. We investigate two CNN architectures: First, a U-Net¹⁷ like architecture is used and second, a RetinaNet²⁰ based object detector is trained. For the U-net architecture, training was performed for 400 epochs with 3200 steps per epoch and a batch size of 1. Adam optimizer parameters were set to $\beta_1 = 0.9$ and $\beta_2 = 0.999$ with a learning rate of 10^{-3} . Applied loss function was a Dice loss, as suggested by Milletari et al.²⁹. In order to retrieve per lesion score for the U-Net, we trained a second helper network for the U-Net: A lesion scoring network, which inputs a patch centered on the lesion, was trained for 500 epochs with a learning rate of 10^{-5} , a batch size of 32 and 87 steps per epoch. Augmentation included rotation, shift and flip operations. Positive and negative patches were equally sampled. Positive patches for the lesion scoring network were extracted from the available training segmentations. Using a hard-negative mining^{30,31} approach, negative patches for the lesion scoring network were extracted from positions, where the U-Net yielded a prediction on healthy radiographs. The overall architecture for the U-Net based approach is illustrated in Fig 7.

For RetinaNet, training was performed for 50 epochs with a step size of 1000. The batch size was set to 1. The backbone was set to ResNet-101³². Loss function hyperparameters were set to $\alpha = 0.25$, $\gamma = 2.0$. The learning rate was set to 10^{-5} . It was reduced by factor 0.1 after the loss did not change for more than 3 epochs ($\delta = 0.0001$). Data augmentation transformations for RetinaNet included contrast, brightness, shear, scale, flip, and translation. Models were implemented using Tensorflow³³ and Keras³⁴. Plots were generated using Matplotlib²⁷ and RJafroc²⁶. Furthermore, RetinaNet models are based on *keras-retinanet*³⁵. Weights were obtained from the epoch with the best validation loss for both architectures.

Data analysis. Usually the tradeoff between sensitivity and specificity can be analyzed using receiver operating characteristics (ROC). As this technique is only applicable for binary classification tasks on case-level, free response ROC (FROC) methods to evaluate detection performance on lesion level were introduced^{36–38}. Here, a FROC plot consists of two axes: (1) The lesion localization fraction (LLF), defined as the number of true positives divided by the total number of lesions. (2) The non lesion localization fraction (NLF) as the total number of false positives divided by the total number of cases. However, in this method patients with more lesions are weighted more. To compensate for this, the weighted alternative FROC (wAFROC)^{38–40} is used: it assigns a weight w to each lesion, which sum up to unity on patient level and therefore ensures each patient is an equal representative of the population.

Given a threshold ζ , above which nodules are counted as positives, the number of nodule-containing cases K_N , the number of Lesions L_k for each case k , the lesion weight W_{kl} and the indicator function I , which returns 1 if the argument is true and zero otherwise,

$$wLLF_r(\zeta) = \frac{1}{K_N} \sum_{k=1}^{K_N} \sum_{l=1}^{L_k} W_{kl} I(z_{kl} \geq \zeta),$$

and a false positive fraction (FPF)

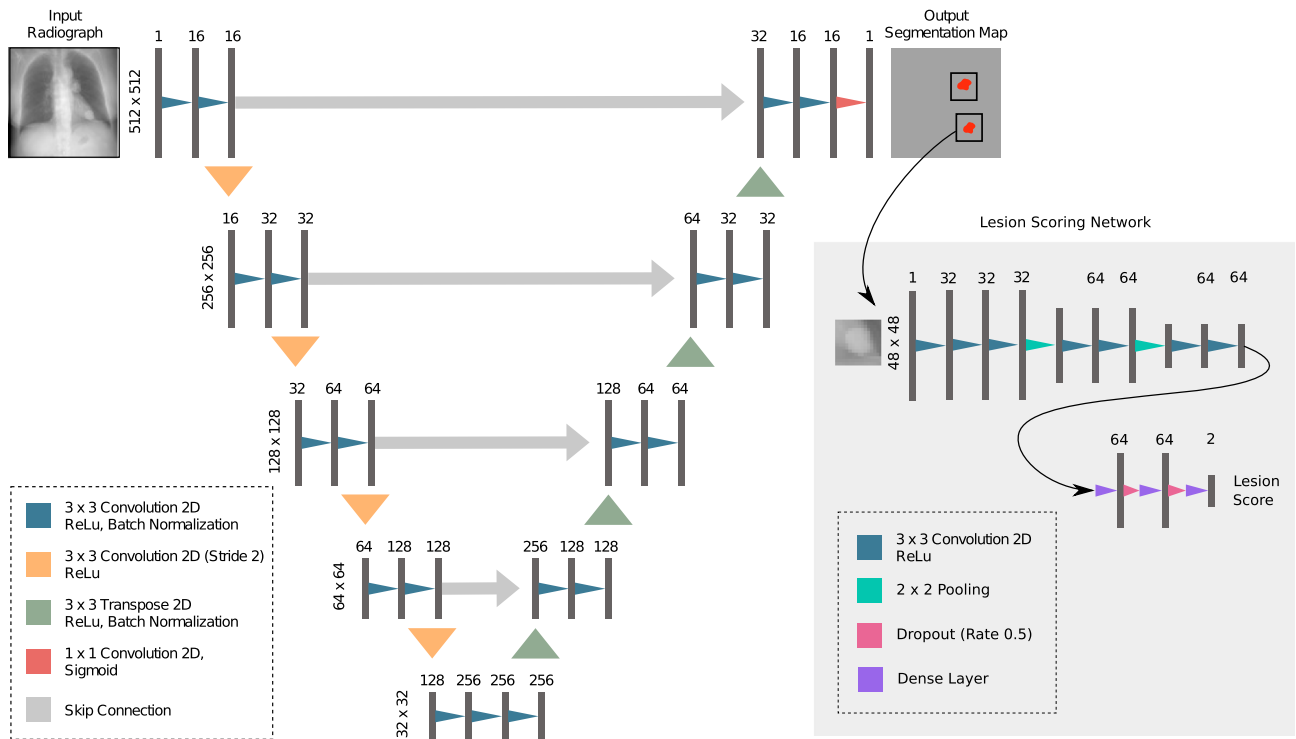


Figure 7. Architecture of the utilized U-Net. Input and output were a single channel 512×512 matrix. Downsampling was performed using convolutional layers with a stride of 2. A second network (lesion scoring network) was used to retrieve a per-lesion score of the segmented nodules. Numbers above layers indicate convolution filters for convolution layers and number of neurons for dense layers.

$$FPF(\zeta) = \frac{1}{K_F} \sum_{k=1}^{K_F} I(FP_k \geq \zeta),$$

given the number of nodule-free cases K_F , which sums up nodule-free cases with false positive detections FP_k ^{39,40}.

To count a prediction as true positive, a distance criterion needs to be defined. In our study, a true positive is counted if the distance of the prediction center of mass (COM) is below 30 pixels to the ground-truth and the score is above or equal ζ . Furthermore, we define the weight w of all lesions in a single patient as equal (e.g. if a patient has 4 lesions, the weight of every lesion is 0.25). Software used for evaluation was RJAfroc²⁶. Significance testing was done with $\alpha = 0.05$.

Discussion

Research on nodule location detection in radiographs was primarily performed using large sets of manually annotated data. To reach human observer performance, McKineey et al. reported 11734 annotated radiographs in their training dataset^{18,41}. However, it requires a lot of time and cost intensive work to annotate and delineate nodules. While approaches are available that work without pixel-level or box-level annotations, e.g. weakly supervised learning^{12,13,15,42,43}, the provided output locations are usually not very accurate, compared to segmentation or bounding box approaches. Hence, the presented method is a potential alternative method for pre-clinical evaluation of deep learning systems without the need of large sets of manually annotated data. The benefit here is that, contrary to manually annotated data, the ground-truth delineations of tumors are perfectly accurate, and single tumors can not be missed or wrongly delineated.

However, the application of the trained model on non-simulated radiographs remains challenging and is still under development, referred to as domain randomization: Previous investigations tried to generate real-world car detection or robotic systems by use of synthetic images⁴⁴⁻⁴⁶. This can be achieved by applying a large amount of unrealistic perturbations to the training domain. Studies in the medical field were performed by Toth et al.⁴⁷, who registered cardiac models to radiographs. In our work we already did first preparations to transfer the CT generated data to the CXR domain by removing the patient table from the CT scan. Further challenges are the different arm positions in CT scans and higher resolution of radiographs. The arms can not be simply removed or masked out, as the different pose affects the position of the scapula and thus the visibility of the lungs is different for the two modalities.

To scale up resolution a superresolution network like Yamanaka et al.⁴⁸ can be applied. Superresolution networks already have been applied successfully to chest-radiographs⁴⁹. However, in our opinion, the main challenge lies in the modelling of the tumor shape. In this work, we used a shape from a pool of 19 tumors, which is augmented by rotation and rescaling to model more tumor shapes. However, this approach does most likely

not capture the complete variance of tumor shapes, as tumor shapes vary broadly. Here, a spatial tumor model as described by Vogelstein et al.⁵⁰ could be helpful.

The employed framework facilitates to analyze locations of false negative detections, and doing so showed some differences between CNN and radiologists: While most false negative detections by radiologists and by the RetinaNet CNN were located in paramediastinal positions, the U-Net CNN showed false negatives more uniformly distributed across the lung. Increased false negative rates along the mediastinum were already found in prior studies of blind-spot detection on CXRs^{51–53}. The concentration of observer false negatives in paramediastinal positions could be due higher absorption coefficients in this area and therefore less contrast. Another interesting observation in our study was that human readers did not fully utilise the 1–100 scale, but 86% of ratings were provided in increments of 10 (e.g. 10, 20, 30,...).

This study has some limitations: Above all, as stated before, the simulation does not completely resemble the real setting: The tumor shape generation was implemented using simple augmentation model, due to less implementation effort. Forward-projections were performed by a parallel beam projector, as cone-beam projectors are computationally more challenging. The radiograph resolution is further limited to 512 pixel width, as the CT scan resolution is not higher. Furthermore, while the test set was checked to be nodule-free, the potential presence of additional tumors in the training data set may impact the performance. However, as the true non-nodulous areas occur with a much higher frequency than falsely marked regions, this is probably compensated by the class imbalance effect⁵⁴. Another limitation is the use of absorption values derived from HU values: here, future work could further improve the simulation model by using a polychromatic spectrum with different kVp settings. Moreover, a limitation is that the number of different nodules used in the study still was low. This number could be increased, in order to have more variation between the different nodules. Also, the number of healthy cases was low. As these were generated from healthy CT scans, only one X-ray image per scan was generated in order to avoid duplicate radiographs.

Since a simulation does never completely resemble the real setting, the radiologist's performance may be slightly worse than an evaluation on real radiographs. Here, the expression of effects such as the silhouette sign or differences in the mediastinal area between upright and supine patient positions could play a role.

Conclusion

In this study, we presented a framework that generated realistic looking radiographs by inserting nodules into existing CT volumes. The radiographs generated by the framework were used to train multiple CNNs and to evaluate the CNN performance against radiologists. We found our method to be adequate for initial CNN and observer performance evaluation. Thus, it could serve as an additional performance indicator for CNNs, as, contrary to manually annotated data, the groundtruth segmentations are perfectly accurate. Furthermore, our method allows to find positions where observers have problems identifying nodules: we found critical positions in paramediastinal positions.

Data availability

The .xlsx table file used for R/Jafroc evaluation is available as supplementary material. The Luna16 dataset used for training is available on²⁸.

Received: 26 August 2020; Accepted: 15 July 2021

Published online: 04 August 2021

References

- Bray, F. et al. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* **68**, 394–424. <https://doi.org/10.3322/caac.21492> (2018).
- Manser, R. et al. Screening for lung cancer. *Cochrane Database Syst. Rev.* <https://doi.org/10.1002/14651858.CD001991.pub3> (2013).
- van Beek, E. J. Lung cancer screening: Computed tomography or chest radiographs?. *World J. Radiol.* **7**, 189. <https://doi.org/10.4329/wjrv.7.i8.189> (2015).
- de Koning, H. J. et al. Reduced lung-cancer mortality with volume CT screening in a randomized trial. *N. Engl. J. Med.* **382**, 503–513. <https://doi.org/10.1056/NEJMoa1911793> (2020).
- Mettler, F. A., Huda, W., Yoshizumi, T. T. & Mahesh, M. Effective doses in radiology and diagnostic nuclear medicine: A catalog. *Radiology* **248**, 254–263. <https://doi.org/10.1148/radiol.2481071451> (2008).
- The National Lung Screening Trial Research Team. Reduced lung-cancer mortality with low-dose computed tomographic screening. *N. Engl. J. Med.* **365**, 395–409. <https://doi.org/10.1056/NEJMoa1102873> (2011).
- Mohammed, T. L. H. et al. ACR appropriateness criteria* screening for pulmonary metastases. *J. Thorac. Imaging* **26**, W1–W3. <https://doi.org/10.1097/RTI.0b013e3182010bf9> (2011).
- Davidson, R. S., Nwogu, C. E., Brentjens, M. J. & Anderson, T. M. The surgical management of pulmonary metastasis: Current concepts. *Surg. Oncol.* **10**, 35–42. [https://doi.org/10.1016/s0960-7404\(01\)00013-5](https://doi.org/10.1016/s0960-7404(01)00013-5) (2001).
- Stella, G. M., Kolling, S., Benvenuti, S. & Bortolotto, C. Lung-seeking metastases. *Cancers* **11**, 1–18. <https://doi.org/10.3390/cancers11071010> (2019).
- Schalekamp, S. et al. Computer-aided detection improves detection of pulmonary nodules in chest radiographs beyond the support by bone-suppressed images. *Radiology* **272**, 252–261. <https://doi.org/10.1148/radiol.14131315> (2014).
- Li, F., Engelmann, R., Metz, C. E., Doi, K. & MacMahon, H. Lung cancers missed on chest radiographs: Results obtained with a commercial computer-aided detection program. *Radiology* **246**, 273–280. <https://doi.org/10.1148/radiol.2461061848> (2008).
- Wang, X. et al. ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings—30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017* **2017-Janua**, 3462–3471, <https://doi.org/10.1109/CVPR.2017.369> (2017). [arXiv:1705.02315](https://arxiv.org/abs/1705.02315).
- Rajpurkar, P. et al. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. 3–9, (2017). [arXiv:1711.05225](https://arxiv.org/abs/1711.05225).
- Ausawalaithong, W., Marukatat, S., Thirach, A. & Wilaiprasitporn, T. Automatic Lung Cancer Prediction from Chest X-ray Images Using Deep Learning Approach. (2018). [arXiv:1808.10858](https://arxiv.org/abs/1808.10858).

15. Geras, K. J., Wolfson, S., Kim, S. G., Moy, L. & Cho, K. High-Resolution Breast Cancer Screening with Multi-View Deep Convolutional Neural Networks. 1–7 (2017). [arXiv:1703.07047](https://arxiv.org/abs/1703.07047).
16. Tang, Y., Tang, Y., Xiao, J. & Summers, R. M. XLSor: A Robust and Accurate Lung Segmentor on Chest X-Rays Using Criss-Cross Attention and Customized Radiorealistc Abnormalities Generation. 457–467 (2019). [arXiv:1904.09229](https://arxiv.org/abs/1904.09229).
17. Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* **9351**, 234–241. https://doi.org/10.1007/978-3-319-24574-4_28 (2015) [arXiv:1505.04597](https://arxiv.org/abs/1505.04597).
18. McKinney, S. M. *et al.* International evaluation of an AI system for breast cancer screening. *Nature* <https://doi.org/10.1038/s41586-019-1799-6> (2020).
19. Pan, I., Cadrin-Chênevert, A. & Cheng, P. M. Tackling the radiological society of North America pneumonia detection challenge. *Am. J. Roentgenol.* **213**, 568–574. <https://doi.org/10.2214/AJR.19.21512> (2019).
20. Lin, T.-Y., Goyal, P., Girshick, R., He, K. & Dollár, P. Focal Loss for Dense Object Detection. <https://doi.org/10.1016/j.jado.2005.02.022> (2017). [arXiv:1708.02002](https://arxiv.org/abs/1708.02002).
21. von Schacky, C. E. *et al.* Development and validation of a multitask deep learning model for severity grading of hip osteoarthritis features on radiographs. *Radiology* **295**, 136–145. <https://doi.org/10.1148/radiol.2020190925> (2020).
22. Abadi, E. *et al.* Virtual clinical trials in medical imaging: A review. *J. Med. Imaging* **7**, 1. <https://doi.org/10.1117/1.JMI.7.4.042805> (2020).
23. Abadi, E. *et al.* DukeSim: A realistic, rapid, and scanner-specific simulation framework in computed tomography. *IEEE Trans. Med. Imaging* **38**, 1457–1465. <https://doi.org/10.1109/TMI.2018.2886530> (2019).
24. Barufaldi, B., Bakic, P. R., Higginbotham, D. & Maidment, A. D. A. OpenVCT: A GPU-accelerated virtual clinical trial pipeline for mammography and digital breast tomosynthesis. In *Medical Imaging 2018: Physics of Medical Imaging* (eds Chen, G.-H. *et al.*) 194 (SPIE, Berlin, 2018). <https://doi.org/10.1117/12.2294935>.
25. Yu, L. *et al.* A virtual clinical trial using projection-based nodule insertion to determine radiologist reader performance in lung cancer screening CT. In *Proc. SPIE 10132, Medical Imaging 2017: Physics of Medical Imaging*, 101321R, <https://doi.org/10.1117/12.2255593> (2017).
26. Chakraborty, D. P. RJaFroc, <https://github.com/dpc10ster/RJaFroc>, Accessed 23 March 2021.
27. Hunter, J. D. Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.* **9**, 90–95. <https://doi.org/10.1109/MCSE.2007.55> (2007).
28. Luna 16 Dataset. <https://luna16.grand-challenge.org/data/>. Accessed 16 Jan 2020.
29. Milletari, F., Navab, N. & Ahmadi, S.-A. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. In *IEEE International Conference on 3D Vision* [arXiv:1606.04797](https://arxiv.org/abs/1606.04797) 1–11 (2016).
30. Liao, F., Liang, M., Li, Z., Hu, X. & Song, S. Evaluate the malignancy of pulmonary nodules using the 3D deep leaky noisy-or network. **14**, 1–12, <https://doi.org/10.1109/TNNLS.2019.2892409> (2017). [arXiv:1711.08324](https://arxiv.org/abs/1711.08324).
31. Lisowska, A., Beveridge, E., Muir, K. & Poole, I. Thrombus detection in CT brain scans using a convolutional neural network. *Proceedings of the 10th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2017)* 24–33, <https://doi.org/10.5220/0006114600240033> (2017).
32. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. <https://doi.org/10.1109/CVPR.2016.90> (2015). [arXiv:1512.03385](https://arxiv.org/abs/1512.03385).
33. Abadi, M. *et al.* TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems (2016). [arXiv:1603.04467](https://arxiv.org/abs/1603.04467).
34. Chollet, F. *et al.* Keras. (2015). <https://github.com/fchollet/keras>. Accessed 7 Dec 2018.
35. Gaiser, H. Keras-Retinanet. <https://github.com/fizyr/keras-retinanet>, <https://doi.org/10.5281/zenodo.1188105>. Accessed 2 Jan 2020.
36. Egan, J. P., Greenberg, G. Z. & Schulman, A. I. Operating characteristics, signal detectability, and the method of free response. *J. Acoust. Soc. Am.* **33**, 993–1007. <https://doi.org/10.1121/1.1908935> (1961).
37. Bunch, P. C., Hamilton, J. F., Sanderson, G. K. & Simmons, A. H. A free response approach to the measurement and characterization of radiographic observer performance. In *Proc. SPIE 0127, Application of Optical Instrumentation in Medicine VI*. <https://doi.org/10.1117/12.955926> (1977).
38. Chakraborty, D. P. & Zhai, X. On the meaning of the weighted alternative free-response operating characteristic figure of merit. *Med. Phys.* **43**, 2548–2557. <https://doi.org/10.1118/1.4947125> (2016).
39. Chakraborty, D. P. & Berbaum, K. S. Observer studies involving detection and localization: Modeling, analysis, and validation. *Med. Phys.* **31**, 2313–2330. <https://doi.org/10.1118/1.1769352> (2004).
40. Chakraborty, D. P. Observer performance methods for diagnostic imaging: Foundations, modeling, and applications with R-based examples. *Imaging in Medical Diagnosis and Therapy* (CRC Press, 2017).
41. Kim, Y. G. *et al.* Short-term reproducibility of pulmonary nodule and mass detection in chest radiographs: Comparison among radiologists and four different computer-aided detections with convolutional neural net. *Sci. Rep.* **9**, 1–9. <https://doi.org/10.1038/s41598-019-55373-7> (2019).
42. Shapira, N. *et al.* Liver lesion localisation and classification with convolutional neural Networks: A comparison between conventional and spectral computed tomography. *Biomed. Phys. Eng. Express* <https://doi.org/10.1088/2057-1976/ab6e18> (2020).
43. Dubost, F. *et al.* Gp-Unet: Lesion detection from weak labels with a 3D regression network. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 10435 LNCS, 214–221, https://doi.org/10.1007/978-3-319-66179-7_25 (2017). [arXiv:1705.07999](https://arxiv.org/abs/1705.07999).
44. Prakash, A. *et al.* Structured Domain Randomization: Bridging the Reality Gap by Context-Aware Synthetic Data. [arXiv:1810.10093v1](https://arxiv.org/abs/1810.10093v1) (2018).
45. Tremblay, J. *et al.* Training deep networks with synthetic data: Bridging the reality gap by domain randomization. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops* 2018–June, 1082–1090, <https://doi.org/10.1109/CVPRW.2018.00143> (2018). [arXiv:1804.06516](https://arxiv.org/abs/1804.06516).
46. Tobin, J. *et al.* Domain randomization for transferring deep neural networks from simulation to the real world. In *IEEE International Conference on Intelligent Robots and Systems* 2017–Sept, 23–30, <https://doi.org/10.1109/IROS.2017.8202133> (2017). [arXiv:1703.06907](https://arxiv.org/abs/1703.06907).
47. Toth, D., Cimen, S., Ceccaldi, P., Kurzendorfer, T., Rhode, K. & Mountney, P. Training deep networks on domain randomized synthetic X-ray data for cardiac interventions. *Proc. Mach. Learn. Res.* **102**, 468–482, (2019).
48. Yamanaka, J., Kuwashima, S. & Kurita, T. Fast and Accurate Image Super Resolution by Deep CNN with Skip Connection and Network in Network. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 10635 LNCS, 217–225, https://doi.org/10.1007/978-3-319-70096-0_23 (2017). [arXiv:1707.05425](https://arxiv.org/abs/1707.05425).
49. Umehara, K. *et al.* Super-resolution convolutional neural network for the improvement of the image quality of magnified images in chest radiographs. In *Proc. SPIE 10133, Medical Imaging 2017: Image Processing*, 101331P, <https://doi.org/10.1117/12.2249969> (2017).
50. Vogelstein, B. *et al.* A spatial model predicts that dispersal and cell turnover limit intratumour heterogeneity. *Nature* **525**, 261–264. <https://doi.org/10.1038/nature14971> (2015).
51. Chakraborty, D. P. *et al.* Digital and conventional chest imaging: A modified ROC study of observer performance using simulated nodules. *Radiology* **158**, 35–39. <https://doi.org/10.1148/radiology.158.1.3940394> (1986).

52. Monnier-Cholley, L. *et al.* Characteristics of missed lung cancer on chest radiographs: A French experience. *Eur. Radiol.* **11**, 597–605. <https://doi.org/10.1007/s003300000595> (2001).
53. de Groot, P. M., Carter, B. W., Abbott, G. F. & Wu, C. C. Pitfalls in chest radiographic interpretation: Blind spots. *Semin. Roentgenol.* **50**, 197–209. <https://doi.org/10.1053/j.ro.2015.01.008> (2015).
54. Buda, M., Maki, A. & Mazurowski, M. A. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks* **106**, 249–259. <https://doi.org/10.1016/j.neunet.2018.07.011> (2017) [arXiv:1710.05381](https://arxiv.org/abs/1710.05381).

Author contributions

F.P. D.P. designed the research study. F.P. D.P. and M.M. supervised the Project. M.S. drafted the manuscript. P.S., M.S. and F.D. generated the simulated radiographs and analysed the data. J.B., J.A., C.M.L., F.G.G., F.T.G., J.F.G., E.C.H., D.S., C.E.v.S., S.Z. participated in the initial and/or the revised reader study. M.S. and P.S. implemented network training pipelines for RetinaNet and U-Net. M.S. developed the tools used to perform the reader study. All authors reviewed the manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-021-94750-z>.

Correspondence and requests for materials should be addressed to M.S.

Reprints and permissions information is available at www.nature.com/reprints.

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



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021