Automating chest radiograph imaging quality control

Katri Nousiainen^{a,b}, Teemu Mäkelä^{a,b}, Anneli Piilonen^a, Juha I. Peltonen^{a,*}

^a HUS Medical Imaging Center, Radiology, University of Helsinki and Helsinki University Hospital, P.O. Box 340, FI-00029 HUS, Helsinki, Finland

^b Department of Physics, University of Helsinki, P.O. Box 64, FI-00014 Helsinki, Finland

ARTICLE INFO

Keywords:

Artificial intelligence
Quality control
Chest radiography

ABSTRACT

Purpose: To automate diagnostic chest radiograph imaging quality control (lung inclusion at all four edges, patient rotation, and correct inspiration) using convolutional neural network models.

Methods: The data comprised of 2589 postero-anterior chest radiographs imaged in a standing position, which were divided into train, validation, and test sets. We increased the number of images for the inclusion by cropping appropriate images, and for the inclusion and the rotation by flipping the images horizontally. The image histograms were equalized, and the images were resized to a 512×512 resolution. We trained six convolutional neural networks models to detect the image quality features using manual image annotations as training targets. Additionally, we studied the inter-observer variability of the image annotation.

Results: The convolutional neural networks' areas under the receiver operating characteristic curve were >0.88 for the inclusions, and >0.70 and >0.79 for the rotation and the inspiration, respectively. The inter-observer agreement between two human annotators for the assessed image-quality features were: 92%, 90%, 82%, and 88% for the inclusion at patient's left, patient's right, cranial, and caudal edges, and 78% and 89% for the rotation and inspiration, respectively. Higher inter-observer agreement was related to a smaller variance in the network confidence.

Conclusions: The developed models provide automated tools for the quality control in a radiological department. Additionally, the convolutional neural networks could be used to obtain immediate feedback of the chest radiograph image quality, which could serve as an educational instrument.

Introduction

Quality control (QC) is an essential part of diagnostic radiography. Technical QC includes monitoring electrical, mechanical and radiation safety, and verifying the manufacturer specified normal operation of the imaging device. The acceptance criteria can be defined with a relative ease and the compliance either directly measured or otherwise explicitly noted. Imaging QC consists of proper selection and optimization of the imaging parameters, active image quality evaluation for specific clinical conditions, and adherence to good imaging practices. These tasks are generally more subjective than in technical QC and the practices may vary between, or even within, hospitals. In contrast to technical QC, defining acceptable operational limits in imaging QC is less straightforward and tools for automation and management are more limited. Performing imaging QC is still often highly labor-intensive and requires expert evaluations with significant inter-observer variations [1,2]. This restriction substantially limits the number of imaging studies included in systematic QC evaluations decreasing the applicability of the results and

hindering effective follow-ups.

Considering chest radiographs, patient alignment and proper inspiration are among key factors in acquiring images optimal for reading. While the lungs need to be fully included, excessively large fields-of-view should be avoided. The latter increases patient's radiation dose and worsens image quality due to X-ray scattering while offering no additional diagnostic information. Guidelines outlining acceptable diagnostic image quality exist for typical projections [3–5]. Adopting common evaluation guidelines and consistent scoring allows the use of multiple observers and facilitates comparisons between imaging sites. However, due to the highly subjective nature of the visual image assessment, only moderate inter-observer agreements have been reported [6,7].

During recent years, convolutional neural networks (CNN) have shown huge potential in computer vision and image analysis [8]. The CNN-based deep-learning approaches have been successfully applied in practically all imaging modalities for example in pathology detection, classification, segmentation, and image reconstruction [9]. Several

* Corresponding author.

E-mail address: juha.peltonen@hus.fi (J.I. Peltonen).

<https://doi.org/10.1016/j.ejmp.2021.03.014>

Received 2 December 2020; Received in revised form 15 February 2021; Accepted 8 March 2021

Available online 23 March 2021

1120-1797/© 2021 Associazione Italiana di Fisica Medica. Published by Elsevier Ltd. All rights reserved.

Fig. 1. a) The graphical annotator user-interface. The red lines mark two-centimeter distance from the image edges. b) A bird-eye's view of an appropriate (left) and an exaggeratedly not appropriate patient rotation (right). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

studies have demonstrated CNN's applicability to chest radiographs for various clinical tasks [10–12], and for image quality assessments [13]. Outsourcing image quality grading to an artificial intelligence offers many benefits: decrease in manual work, reduction or removal of inter- and intra-observer variations in image interpretation, report standardization, expansion of the QC coverage up to all images acquired, and considerable reduction in the results delay. An instant feedback

following the image acquisition could further be employed in operator training and improving the quality of care.

The aim of this study was to develop convolutional neural network models to assess three quality criteria from posterior-anterior (PA) chest radiographs: correct inclusion of the lungs, patient rotation, and inspiration. We chose the set of studied features based on previous QC efforts to reflect identified targets of development at our imaging center. We

Fig. 2. A representation of the annotation criteria for a) left, right, and caudal edges where the red lines denote the two-centimeter distance from the image edges, b) cranial edge where the first thoracic vertebra is highlighted, c) inspiration where the sixth and tenth rib are highlighted, and d) rotation where the medial ends of the collarbones are highlighted. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Fig. 3. Preprocessing. a) Image where the anatomical dimensions on the right side of the image (patient's left) are preserved and squeezed on the opposite side. b) The image cropped at the right side of the image (patient's left) and stretched on the opposite side, c) image flipped and the right side of the image (patient's right) preserved. d) The cropped image enclosing the medial ends of the collarbones. The image histograms are equalized, and the images resized to a resolution of 512×512 pixels.

created an image annotation tool to aid in the visual assessment of the large datasets and compared the network's decision confidence to human inter-observer agreement. We trained a separate CNN model for each of the features (i.e. four edges of lung inclusion, patient rotation and inspiration).

Material and Methods

Study population

In this study, a total of 2589 PA standing-position adult chest radiographs were retrospectively acquired from our hospital district's picture archiving and communication system, and the images were subsequently anonymized. The data comprised of two image datasets; Dataset A consisted of 2019 consecutive chest radiographs from a single ceiling X-ray system (Samsung Electronics Co., Ltd., Suwon, South Korea). Dataset B consisted of 570 chest radiographs from 44 ceiling X-ray systems manufactured by ten different vendors. The study received research permit from HUS Medical Imaging Center of Helsinki University Hospital (HUS/628/2019). Patient informed consent was waived due to the retrospective nature of this study.

Annotation

We used the European Commission's guidelines on quality criteria for diagnostic radiographic images [4] as our baseline for the annotations. We created a graphical user-interface with MATLAB (The MathWorks, Inc., Natick, MA, USA) for rapid annotation of the images (Fig. 1a). We divided the lung inclusion into four separate tasks: the (patient's) left, right, caudal, and cranial edges. The inclusion was marked either excessive, appropriate or insufficient separately for each

four edges, and the inspiration and the rotation (Fig. 1b) as either appropriate or not appropriate.

Initially, a resident medical physicist (KN) and a medical physicist (JIP) each annotated Dataset B based on the European Commission guidance [4]. We compared the annotations to obtain inter-observer statistics. The observers reviewed roughly 20 of these images with a senior thoracic radiologist (AP) to unify the annotation criteria. Resulting final criteria was as follows: at left and right edges, the inclusion was insufficient if a lung was cut, excessive if there was more than two centimeters of spare space outside the ribcage, and otherwise appropriate (Fig. 2a). At the cranial edge, the inclusion was insufficient if the first thoracic vertebra was not fully visible and excessive if the sixth cervical vertebra was fully visible (Fig. 2b). At the caudal edge, the inclusion was insufficient if the caudal tips of the lungs were cut, and excessive if there was more than two centimeters of spare space below the tips. The inspiration was appropriate if both the anterior side of the sixth rib and the posterior side of the tenth rib were visible over the right lung (Fig. 2c). The rotation was appropriate, if the medial ends of the collarbones were symmetrical around the spine (Fig. 2d). Finally, KN annotated all the images in Dataset A.

Preprocessing

From Dataset A, we randomly extracted 3.5% and 1.5% of the images for validation and testing of inclusion (same images for each edge), and 10% and 5% for validation and testing of inspiration and rotation. Such proportions resulted in test sets of roughly equal in size for different features, after we had artificially increased the number of images for inclusion by image cropping as described below. The validation and test images (including Dataset B) went through the same preprocessing steps as the test data, but the steps varied between inspected features. Dataset

Table 1

Number of images per class for inclusion.

Edge	Data	Insufficient	Appropriate	Excessive	Total
Cranial	Training	3322	2612	502	6436
	Validation	56	44	18	118
	Testing	116	88	24	228
Caudal	Training	454	396	3370	4420
	Validation	26	20	48	94
	Testing	20	20	120	160
Left and Right	Training	1300	1277	2524	5101
	Validation	39	36	35	110
	Testing	51	50	89	190

Table 2

Number of images per class for inspiration and rotation.

Feature	Data	Appropriate	Not appropriate	Total
Inspiration	Training	1293	413	1706
	Validation	87	23	110
	Testing	149	53	202
Rotation	Training	1854	1558	3412
	Validation	116	104	220
	Testing	110	92	202

B served as an external test set to investigate models' generalization capabilities.

For all the images, when necessary (i.e. if the data was stored using inverted grayscale map), the image intensity was inverted and, if present, the black borders (i.e. digital zero padding) were removed, after which the images had 1832–3037 pixels per column and 2058–3039 pixels per row.

In the initial annotated data, images with insufficient inclusion were underrepresented for left, right, and caudal edges. Hence, we increased the amount of the image data for the inclusion by cropping the appropriate-inclusion images by four centimeters at the examined edge to meet the insufficient criterion. Additionally, we doubled the train and validation data size by flipping the images horizontally. We trained a separate network for each edge. All the network inputs were of fixed size, which required scaling of the image. To keep relevant physical dimensions unchanged, images were divided in two depending on the inspected edge. The half containing the target edge remained untouched and the other half was stretched or squeezed in perpendicular direction to the edge to reach square aspect ratio using bilinear interpolation. See Fig. 3a–c for inclusion preprocessing steps.

For rotation we cropped the images so that they contained the area around the medial ends of the collarbones, by keeping an area of five centimeters on both sides around the central line at the top half of the

Fig. 4. The inclusion ROC curves and AUC for the test images of Dataset A. AUC 95% confidence intervals are in parentheses.

Fig. 5. The inspiration and rotation ROC curves, AUC (with 95% confidence intervals) for the test images of Dataset A.

image (Fig. 3d). We doubled the train and validation sets by flipping the cropped images horizontally.

With inspiration, no feature-specific preprocessing steps were applied.

Finally, the image contrasts were standardized using histogram equalization (*equalize_hist* function in scikit-image [14]) and the images were resized to a 512×512 -pixel resolution (*resize* function in OpenCV [15]) for all the features: inclusion, rotation, as well as inspiration. Table 1 and 2 show the final number of images per class after preprocessing.

Training

We trained a variety of ResNet50 [16] and DenseNet121 [17] networks without the pre-trained weights using Python version 3.7, Keras version 2.2.4 [18] and TensorFlow version 1.14 [19]. We tested different learning rates, batch sizes, class balancing, augmentation, dropout, and pooling schemes. The best performing architectures, which we chose based on the validation accuracy, are described below for each task.

With the left and right edges, our ResNet50 had 32 filters and did not use initial padding; it used dropout of 0.2 and global average pooling. During training we used sparse categorical cross-entropy loss function, and Adam optimizer with learning rate of 0.001. With the cranial and caudal edges, our DenseNet121 (four dense layers, with 6, 12, 24, and 16 blocks) had 32 filters, no initial padding, and it used dropout of 0.2 and global maximum pooling, with learning rate 0.001. For rotation and inspiration, we used a similar DenseNet121, but without the dropout.

Based on preliminary testing we used two class balancing schemes for the inclusions: majority class resampling for left, right, and caudal; all-but-minority class resampling for cranial. In addition to previous preprocessing steps, we augmented the inclusion images by randomly rotating them from -5 to $+5$ degrees during training. For rotation and inspiration, we did not use random rotation or class balancing. For ResNet50 (in left and right edges), our batch size was 32, otherwise 16. In each case, the training lasted for 15–25 epochs.

Testing

With the resulting neural networks, we classified the test images from Dataset A, and the complete Dataset B. To assess the potential transfer of observer uncertainty (i.e. the difficulty of interpretation) to the final models, we compared the network confidences (i.e. class probabilities) of the cases in Dataset B, where the two observers had agreed, to those cases, where they had disagreed.

The area under the receiver operating characteristics (ROC) curve (AUC) results were considered excellent for AUC values between 0.9 and 1.0, good for AUC values between 0.8 and 0.9, fair for AUC values between 0.7 and 0.8, poor for AUC values between 0.6 and 0.7 and failed for AUC values between 0.5 and 0.6. We estimated the internal variation in the evaluation metrics by calculating 95% confidence intervals (CI) using a bootstrapping method, where resampling with replacement was repeated 10^5 times.

Results

With the test images of Dataset A, the ROC AUCs were >0.92 for the inclusion detection of all four edges (Fig. 4), and 0.72 and 0.90 for the rotation and the inspiration, respectively (Fig. 5). With Dataset B, AUCs were >0.88 for the inclusion in all four edges (Fig. 6), and 0.70 and 0.79 for the rotation and the inspiration (Fig. 7), respectively. The figures also show 95% CIs. The models performed better with Dataset A than with Dataset B: there was a statistical difference between the datasets for eight out of twelve inclusion classes and for inspiration.

The inter-observer agreement for the 570 images in Dataset B for each assessed feature were: 92.3%, 90.0%, 82.3%, and 87.9% for the inclusion at patient's left, patient's right, cranial, and caudal edges, and 78.1% and 89.3% for the rotation and inspiration, respectively. The distributions of the neural network confidence had lower median values and wider range of the 25th–75th percentile, when the two observers disagreed on cases (Fig. 8).

The processing time per image was approximately six seconds on a CPU. This included the preprocessing and the neural network predictions.

Discussion

In this study, we used a deep-learning-based approach to assess the clinical quality of chest radiographs. Deep learning is an attractive approach to reduce the workload of human observers. It could solve problems originating from limited inter-observer agreement and increase the number of studies included in QC evaluations. Modern CNNs perform well in extracting complex features from radiological images and classifying data based on clinically relevant factors. CNNs offer an intriguing possibility to integrate a large variety of automated QC tests into existing monitoring applications and processes.

Our study reached four major conclusions. 1) Using CNNs, which have been successful in many clinical tasks, is also feasible for automating imaging QC. 2) Estimating inclusion was an easier task than identifying correct rotation or inspiration. This reflects CNNs' power to identify local features over combining information from spatially separated image locations. Inclusion focuses fully on the specific lung edge, whereas the visual scoring criteria for inspiration and rotation depend on more complicated logic. 3) Scoring ambiguity (inter-observer variability) was reflected in the prediction confidences (probabilities). The network confidences had greater variances and lower median values when the observers disagreed compared to the cases where they had agreed on. Some bias in performance is expected as the model is most likely to concur with the observer whose annotations it was trained on. However, as can be seen in Fig. 8, the networks' decision confidences seem to reflect likewise difficulty of the task for the human observers. This could be from objective difficulty to judge certain images, likely to produce randomness as well as systematic bias between observers, or

Fig. 6. The inclusion ROC curves and AUC for the test images of Dataset B. AUC 95% confidence intervals are in parentheses.

from the observer's uncertainty or inconsistency (training label noise). The effect was less pronounced in the more difficult, and overall more subjective, tasks of inspiration and rotation than in the inclusions. 4) The CNN models were able to generalize to unseen data. Although the performance was lower in Dataset B than in Dataset A, the absolute differences were generally small. The models were trained on single-device images to minimize training data variation and to demonstrate the feasibility of the methodology. The models showed good, although mostly significantly lower, performances on the heterogeneous Dataset B. This underlines the well-known problem of deep learning overfitting, unpredictable generalization, and the importance of external validation. In the current study design, we were not able to discern if the lower performance resulted from overfitting that could have been managed during training, or from the underlying differences between the datasets. Using validation images from different sites during training could address this issue.

Data annotation is often the most important task in the artificial intelligence training. Adhering to clear guidelines can improve annotation consistency but does not eliminate all the ambiguity, especially for borderline cases. In our study, we reviewed a subset of images in Dataset B to clarify the annotation criteria, after which a single observer annotated Dataset A. With this process, we tried to minimize the noise in the final training labels. We created a custom-build interface to allow

effortless annotation. The interface provided graphical guidance of the inclusion criteria to further minimize the inter- and intra-observer variability.

Input normalization is a non-trivial task in neural network development. We used several preprocessing steps to make the image sets uniform and to emphasize the desired features of interest. Histogram equalization normalizes the contrast while maintaining desired anatomical features. We also included horizontally flipped images for inclusion and rotation data sets. Although the transformation is not anatomically trivial due to organ asymmetries, we chose this approach to discourage the usage of certain image components, such as heart, large arteries, and possible orientation markings, which we had assumed to be irrelevant for the task. For example, the inclusion models should primarily focus on the ribcage-lung borders and the model should therefore perform equally well for normal and flipped images. Focusing the networks' attention could also hinder possible detrimental effects from anatomical variation.

Due to the chosen neural network architecture and training by batches, the input images need to be of the same size. In addition, the chosen criteria for inclusion were based on exact distances. Therefore, the resizing operations to reach square aspect ratio was carried out while preserving physical dimensions in the region of interest. In classifying the rotations, the images were cropped to hinder superfluous

Fig. 7. The inspiration and rotation ROC curves, AUC (with 95% confidence intervals) for the test images of Dataset B.

correlations resulting from random and irrelevant image features. We believe that it is important to understand both the properties of the CNN and the clinical task to produce an effective preprocessing workflow. This aspect is especially emphasized if the amount of training data is limited with increased possibility of overfitting and potential problems with generalization. Improving augmentation and training strategies is an important and active area of deep-learning research.

We applied different preprocessing steps and trained separate networks for each task to maximize the models' representational power. We chose the ResNet50 [16] and DenseNet121 [17] network architectures initially because of their reportedly excellent performance on image classification tasks, indicating insensitivity to image noise. Although noise was not a key feature in this work, it is an inseparable part of any radiological image and the noise level and distribution vary between radiographs. Thus, a network should be able to separate it effectively from the relevant anatomical features. The aforementioned architectures encourage feature reuse and propagation and facilitate deep networks without explosion in the number of trainable parameters. We chose the final optimized architectures task-by-task based on the

validation accuracy of images from the single X-ray device. Although we used a different model for each task to meet the research aims, it is possible that a similar performance could be achieved by a single model with standard preprocessing. This alternative approach could benefit from sharing features between tasks, and expedite training and inference.

We evaluated the inclusion performance solely based on the independent AUC values and did not choose any representative (or a combination of) operating points on the ROC curves. The performance for the appropriate inclusion was from good to excellent. The AUCs for all the edges were >0.92 with Dataset A and >0.88 with Dataset B providing very convincing ability to detect these features automatically. Deployment into practice would require fixed operating points to be defined, based on the judgement call for the sensitivity–specificity tradeoff. This choice would preferably be based on a multi-device image set.

The inspiration detection performance on Dataset A test images was excellent especially considering the annotation criteria and the inter-observer variability. There is no golden standard for determining correct inspiration from images that would cover all differences between anatomies. In addition to a possible overfitting, the underlying uncertainty and interpretation complexity may contribute to the significantly lower AUC in Dataset B: the inspiration model had a greater difficulty to generalize than the models for the other features.

The rotation detection resulted in the lowest AUC value among the studied parameters: 0.72 with Dataset A and 0.70 with Dataset B. This was considered a fair performance. The rotation had also the lowest inter-observer score due to the subjective annotation criteria, and the widest interquartile range in the network confidence. It is likely that the CNN model can detect coarse deviations from the optimal patient alignment, but minor misalignment results in ambiguous outcome. The results could possibly be improved by an approach with intermediate anatomical segmentations detecting the collarbone location relative to the spine.

When comparing Dataset B from 44 X-ray devices with Dataset A from a single X-ray machine the AUC values differed no >0.06 for the inclusions and the rotation, and by 0.11 for the inspiration. We consider these differences relatively minor and conclude that the chosen architectures and the applied preprocessing steps were moderately insensitive to confounding factors from patient- and device-specific differences. Regarding the manual annotation, minimal variation in image outlook of Dataset A (one X-ray device and uniform image contrast) may also improve consistency.

The annotation criteria used in our study for the different features, despite being widely approved, are rather strict for clinical use. The proportion of the appropriate images in our datasets are minor

Fig. 8. The box plots of network confidence distributions for the convolutional neural networks when two observers agreed or disagreed on a case.

compared to the inappropriate images in respect of most of the features. For operator-teaching purposes, and to better match clinical requirements, it would be beneficial to include an additional category representing clinically acceptable images. This category would indicate when no re-exposure is needed while the image still has room for improvement.

Additionally, the use of strict annotation criteria leads to a binary classification problem, which may not be optimal when considering borderline cases: miniscule feature changes can lead to a change in the target class. Using continuous criteria would hinder this problem but creating such training targets may become prohibitively expensive. In practice, the networks could be trained with intermediate tasks (e.g. lung and ribcage segmentation) from which the adherence to acceptance criteria could be calculated. This approach might not be suitable in all situations, for example detecting insufficient inclusion (a part of the lung that is not imaged cannot be segmented).

Another possible approach, presented by von Berg et al. [13], is to use the combination of CNN and anatomical atlases to directly detect and segment anatomical features in the image. This approach allows the user to change the QC criteria afterwards without a need to train CNN again. The discrete-decision training approach we have presented requires carefully pre-selected annotation criteria and may suffer from imbalanced training data. Benefits, on the other hand, include easy applicability to wide range of studies and relatively easy generation of large training datasets.

We trained and evaluated the models solely on clinical images. This resulted in a natural data imbalance: not all features, error types and combinations were present in equal numbers. Data size and variety, for both training and testing, could be expanded by acquiring exposures of anthropomorphic chest phantoms. This would allow repeated acquisitions with different orientations, exposure settings, and image processing. The images could be utilized in estimating model's accuracy, in QC when developing new models, and verifying consistency with well-defined ground truths. One could also assess, to a limited extent, models' generalizability to new imaging devices without requiring access to patient data. Furthermore, besides physical phantoms with fixed geometry, use of adjustable digital anthropomorphic phantoms with simulated radiographs could offer various body models, inspiration phases, and postures.

The presented methodology can be used to obtain imaging statistic and could be used as a training and feedback tool. For example, it could provide information on the differences between imaging sites in reaching optimal image quality and help in analyzing these discrepancies. The effect of imaging time and date, device, or device setup could be studied. Additionally, the same models could provide an anonymous online response about the image quality for the operator.

Conclusions

We demonstrated that a CNN-based system can be used to detect if a chest radiograph follows relevant image quality criteria. Correspondence to human observer was good for inclusion and inspiration and

moderate for rotation. The system could be scaled to provide real-time QC statistics from all the images acquired within an imaging center. Similarly, it could act as a training tool providing immediate anonymous feedback to the operator.

Funding

This work was supported by HUS Medical Imaging Center [grant numbers TK20190001 and M1022TK906].

References

- [1] Reiner BI. Automating quality assurance for digital radiography. *J Am Coll Radio* 2009;6(7):486–90. <https://doi.org/10.1016/j.jacr.2008.12.008>.
- [2] Whaley JS, Pressman BD, Wilson JR, Bravo L, Sehnert WJ, Foos DH. Investigation of the variability in the assessment of digital chest X-ray image quality. *J Digit Imag* 2013;26:217–26. <https://doi.org/10.1007/s10278-012-9515-1>.
- [3] Herrmann TL, Fauber TL, Gill J, Hoffman C, Orth DK, Peterson PA, et al. Best practices in digital radiography. *Radiol Technol* 2012;84:83–9.
- [4] Carmichael JHE, Maccia C, Moores BM, Oestmann JW, Schibilla H, Teunen D, et al. Commission of the European Communities European Guidelines On Quality Criteria For Diagnostic Radiographic Images. Luxembourg: Office of Official Publications of the European Communities; 1996.
- [5] American College of Radiology, ACR-SPR-STR practice parameter for the performance of Chest radiography; 2017.
- [6] Tesselaaar E, Dahlström N, Sandborg M. Clinical audit of image quality in radiology using visual grading characteristics analysis. *Radiat Prot Dosimetry* 2016;169(1–4): 340–6. <https://doi.org/10.1093/rpd/ncv411>.
- [7] Decoster R, Toomey R, Butler ML. Do radiographers base the diagnostic acceptability of a radiograph on anatomical structures? *InMedical Imaging* 2018; 10577:1057703. <https://doi.org/10.1117/12.2293108>.
- [8] Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. Artificial intelligence in radiology. *Nat Rev Cancer* 2018;18(8):500–10. <https://doi.org/10.1038/s41568-018-0016-5>.
- [9] Yamashita R, Nishio M, Do RKG, Togashi K. Convolutional neural networks: an overview and application in radiology. *Insights Imaging* 2018;9(4):611–29. <https://doi.org/10.1007/s13244-018-0639-9>.
- [10] Qin C, Yao D, Shi Y, Song Z. Computer-aided detection in chest radiography based on artificial intelligence: a survey. *BioMed Eng OnLine* 2018;17:113. <https://doi.org/10.1186/s12938-018-0544-y>.
- [11] Zotin A, Hamad Y, Simonov K, Kurako M. Lung boundary detection for chest X-ray images classification based on GLCM and probabilistic neural networks. *Proc Comput Sci* 2019;159:1439–48. <https://doi.org/10.1016/j.procs.2019.09.314>.
- [12] Yan F, Huang X, Yao Y, Lu M, Li M. Combining LSTM and DenseNet for automatic annotation and classification of chest X-ray images. *IEEE Access* 2019;7:74181–9. <https://doi.org/10.1109/Access.628763910.1109/ACCESS.2019.2920397>.
- [13] von Berg J, Krönke S, Gooßen A, Bystrov D, Brück M, Harder T, et al. Robust chest x-ray quality assessment using convolutional neural networks and atlas regularization. *InMed Imag* 2020;11313:113131L. <https://doi.org/10.1117/12.2549541>.
- [14] Van der Walt S, Schönberger JL, Nunez-Iglesias J, Boulogne F, Warner JD, Yager N, Gouillart E, Yu T, et al. scikit-image: Image processing in Python. *PeerJ* 2014;2. <https://doi.org/10.7717/peerj.453>.
- [15] Bradski G. The OpenCV library. Dr. Dobb's J Softw Tools 2000.
- [16] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. *Proc CVPR* 2016:770–8. <https://doi.org/10.1109/cvpr.2016.90>.
- [17] Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. *Proc CVPR* 2017:4700–8.
- [18] Chollet F, Yee A, Keras PR. Deep learning for humans. GitHub Repos 2015.
- [19] Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J. Tensorflow: A system for large-scale machine learning. In: 12th USENIX symposium on operating systems design and implementation (OSDI 16); 2016. p. 265–83.