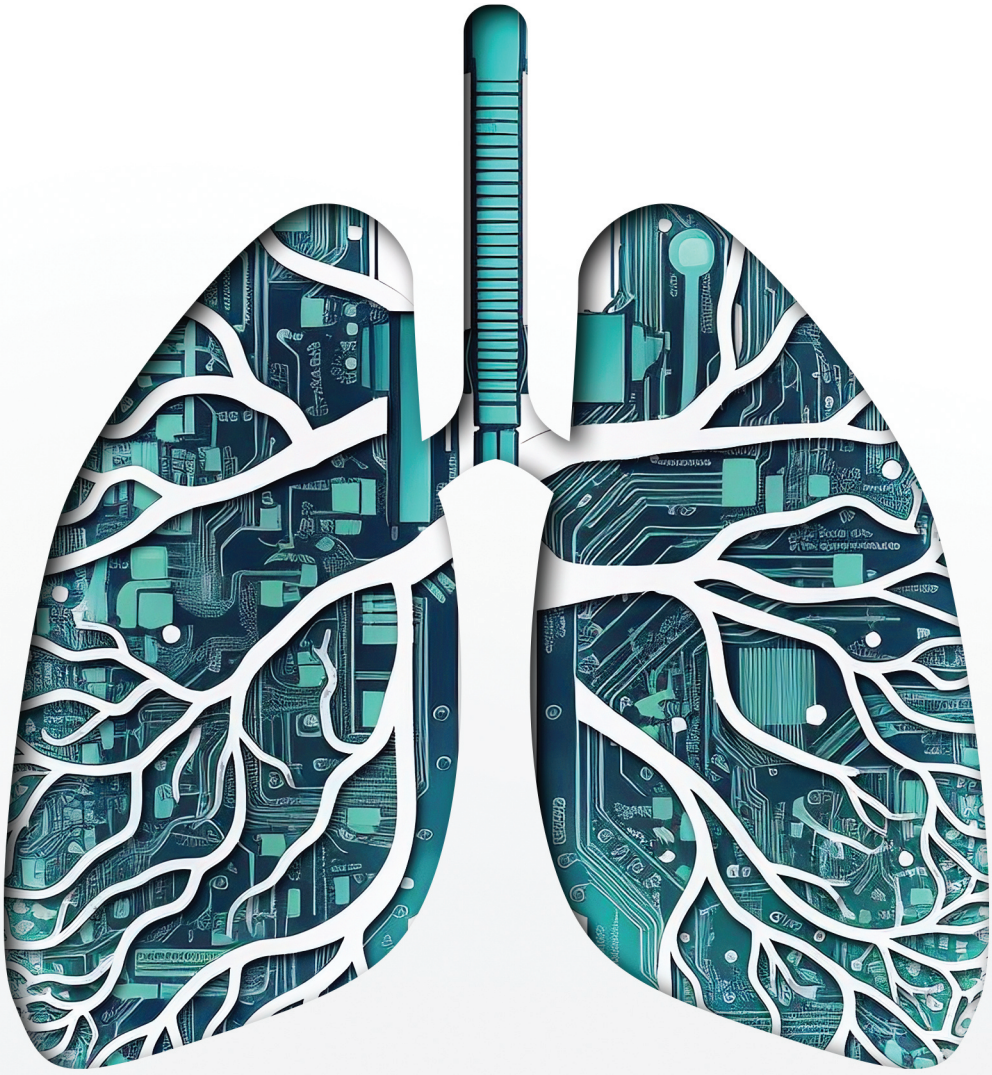
MACHINE LEARNING IN CARDIOTHORACIC RADIOLOGY

From Medical Data Curation to Clinical Application



Thomas Weikert

Machine Learning in Cardiothoracic Radiology -
from Medical Data Curation to Clinical Application

Thomas Weikert

Machine Learning in Cardiothoracic Radiology - from Medical Data Curation to Clinical Application

ISBN/EAN: 978-94-6469-384-3

This thesis was prepared at the Clinic of Radiology and Nuclear Medicine, University Hospital Basel, University of Basel, Switzerland and the Department of Radiology, University Medical Center Utrecht, Utrecht University, the Netherlands.

Cover: Andrea Kühne
Design and layout: Thomas Weikert
Printing: ProefschriftMaken

Machine Learning in Cardiothoracic Radiology - from Medical Data Curation to Clinical Application

Machinaal leren in de cardiothoracale radiologie - van medische gegevenscuratie tot klinische toepassing

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor aan de
Universiteit Utrecht
op gezag van de
rector magnificus, prof.dr. H.R.B.M. Kummeling,
ingevolge het besluit van het college voor promoties
in het openbaar te verdedigen op

maandag 5 juni 2023 des middags te 4.15 uur

door

Thomas Johannes Weikert

geboren op 26 juni 1984
te Wuerzburg, Duitsland

Promotoren:

Prof. dr. T. Leiner
Prof. dr. A.W. Sauter

Copromotor:

Dr. B. Stieltjes

Beoordelingscommissie:

Prof. dr. P.A. de Jong
Prof. dr. B.K. Velthuis
Prof. dr. M.C. Post (voorzitter)
Prof. dr. B. Baessler
Prof. dr. K. Maier-Hein

Contents

CHAPTER 1

Introduction

Introduction

The objective of this thesis is to create and investigate machine learning (ML) algorithms for a broad spectrum of tasks in the field of cardiothoracic radiology. This includes data curation, a fundamental step at the beginning of most scientific projects, exemplified by an algorithm for radiology report classification (Chapter 2). Furthermore, image recognition algorithms for detection (Chapters 3 & 4) and segmentation (Chapter 5) of cardiothoracic findings in cross-sectional imaging will be addressed, some of which are currently clinically deployed. Finally, a model using cardiothoracic imaging biomarkers extracted with ML to predict the clinical course of patients infected with SARS-CoV-2 is presented (Chapter 6). Thereby, the chapters of this thesis follow a chronologic sequence of tasks, beginning with data curation to finding detection to finding segmentation to application of extracted information for clinical decision support. Finally, Chapter 7 consolidates methodological insights gained during the abovementioned ML projects and provides a generalizable framework of best practices for creation and evaluation of ML algorithms in clinical practice.

This first chapter includes a brief introduction to ML, discusses the relevance of ML to radiology, provides examples of current ML applications in radiology with a focus on cardiothoracic imaging, and finally outlines the structure of this thesis.

A brief introduction to Machine Learning

Machine Learning is a subcategory of Artificial Intelligence (AI). AI is a very broad term that describes the ability of computers to perform tasks commonly associated with human intelligence [1]. This includes simple *if-then* rules. The characteristic feature of ML approaches is that they learn from data without being explicitly programmed. Depending on the type of learning, the ML subcategories *supervised learning*, *unsupervised learning* and *reinforcement learning* are differentiated.

In *supervised learning*, the algorithm is trained with input-output pairs. An example for an input from the field of radiology is an image of a chest CT; the output could be the information whether the image shows a lung tumor or not. This is a binary classification task. Upon presentation of a certain number of correctly labelled input-output pairs during training, the model learns to map input data to output. Later, the algorithm is capable to predict the correct output for new, unseen input data (e.g., lung tumor present

or not). The vast majority of ML applications in radiology fall into this category. In *unsupervised learning*, no input-output pairs are provided. Instead, the algorithms' task is to identify patterns within the dataset. A clinical example is a model that identifies feature clusters within all patients that had been referred to a radiology department. In *reinforcement learning*, no input-output pairs are provided, either. Instead, the algorithm receives instant feedback on its decisions, rewarding desirable output. Over time, the model learns which behavior is adequate in a given situation. An example in radiology is ML for speech recognition. Those algorithms maximize adaption to the voice of the individual radiologist over time (feedback being the lack or presence of manual correction of a word proposed by the speech recognition system).

The term ML covers a variety of methods with individual strengths and weaknesses. A task-approach fit is of upmost importance. Highly relevant to radiology are *Support Vector Machines* (SVM), *Random Forest* (RF) and *Deep Learning* (DL). SVMs are used to group data points by identifying a hyperplane in a high-dimensional space that separates them. This computationally advantageous approach is very useful in classification tasks such as the one on radiology report classification provided in **Chapter 2**. Another approach to classification is RF. The forest consists of many decision trees, randomly built to cover different sets of features. The decision of the RF model is then reached by collecting the votes of all trees. This approach is used in **Chapter 2** as well. However, the vast majority of applications in cardiothoracic radiology falls into the category of DL, which includes state-of-the-art approaches to image recognition and segmentation tasks such as U-Net [2]. It is based on models with multiple interconnected layers consisting of "neurons", that is nodes that process and forward information. Basic components of DL networks are an input layer, an output layer, and a defined number of layers in between, so called *hidden layers*. To provide an example from radiology, pixel grayscale levels of a CT scan can be fed into the input layer and are then processed through the hidden layers. Finally, the output layer might provide a class (e.g.: image shows the left ventricle, or not). As the number of hidden layers has increased over the years, the term deep learning has evolved. The networks architecture is part of the so called *hyperparameters*, that is parameters that control the learning process, which are set before interaction with data (number of layers being one example). So called node *weights* are adapted during the exposure to training data so that prediction error is minimized. After training, these weights are fixed and the DL model is able to predict the correct outcome upon presentation of new, unseen data.

The advantage of this approach is its enormous capacity to incorporate relations between data. DL is applied in **all Chapters**.

ML task categories and standard performance measures

Machine learning has the potential to solve many clinical tasks. To radiology, the following task categories are of upmost importance: Classification, detection, and segmentation. Classification can be performed on a per-image level (does the chest X-ray show pneumothorax?), per-scan (does the CT scan show intracranial hemorrhage?) or per pixel/voxel (does this pixel belong to class "pulmonary embolus"?). An example for finding detection is an algorithm that marks coronary stenoses [3]. Results in this task category are often visualized with *bounding boxes*, that is rectangles surrounding a finding of interest. The result of a segmentation task is a label attached to each pixel/voxel signifying to which class it belongs to (e.g., "aorta – lumen", "aorta – wall", "mediastinal adipose tissue", and so on). Segmentations are the basis for many secondary analyses such as volumetry and computer-assisted reporting.

Each task category has specific metrics and visualizations of performance, and only the most relevant and established ones shall briefly be mentioned. Regarding binary classification and detection, those include numbers of true-positive (TP), true-negative (TN), false-positive (FP) and false-negative (FN) predictions. The central performance measures calculated with this information are accuracy [(TP+TN)/(TP+TN+FP+FN)], sensitivity [=TP/(TP+FN)], specificity [=TN/(TN+FP)] the positive predictive value [PPV; =TP/(TP+FP)] and the F1-score [=2*(sensitivity*PPV)/(sensitivity+PPV)]. The most frequently used visualization for binary classifiers is the receiver operating characteristic (ROC) curve, which plots true positive rates against false positive rates over the range of classifier thresholds. The area under the ROC (=AUC) is used as aggregate measure of performance across all classifier thresholds [4]. The free response ROC (FROC) plots sensitivity against the average number of FPs per case and is very useful in radiology, because it visualizes the trade-off between sensitivity and false alerts, which both can render a solution inappropriate for clinical use. For segmentation tasks, the Sørensen–Dice coefficient (Dice score) is standard, which is calculated as $2 * |X \cap Y| / (|X| + |Y|)$, where X represents the pixels/voxels included in the ground truth mask and Y the pixels/voxels included in the predicted segmentation mask. It ranges from 0 (no overlap at all) to 1 (complete overlap). Another standard measure is the Hausdorff distance, which measures

the distance of two outer measures in a metric space (e.g., in segmentation tasks the boundaries of the ground truth and the predicted mask) [5].

Relevance of ML in the field of radiology and current applications

There are three main reasons why ML and radiology are an excellent match: First, text and image recognition are among the core capabilities of modern ML algorithms. In fact, ML originates from these two fields that are so important to radiology. Second, data in radiology has become predominantly digital and therefore machine-readable, a prerequisite for ML. And third, the capacities of modern computer processing units have greatly increased over the last years, which is why fast processing of large amounts of data, such as DICOM image datasets, has become feasible. This enables clinical application. Additionally, ML comes at the right time: both the numbers of imaging exams and the demand for advanced analyses requiring complex post-processing are steadily increasing [6]. This is especially true in cardiothoracic imaging. The resulting increase in workload is not matched by an adequate increase in professional staff due to economic pressures. ML shows a way out of this dilemma by supporting radiologists in their work.

The field of image analysis provides some of the most apparent examples of application of ML in radiology: Based on finding detection, ML can timely warn radiologists of suspected critical findings such as pulmonary embolism [7], which is becoming more relevant as the total number of examinations increases. This information can also be used for worklist prioritization. Furthermore, ML segmentations can enhance the quality of reports, e.g., by providing volumetric information on organs and imaging findings [8]. Again, this is especially relevant to cardiothoracic imaging with time-consuming reading processes, e.g., for advanced cardiac MRI [9]. The information can also be used to prepopulate radiology reports. However, the utility of ML goes beyond pure image analysis and in fact includes the whole pipeline of radiology tasks. Algorithms are used for scheduling patients [10] as well as for image reconstruction in MRI [11]. They can help to optimize the display of series in medical image viewers [12] and provide support for clinical decision-making by integrating information from radiology, laboratory medicine, and other sources [13]. Furthermore, they can be used for automated image quality analysis [14]. Some cardiothoracic-specific applications include automated curved multiplanar reformations of the coronaries in CT angiograms [15], automated segmentation of the cardiac chambers with DL [16,17] and prediction of coronary stenoses

requiring invasive coronary angiography from CT [18]. Other examples are provided in the following chapters of this thesis.

Structure of the thesis

As mentioned previously, the structure of this thesis follows the chronology of ML application during a research project in radiology, from data curation to clinical application.

Chapter 2 demonstrates how ML can be applied for data curation. Natural language processing is used to determine whether radiology reports of CT pulmonary angiograms describe the presence of pulmonary emboli or not. After manual annotation of a small subset of reports, an infinite number of reports can be automatically classified. The resulting structured datasets build the foundation for scientific projects, case collections, and quality control measures. This study demonstrates the potential of NLP for fast data curation and at the same time stimulates a discussion about the advantages of structured reporting.

Chapter 3 evaluates the performance of a deep convolutional neural network in detecting pulmonary embolism on the 1-mm series of CT pulmonary angiograms. For this purpose, 1465 CTPAs are processed and results reviewed. Besides general performance measures, an in-depth analysis of reasons of false-positive findings is conducted to generate ideas for algorithm improvement. The resulting algorithm is currently in clinical use to instantly notify radiologists about CTPAs with pulmonary embolism.

Chapter 4 investigates the performance of a convolutional neural network for detection of acute and chronic rib fractures on a large dataset of whole-body trauma CTs acquired at a level-1 trauma center. The analysis distinguishes per-examination and per-finding performance and analyzes the influence of fracture features on detection probability. This algorithm can help radiologists to avoid missing rib fractures, which, despite not being the main concern in an emergency situation with life-threatening conditions, are known to be associated with an elevated risks for asynchronous complications such as pneumothorax and pneumonia.

Chapter 5 investigates the performance of an algorithm pipeline initially trained to detect and segment lung nodules on chest CT for detection and segmentation of pulmonary

tumors of various size on FDG-PET/CT. For this purpose, the 320 tumors in the test dataset were manually labelled in 3D and classified into T-subcategories according to the 8th edition of the TNM lung cancer classification. This study emphasizes the importance of investigating the performance of ML algorithms for different stages of disease.

Chapter 6 uses multiple deep convolutional neural networks to extract cardiovascular and pulmonary imaging parameters from 120 initial chest CTs of patients with RT-PCR confirmed SARS-CoV-2 infection. The predictive potential of those imaging parameters, six standard laboratory parameters and demographic information regarding a patient's treatment journey (outpatient treatment vs. hospital admission vs. ICU admission) was assessed. The algorithm could identify patients at high risk of needing intensified treatment within the following weeks, already at the time of initial hospital presentation.

Chapter 7 is based on the experience gained during the abovementioned scientific projects and provides a framework for the design and evaluation of ML studies in cardiothoracic radiology. This includes prerequisites of ML projects with regard to hardware, software, and expertise, but also a checklist of items that should be reported in research articles in the field of cardiothoracic imaging.

Finally, **Chapter 8** summarizes and discusses the main findings. Furthermore, future directions of ML in radiology with a focus on cardiothoracic imaging are outlined. The valuable contributions of other persons to this thesis are acknowledged in the Appendix.

References

1. Artificial Intelligence | Definition, Examples, Types, Applications, Companies, & Facts | Britannica. https://www.britannica.com/technology/artificial-intelligence. Accessed 4 Feb 2022
2. Ronneberger O, Fischer P, Brox T (2015) U-Net: Convolutional Networks for Biomedical Image Segmentation. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 9351:234–241
3. Zreik M, van Hamersvelt RW, Wolterink JM, Leiner T, Viergever MA, Išgum I (2019) A Recurrent CNN for Automatic Detection and Classification of Coronary Artery Plaque and Stenosis in Coronary CT Angiography. IEEE transactions on medical imaging 38:1588–1598
4. Chen C-P, Lin Y-M (2010) Bland-Altman Plots and Receiver Operating Characteristic Curves Are Preferred. Radiology 257:896–896
5. Huttenlocher DP, Klanderman GA, Rucklidge WJ (1993) Comparing Images Using the Hausdorff Distance. IEEE Transactions on Pattern Analysis and Machine Intelligence 15:850–863
6. Bruls RJM, Kwee RM (2020) Workload for radiologists during on-call hours: dramatic increase in the past 15 years. Insights into imaging. https://doi.org/10.1186/S13244-020-00925-Z
7. Weikert T, Winkel DJ, Bremerich J, Stieltjes B, Parmar V, Sauter AW, Sommer G (2020) Automated detection of pulmonary embolism in CT pulmonary angiograms using an AI-powered algorithm. European Radiology 30:6545–6553
8. Winkel DJ, Weikert TJ, Breit H-C, Chabin G, Gibson E, Heye TJ, Comaniciu D, Boll DT (2020) Validation of a fully automated liver segmentation algorithm using multi-scale deep reinforcement learning and comparison versus manual segmentation. European Journal of Radiology. https://doi.org/10.1016/j.ejrad.2020.108918
9. Tao Q, Lelieveldt BPF, van der Geest RJ, Q. T, B.P.F. L, R.J. van der G, Tao Q, Lelieveldt BPF, van der Geest RJ (2019) Deep Learning for Quantitative Cardiac MRI. AJR American journal of roentgenology 1–7
10. Kapoor N, Lacson R, Khorasani R (2020) Workflow Applications of Artificial Intelligence in Radiology and an Overview of Available Tools. Journal of the American College of Radiology : JACR 17:1363–1370
11. Qin C, Schlemper J, Caballero J, et al (2019) Convolutional recurrent neural networks for dynamic MR image reconstruction. IEEE Transactions on Medical Imaging 38:280–290
12. Filice RW, Stein A, Pan I, Shih G (2022) Federated Deep Learning to More Reliably Detect Body Part for Hanging Protocols, Relevant Priors, and Workflow Optimization. Journal of digital imaging. https://doi.org/10.1007/S10278-021-00547-X
13. Weikert T, Rapaka S, Grbic S, et al (2021) Prediction of Patient Management in COVID-19 Using Deep Learning-Based Fully Automated Extraction of Cardiothoracic CT Metrics and Laboratory Findings. Korean journal of radiology 22:994–1004
14. Fung A, Moulson N, Balthazaar S, et al (2019) Artificial Intelligence Application for Assessing Point-of-Care Ultrasound Image Quality. Journal of the American Society of Echocardiography 32:B122
15. Tatsugami F, Higaki T, Nakamura Y, et al (2019) Deep learning-based image restoration algorithm for coronary CT angiography. European radiology 29:5322–5329
16. Romaguera LV, Romero FP, Fernandes Costa Filho CF, Fernandes Costa MG (2018) Myocardial segmentation in cardiac magnetic resonance images using fully convolutional neural networks. Biomedical Signal Processing and Control 44:48–57
17. Bai W, Sinclair M, Tarroni G, et al (2018) Automated cardiovascular magnetic resonance image analysis with fully convolutional networks. Journal of cardiovascular magnetic resonance : official journal of the Society for Cardiovascular Magnetic Resonance 20:65
18. Zreik M, van Hamersvelt RW, Khalili N, et al (2019) Deep learning analysis of coronary arteries in cardiac CT angiography for detection of patients requiring invasive coronary angiography. IEEE transactions on medical imaging. https://doi.org/10.1109/TMI.2019.2953054

CHAPTER 2

Automated Generation of Curated Datasets in Radiology: Application of Natural Language Processing to Unstructured Reports Exemplified on CT for Pulmonary Embolism

BASED ON: Weikert T, Nesic I, Cyriac J, Bremerich J, Sauter AW, Sommer G, Stieltjes B. Towards automated generation of curated datasets in radiology: Application of natural language processing to unstructured reports exemplified on CT for pulmonary embolism. Eur J Radiol. 2020 Apr;125:108862. doi: 10.1016/j.ejrad.2020.108862.

Abstract

Purpose: To design and evaluate a self-trainable natural language processing (NLP)-based procedure to classify unstructured radiology reports. The method enabling the generation of curated datasets is exemplified on CT pulmonary angiogram (CTPA) reports.

Materials and Methods: We extracted the impressions of CTPA reports created at our institution from 2016 to 2018 (n=4397; language: German). The status (pulmonary embolism: yes/no) was manually labelled for all exams. Data from 2016/2017 (n = 2801) served as a ground truth to train three NLP architectures that only require a subset of reference datasets for training to be operative. The three architectures were as follows: a convolutional neural network (CNN), a support vector machine (SVM) and a random forest (RF) classifier. Impressions of 2018 (n = 1377) were kept aside and used for general performance measurements. Furthermore, we investigated the dependence of classification performance on the amount of training data with multiple simulations.

Results: The classification performance of all three models was excellent (accuracies: 97—99 %; F1 scores 0.88-0.97; AUCs: 0.993-0.997). Highest accuracy was reached by the CNN with 99.1 % (95 % CI: 98.5-99.6 %). Training with 470 labelled impressions was sufficient to reach an accuracy of > 93 % with all three NLP architectures.

Conclusion: Our NLP-based approaches allow for an automated and highly accurate retrospective classification of CTPA reports with manageable effort solely using unstructured impression sections. We demonstrated that this approach is useful for the classification of radiology reports not written in English. Moreover, excellent classification performance is achieved at relatively small training set sizes.

1. Introduction

The retrospective assessment of whether a certain finding is present or not is a necessary first step for the secondary utilization of data accumulated in radiology imaging archives. This is true for scientific projects as well as efforts targeted at the development of applications that improve clinical workflows. Presuming that the information needed is contained in the radiology reports, its extraction is hampered by the fact that these mostly are continuous, variable texts, despite efforts towards standardization in recent years [1]. This lack of structured content becomes a problem in an increasingly data-driven world and with the emergence of methods like neural networks for image analysis that need substantial amounts of data to be operational. This is especially pressuring in hospitals, where sparse time resources prevent manual classification at large scale [2].

In search of ways to transfer unstructured information to meaningful categories, natural language processing (NLP) is a promising approach with beginnings dating back to the 1950s [3,4]. It has demonstrated its maturity for text mining in radiology in many fields of use, including the detection of critical findings [5–8], decision support for clinicians [9,10], quality assessment of reports [11,12] and the generation of curated datasets [13–19]. While in most of these cases, lexical rule systems designed by experts built the backbone of NLP, recently methods that allow for an automated model generation based on training data were incorporated and demonstrated promising results for the classification of radiology reports written in English [17,20–22]. One of the drawbacks of lexical rule systems is that they are limited to the language they were created in. That is why the use of automated model generation represents a step forward particularly for application in Europe with multiple official languages. Once such a model is set up, it needs only to be trained with reports in a given language and can therefore be used by radiologists in many countries.

To further investigate the potential of those automated NLP solutions in radiology and demonstrate their applicability to languages other than English, we developed and tested three NLP architectures that have shown good results in classification elsewhere, namely based on random forest (RF) [21,23], a support vector machine (SVM) [21,24,25] and a convolutional neural network (CNN) [26]. The evaluation was performed on the specific task of generating curated datasets from unstructured radiology reports concerning the detection of pulmonary embolism (PE) in CT pulmonary angiograms (CTPAs).

Furthermore, we assessed the dependence of classification performance on the number of impressions used for training.

2. Materials and Methods

This study was conducted under the provisions of the local ethics committee. Informed consent was waived due to the retrospective nature of this project and full anonymization of the data.

2.1. Data retrieval, pre-processing and annotation

To build a study dataset, we retrospectively identified all reports of CTPAs conducted at our institution between 2016 and 2018 using an in-house developed RIS search engine (n = 4397). The only criteria were the procedure code and the time period (*01.01.2016 – 31.12.2018*). Impression sections were automatically extracted with an in-house report segmentation tool based on a keyword search. The rationale for using impressions only was that the clinical question of interest is answered in this section, and therefore the rest of the report is not needed. We deliberately refrained from performing document normalization or including pre-filtering of the reports based on lexical rules as we wanted to keep manual input as small as possible. In a next step, all impressions were manually reviewed by a radiology resident with 3 years of experience (TW). Exams with a clinical question deviating from that of the presence of PE (n = 128) and exams with an uncertain conclusion (n = 91) were excluded. All remaining impressions were classified as either positive or negative for PE. Finally, we uploaded the texts to our in-house developed web-based NLP platform with the three NLP architectures (CNN, SVM, RF). The study workflow is shown in Figure 1.

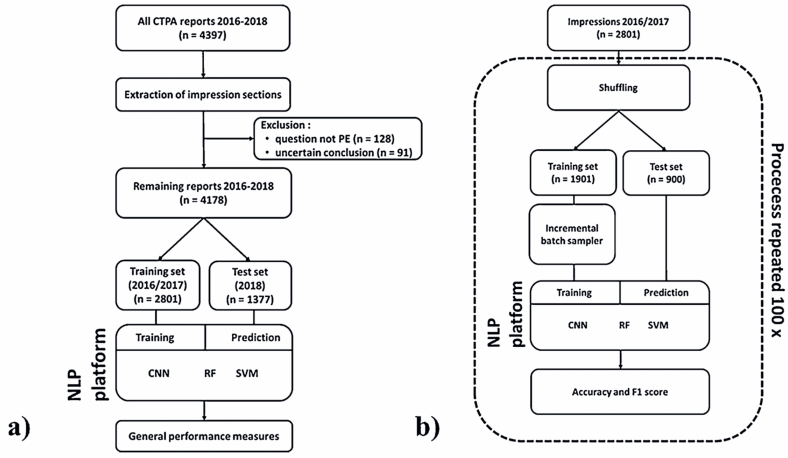


Fig. 1. (a) Workflow of data preprocessing and general performance measurements. For training, all impressions of 2016/2017 were used. Testing was done on a completely separate test set comprising all impression of 2018. (b) workflow of performance measurements depending on training batch size. After shuffling the order of impressions, a test set (n = 900) was separated. From the training set (n = 1901), a subset with an increasing number of randomly selected impressions was chosen by a batch sampler and accuracy and F1 score were calculated for all three approaches. This procedure was repeated 100 times.

2.2. Modelling

For model generation, we used Scikit-learn Version 0.19.1 [27] (SVM, RF) and Tensor Flow version 1.9.0 [28] (CNN). All models do not require previous definition of lexical rules and were downloaded from these platforms; only minor modifications were made. Parameters that differed from default are indicated below. We defined 0.5 as fixed classifier threshold for all three models. We used two data representations: term frequency-inverse document frequency (tf-idf) [29,30] and the Word2vec model [31]. For hyperparameter tuning, we assessed the best accuracy on a 3-fold cross-validation.

CNNs are neural networks that rely on the mathematical operation of convolution. They consist of multiple connected layers. Beginning with an input layer, the representation of report texts is propagated through multiple so called “hidden layers” that perform computations by which they capture more and more abstract features of the input information. Finally, the output of the network we used is a classifier indicating whether a report describes pulmonary embolism or not. CNNs have shown good results when applied to imaging problems [32]. With the Word2vec method [31], it is possible to apply CNNs to textual problems in a similar manner. Our neural network architecture is conceptually similar to Kim et al. [26], but with average global pooling layer [33] instead of max pooling layer. As reported by Duque et al., networks that utilize global average pooling are more efficient in terms of speed and memory usage while yielding comparably

good results with those that rely on fully connected layers after the other types of pooling [34]. No hyperparameter tuning was performed.

Support vector machines are supervised machine learning algorithms primarily used for binary classification [35]. After transferring data input to feature vectors in a high-dimensional vector space, a hyperplane that best separates the different classes is calculated. We use tf-idf data representation with this algorithm. The parameters were selected by performing the grid search on the training dataset. Parameter C was set to 2. For gamma we used an inverse number of features (1/n). The kernel type used is linear. We also used L2 regularization to penalize the weights. The loss function is the squared Hinge loss.

The random forest approach was first applied in the form of algorithms in 2001 [36]. The idea is to randomly generate a large number decision trees assembled by a subset of decision relevant variables. We used tf-idf data representation with this algorithm. Parameters were selected by performing a grid search. Near optimal parameters on the training set: minimum samples split = 3, number of estimators = 20. Time needed to train the three models was recorded.

2.3. Performance measurements

The clinically approved written report was defined as reference standard. The index test was the prediction made by the three NLP architectures.

First, for general performance measurements, the impressions were grouped into a training dataset (all exams of 2016 and 2017, n = 2801) and a fully separate testing dataset (all exams of 2018, n = 1377). As performance measures, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), accuracy and the F1 score were calculated. The F1 score is a harmonic mean of PPV and sensitivity, with values ranging between 0 and 1.

$$\mathrm{F1} = \left(\frac{PPV^{-1} + Sensitivity^{-1}}{2} \right)^{-1}$$

Furthermore, receiver operating curves (ROCs) were computed. The corresponding areas under the ROCs (AUCs) were calculated to illustrate the discriminative performance of the models. Second, to investigate the dependence of performance on the amount of training data, we repeatedly calculated accuracies and F1 scores for the three architectures

using different training subset sizes. To this end, we used all impressions of 2016/2017 (n = 2801) and separated a test dataset comprising 900 impressions (~1/3). The remaining impressions (n = 1901; ~2/3) served as training data. This ratio of training and testing dataset is common and well-suited for the size of our dataset [37]. We started training with 50 impressions and then incrementally added 60 impressions at each training cycle in a random fashion until all impressions of the test dataset were included. Concretely, the first training was performed with 50 impressions and the performance on the test dataset was determined. Then training was performed with 110 impressions, the performance was determined, and so on. To account for the effect of different impressions included into the training subsets, this procedure was repeated 100 times, each time starting with a randomly selected subset. The workflow for the performance depending on training batch size is displayed in Fig. 1b.

2.4. Statistical analysis

Statistical analysis was performed with IBM SPSS Statistics for Windows, Version 24.0 (IBM Corporation). To test for differences between the training and the test cohort, a Mann-Whitney-U-Test for age and Chi-Square Tests for gender and PE status were calculated. Furthermore, we calculated the percentage of PE-negative reports that contained the phrasing "no pulmonary embolism" to determine how many impressions could have been identified with a simple word search. Time needed to manually label the impressions was recorded. Number of words contained in correctly classified and misclassified impressions was determined for the approach with the highest accuracy and a Mann-Whitney-U-Test was performed to test for differences. Measures of diagnostic accuracy and their 95 % confidence intervals were calculated as mentioned above.

3. Results

The training dataset consisted of 2801 CTPA impressions created in 2016 and 2017. Time needed to train the models with the complete training dataset was < 1 s for RF and SVM and < 5 min for the CNN. There was no missing data; all impressions were processed by all three NLP architectures. In 2018, 68.5 % of the PE-negative impressions contained the phrasing "no pulmonary embolism", while 336 did not. The ratio of impressions of PE negative exams containing "no pulmonary embolism" was even lower in 2016 (44.8 %) and 2017 (66.8 %).

3.1. General performance

The testing dataset covered 1377 reports from 2018. The datasets did not differ statistically significantly in respect to age and the ratio of positive and negative exams (see Table 1). However, the percentage of men in the testing dataset (50.5 %) was significantly higher as compared to the training dataset (46.0 %). Performance results of the models with 95 % confidence intervals are listed in Table 2. The CNN showed the best results regarding sensitivity, NPV, accuracy and F1 score. The random forest approach achieved the highest specificity and PPV. Receiver operating curves for all three approaches are displayed in Fig. 2. The AUC values were close to 1 for all three approaches (CNN: 0.997; SVM: 0.993; RF: 0.996). Fig. 3 displays example CTPA images with related report impression sections and the class attributed to the impressions by the three NLP approaches. Impressions that were misclassified by the CNN had a higher mean word count (54.6; SD: 34.4) than those with correct predictions (38.8; SD: 21.5). This difference was statistically significant (Mann-Whitney-U-Test; p = 0.01).

Table 1
Patient characteristics of the training set (2016/2017) and the general test set (2018).

	Training set 2016/2017 (n = 2801)	Test set 2018 (n = 1377)	Statistical test for group difference	Result
Age (mean / SD)	65.5 (17.1)	65.4 (17.6)	Mann –Whitney U	p = 0.792
Gender				
Male (n, %)	1288 (46.0 %)	695 (50.5 %)	Chi-Square	p = 0.006
Female (n, %)	1513 (54.0 %)	682 (49.5 %)		
PE status				
Positive (n, %)	431 (15.4 %)	214 (15.5 %)	Chi-Square	p = 0.897
Negative (n, %)	2370 (84.6 %)	1163 (84.5 %)		

Table 2
Performance measurements of the three NLP architectures. Sensitivity, specificity, PPV, NPV and accuracy in percent, F1 score is dimensionless. 95 % confidence intervals in brackets. The best result observed for each measure is highlighted with an underscore.

Approach	Sensitivity	Specificity	PPV	NPV	Accuracy	F1 score
CNN	97.7 %	99.4%	96.8%	99.6%	99.1%	0.972
	(94.6–99.2 %)	(98.8–99.8 %)	(93.5–98.4 %)	(99.0–99.8 %)	(98.5–99.6 %)	(0.963-0.981)
SVM	93.0 %	97.7%	88.1%	98.7%	97.0%	0.905
	(88.7–96.0)	(96.6–98.5)	(83.5–91.5)	(97.9–99.2)	(95.9–97.8)	(0.890-0.920)
RF	80.8 %	99.6%	97.2%	96.6%	96.7%	0.883
	(74.9–85.9)	(99.0–99.9)	(93.5–98.8)	(95.6–97.4)	(95.6–97.5)	(0.866-0.900)

PPV: positive predictive value; NPV: negative predictive value; RF: random forest; SVM: support vector machine; CCN: convolutional neural network.

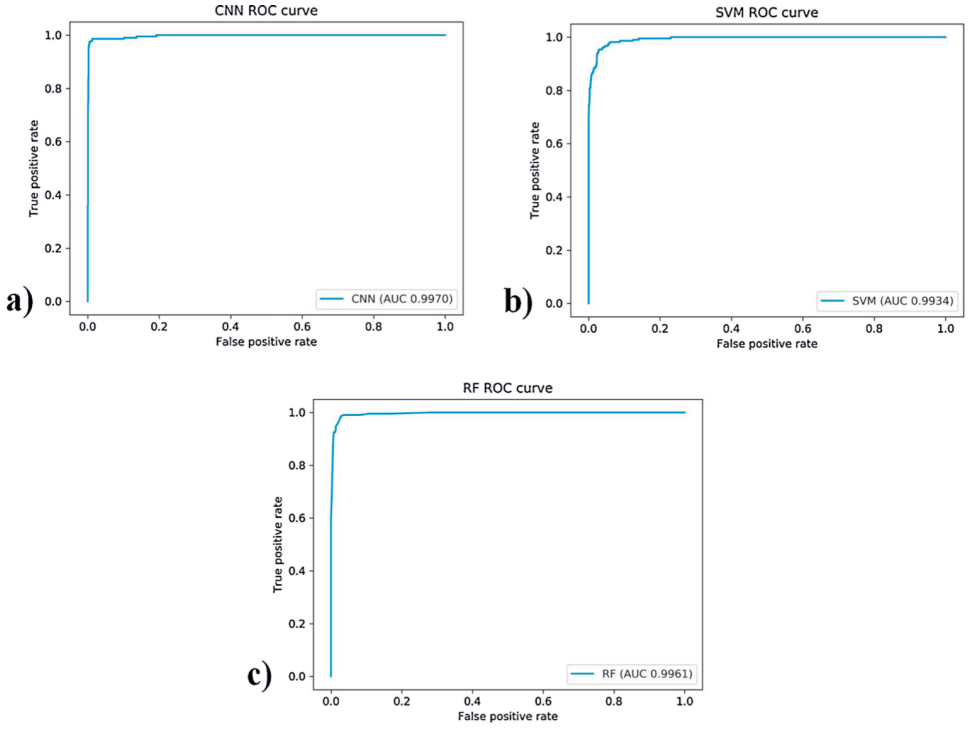


Fig. 2. Receiver operating curves of the different NLP approaches for differentiating reports with and without PE. Corresponding areas under the curve are (a) 0.997 the CNN, (b) 0.993 for SVM and (c) 0.996 for RF.

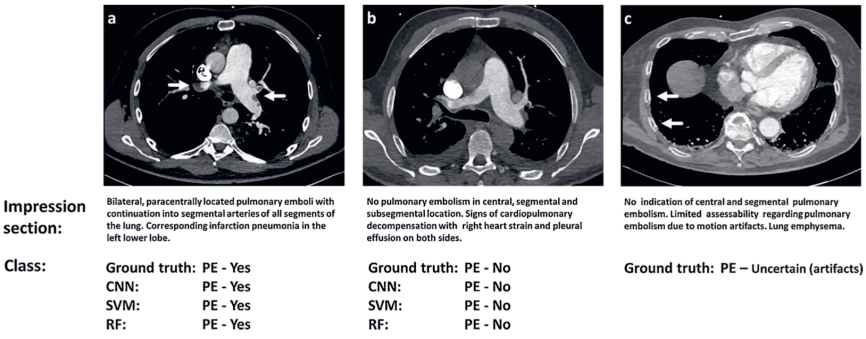


Fig. 3. Three examples of CTPAs with corresponding impression sections as used for testing (English translation of German reports) and the class as determined by the convolutional neural network approach (CNN), the support vector machine (SVM) and the random forest approach (RF). (a) CTPA with pulmonary embolism (emboli marked with white arrows), (b) CTPA negative for pulmonary embolism and (c) CTPA with uncertain conclusion due to breathing artifacts (as indicated by blurred ribs marked by white arrows).

3.2. Training data size effect on performance

Fig. 4 displays the performance of the models depending on the number of impressions used for training. After an initial sharp increase in accuracy and F1 score by adding new

impressions to the training dataset, the increase in performance reached a plateau at relatively small subset sizes. An accuracy of over 93 % was reached at training subset sizes of 170 (SVM), 470 (RF) and 470 (CNN), respectively. By labelling this smaller number of impressions instead of the whole training dataset, 223 min (SVM), 198 min (RF) and 198 min (CNN) would have been saved. This estimation is based on the average time needed to manually label one impression (5.1 s). Labelling all 2810 impressions of 2016/2017 took 238 min.

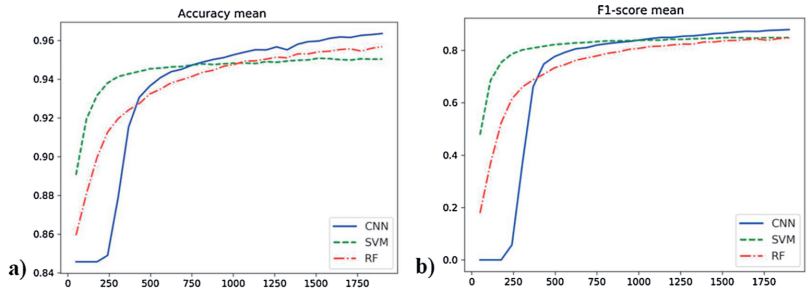


Fig. 4. (a) Mean accuracy and (b) mean F1 score depending on the amount of training data, based on 100 simulations for each model. The amount of impressions used for training is plotted against the X-axis.

4. Discussion

We found that all three models achieved very good classification performance with accuracies ranging from 97%–99% and area under the curve values above 0.99. These algorithms therefore very well predict the mentioning of PE in a given report solely based on the impression section. A second relevant finding is that excellent classification performance is reached at relatively small batch sizes. Labelling more data for training beyond that point did only slightly improve the performance. The optimization of the number of impressions used for training saves time. In our case, e.g., an accuracy of > 93 % is achieved by labelling a subset of 170 impressions (SVM), which would have saved 223 min when compared to a full labelling of the 2016/2017 dataset. The specific classification performance warranted depends on the research question.

Interpreting these results, it is important to keep in mind that 68.5 % of PE-negative impressions in 2018 contained the standard negative phrasing “no pulmonary embolism”, which means they would have been identified by a simple word-based search. This is due to efforts towards structured reporting in recent years. However, a relevant number of

exams would have been missed by this simplistic approach and our NLP architectures are superior classifiers by far. Moreover, the ratio of PE negative impressions that contain "no pulmonary embolism" had been lower in previous years (e.g., 44.8 % in 2016), underlining the usefulness of NLP for reviewing archives. At the same time, this emphasizes the utility of highly structured reporting that, if consistently applied, can render secondary classification of reports redundant but is currently far from being standard in radiological practice.

Previous studies on classification of CTPA reports regarding the presence of PE likewise reported excellent results comparable or only slightly inferior to ours: Yu and colleagues applied a combination of customized lexical rules and machine learning based on the Narrative Information Linear Extraction system (NILE) to achieve a F1 score of 0.96 [19]. Chapman et al. introduced the tool peFinder that is looking for lexical cues in a text, achieving an accuracy of 92 % and a F1 score of 0.90 [16]. However, the approaches presented by Yu and Chapman both need significant manual contributions in the form of definition of terms and lexical rules. This implies five important drawbacks: high expenditure of time, dependence on the expertise of the people involved, inflexibility with regard to different language styles, limitation to the language they were created in and constriction to the specific question they were designed for. This restricts practical applicability, especially in research where many questions arise at fast pace and thus adjustability of algorithms to new challenges is key.

Gerstmair et al. presented a NLP-tool called RadMiner [38]. Based on a test set of 108 reports, they reported a sensitivity of 0.93 and a PPV of 0.95. As the approach is based on an extensive processing pipeline including among others morpho-sematic analysis, abbreviation detection and a dictionary of medical terms, the setup of the platform is complex and time-consuming. Regarding fully automated algorithms for the classification of English reports regarding presence of PE, Chen et al. presented a convolutional neural network achieving an accuracy of 99 % and a F1 score of 0.94 [17]. They used a CNN with a very similar architecture to ours. However, they did not test other architectures and did not evaluate the dependence of performance on the amount of training data. The usefulness of automated NLP architectures for labelling radiology reports was also demonstrated for other tasks: Brown et al. reported an accuracy of up to 92 % and a sensitivity of 83 % for the prediction of downstream utilization of radiology resources in patients with hepatocellular carcinoma using a SVM and – in accordance with our results

– a moderately worse performance of a RF approach [22]. Li and Elliot reported an accuracy of 85 % and a specificity of 95 % for the identification of a group of patients with ureteric stones [18]. The excellent classification performance that we found is in line with these studies.

We trained and tested on impressions in German language. This demonstrates that even German language with peculiarities like the frequent composition of nouns is fully accessible to NLP approaches. Despite the fact that there are approximately 400 million English native speakers in the world [39] and English is the undisputed *lingua franca* in science, it is important to develop NLP solutions for other languages, too, to facilitate the secondary usage of medical texts worldwide. This is why another main insight of our study is that NLP works for the generation of curated datasets in radiology for German texts at accuracy levels comparable to those reported for English reports. Furthermore, we found that wrongly classified impressions had a higher mean word count than those with correct predictions. A higher word count might indicate a higher number of secondary diagnoses that lower the signal-to-noise ratio of the text.

At this point, we want to give a short practical, four-step-guidance on the application of NLP in radiology: In a first step, one should ask if NLP is the right method for a given radiology report classification task. If one wishes to classify a low number of, e.g., 100 reports, the effort for putting up the NLP model clearly exceeds that of manually classifying those reports. However, in large sample size studies consisting of several thousand data samples, as they are more and more common, application of NLP is time-saving. Second, we suggest to start with an easily implementable approach like support vector machines which demand only moderate computing power and showed good performance in our study. Third, a researcher should manually label a limited number of e.g., 300 reports (250 for training & validation of the algorithm, 50 independent cases for performance testing). The question of how many labelled reports are needed depends on many factors as the intended accuracy and report heterogeneity. Fourth, the performance of the model should be tested using an independent test set. If it is sufficient for the needs of the project, it can be applied to the rest of the data. If not, more training data or the use of other NLP models are required.

Our study has several limitations. First, for training and general testing, we excluded reports that did not come to a distinct conclusion due to mediocre image quality. However,

these reports constitute only 2.1 % of all cases. Interestingly, this number is lower than reported elsewhere (e.g., 4.2 % reported by Bates and colleagues [40]). While this might be partly due to optimized CTPA protocols used at our institution, in some cases, the information on suboptimal image quality might just not have been added to the impression section, which is another limitation of our study. Second, we developed and tested solely on data of one medical center. The external validity of this approach has to be proven. However, this approach can easily be adapted by other centers with manageable effort to create a customized, locally applicable NLP solution. Third, difference between the gender ratio of the test dataset and that of the training dataset was statistically significant. However, we used only the impression sections, where the gender of a given patient was never mentioned. Therefore, we do not expect any gender bias. Fourth, the question we chose to illustrate the performance of NLP was whether there is pulmonary embolism or not. The performance of our NLP systems for other questions might vary. Furthermore, the reports on CTPAs with the question of PE at our institution are partially standardized with a suggested standard sentence for exams negative for PE. The performance on less standardized impressions might vary. Fifth, reports were labelled only by one radiologist. Given the low grade of complexity of this task ("is PE described in the given impression section or not") we deliberately refrained from a second reading since we expect no variation between readers. Sixth, we used only impression sections to train, validate and test our algorithms. For other questions that are not answered in impressions on a regular basis, the whole report text has to be included. Seventh, NLP is a dynamic research field with a plethora of methodological approaches. We limited our analysis to three NLP approaches: a CNN, a SVM, and RF. We did so because these are standard approaches readily available to interested researchers. However, we recommend to closely follow new developments and to consider other approaches, too.

In conclusion, we developed three self-trainable NLP algorithms that classify CTPAs based on their impression sections with high accuracy. This enables swift retrospective analysis of large amounts of data in radiology archives. Furthermore, we demonstrated that small training datasets can be sufficient to reach excellent classification performance.

References

1. D. Ganeshan, P.-A.T. Duong, L. Probyn, L. Lenchik, T.A. McArthur, M. Retrouvey, E.H. Ghobadi, S.L. Desouches, D. Pastel, I.R. Francis, Structured Reporting in Radiology, Acad. Radiol. 25 (2018) 66–73. https://doi.org/10.1016/j.acra.2017.08.005.
2. NHS, Hospital Admitted Patient Care and Adult Critical Care Activity, 2018.
3. P.M. Nadkarni, L. Ohno-Machado, W.W. Chapman, Natural language processing: an introduction., J. Am. Med. Inform. Assoc. 18 (2011) 544–51. https://doi.org/10.1136/amiajnl-2011-000464.
4. C. Liew, The future of radiology augmented with Artificial Intelligence: A strategy for success, (2018). https://doi.org/10.1016/j.ejrad.2018.03.019.
5. B. Rink, K. Roberts, S. Harabagiu, R.H. Scheuermann, S. Toomay, T. Browning, T. Bosler, R. Peshock, Extracting actionable findings of appendicitis from radiology reports using natural language processing., AMIA Jt. Summits Transl. Sci. Proceedings. AMIA Jt. Summits Transl. Sci. 2013 (2013) 221.
6. P. Lakhani, W. Kim, C.P. Langlotz, Automated Detection of Critical Results in Radiology Reports, J. Digit. Imaging. 25 (2012) 30–36. https://doi.org/10.1007/s10278-011-9426-6.
7. A.B. Chapman, D.L. Mowery, D.S. Swords, W.W. Chapman, B.T. Bucher, Detecting Evidence of Intra-abdominal Surgical Site Infections from Radiology Reports Using Natural Language Processing., AMIA Annu. Symp. Proceedings. AMIA Symp (2017) 515–524.
8. J. Zech, M. Pain, J. Titano, M. Badgeley, J. Schefflein, A. Su, A. Costa, J. Bederson, J. Lehar, E.K. Oermann, Natural Language– based Machine Learning Models for the Annotation of Clinical Radiology Reports, Radiology. 287 (2018) 570–580. https://doi.org/10.1148/radiol.2018171093.
9. M. Yetisgen-Yildiz, M.L. Gunn, F. Xia, T.H. Payne, A text processing pipeline to extract recommendations from radiology reports, J. Biomed. Inform. 46 (2013) 354–362. https://doi.org/10.1016/j.jbi.2012.12.005.
10. S. Dutta, W.J. Long, D.F.M. Brown, A.T. Reisner, Automated detection using natural language processing of radiologists recommendations for additional imaging of incidental findings., Ann. Emerg. Med. 62 (2013) 162–9. https://doi.org/10.1016/j.annemergmed.2013.02.001.
11. R. Lacson, L.M. Prevedello, K.P. Andriole, R. Gill, J. Lenoci-Edwards, C. Roy, T.K. Gandhi, R. Khorasani, Fleischner Society, Factors Associated With Radiologists' Adherence to Fleischner Society Guidelines for Management of Pulmonary Nodules, J. Am. Coll. Radiol. 9 (2012) 468–473. https://doi.org/10.1016/j.jacr.2012.03.009.
12. R. Duszak, M. Nossal, L. Schofield, D. Picus, Physician Documentation Deficiencies in Abdominal Ultrasound Reports: Frequency, Characteristics, and Financial Impact, J. Am. Coll. Radiol. 9 (2012) 403–408. https://doi.org/10.1016/j.jacr.2012.01.006.
13. G. Hripcsak, J.H.M. Austin, P.O. Alderson, C. Friedman, Use of Natural Language Processing to Translate Clinical Information from a Database of 889,921 Chest Radiographic Reports, Radiology. 224 (2002) 157–163. https://doi.org/10.1148/radiol.2241011118.
14. M.J. Schuemie, E. Sen, G.W. 't Jong, E.M. Soest, M.C. Sturkenboom, J.A. Kors, Automating classification of free-text electronic health records for epidemiological studies, Pharmacoepidemiol. Drug Saf. 21 (2012) 651–658. https://doi.org/10.1002/pds.3205.
15. Y. Zhou, P.K. Amundson, F. Yu, M.M. Kessler, T.L.S. Benzinger, F.J. Wippold, Automated classification of radiology reports to facilitate retrospective study in radiology., J. Digit. Imaging. 27 (2014) 730–6. https://doi.org/10.1007/s10278-014-9708-x.
16. B.E. Chapman, S. Lee, H.P. Kang, W.W. Chapman, Document-level classification of CT pulmonary angiography reports based on an extension of the ConText algorithm., J. Biomed. Inform. 44 (2011) 728–37. https://doi.org/10.1016/j.jbi.2011.03.011.

17. M.C. Chen, R.L. Ball, L. Yang, N. Moradzadeh, B.E. Chapman, D.B. Larson, C.P. Langlotz, T.J. Amrhein, M.P. Lungren, Deep Learning to Classify Radiology Free-Text Reports, Radiology. 286 (2018) 845–852. https://doi.org/10.1148/radiol.2017171115.
18. A.Y. Li, N. Elliot, Natural language processing to identify ureteric stones in radiology reports, J. Med. Imaging Radiat. Oncol. (2019) 1754–9485.12861. https://doi.org/10.1111/1754-9485.12861.
19. S. Yu, K.K. Kumamaru, E. George, R.M. Dunne, A. Bedayat, M. Neykov, A.R. Hunsaker, K.E. Dill, T. Cai, F.J. Rybicki, Classification of CT pulmonary angiography reports by presence, chronicity, and location of pulmonary embolism with natural language processing, J. Biomed. Inform. 52 (2014) 386–393. https://doi.org/10.1016/j.jbi.2014.08.001.
20. J. Swartz, C. Koziatek, J. Theobald, S. Smith, E. Iturrate, Creation of a simple natural language processing tool to support an imaging utilization quality dashboard., Int. J. Med. Inform. 101 (2017) 93–99. https://doi.org/10.1016/j.ijmedinf.2017.02.011.
21. P.-H. Chen, H. Zafar, M. Galperin-Aizenberg, T. Cook, Integrating Natural Language Processing and Machine Learning Algorithms to Categorize Oncologic Response in Radiology Reports, J. Digit. Imaging. 31 (2018) 178–184. https://doi.org/10.1007/s10278-017-0027-x.
22. A.D. Brown, J.R. Kachura, Natural Language Processing of Radiology Reports in Patients With Hepatocellular Carcinoma to Predict Radiology Resource Utilization, J. Am. Coll. Radiol. (2019). https://doi.org/10.1016/J.JACR.2018.12.004.
23. Y. Li, Y. Dai, L. Deng, N. Yu, Y. Guo, Computer-aided detection for the automated evaluation of pulmonary embolism, Technol. Heal. Care. 25 (2017) 135–142. https://doi.org/10.3233/THC-171315.
24. S.M. Castro, E. Tseytlin, O. Medvedeva, K. Mitchell, S. Visweswaran, T. Bekhuis, R.S. Jacobson, Automated annotation and classification of BI-RADS assessment from radiology reports, J. Biomed. Inform. 69 (2017) 177–187.
25. C.M. Rochefort, A.D. Verma, T. Eguale, T.C. Lee, D.L. Buckeridge, A novel method of adverse event detection can accurately identify venous thromboembolisms (VTEs) from narrative electronic health record data., J. Am. Med. Inform. Assoc. 22 (2015) 155–65. https://doi.org/10.1136/amiajnl-2014-002768.
26. Y. Kim, Convolutional Neural Networks for Sentence Classification, (2014). http://arxiv.org/abs/1408.5882 (accessed March 20, 2019).
27. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, É. Duchesnay, Scikit-learn: Machine Learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
28. M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G.S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mane, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viegas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, X. Zheng, TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems, (2016).
29. K.S. Jones, K.S. Jones, A statistical interpretation of term specificity and its application in retrieval, J. Doc. 28 (1972) 11--21.
30. A. Rajaraman, J.D. Ullman, Mining of massive datasets, Cambridge University Press, 2012.
31. T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient Estimation of Word Representations in Vector Space, (2013). http://arxiv.org/abs/1301.3781 (accessed November 16, 2018).
32. W. Rawat, Z. Wang, Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review, Neural Comput. 29 (2017) 2352–2449. https://doi.org/10.1162/neco_a_00990.
33. M. Lin, Q. Chen, S. Yan, Network In Network, n.d. https://arxiv.org/pdf/1312.4400.pdf (accessed March 20, 2019).

34. A.B. Duque, D. Macêdo, L. Lázaro, J. Santos, C. Zanchettin, Squeezed Very Deep Convolutional Neural Networks for Text Classification, n.d. https://arxiv.org/pdf/1901.09821.pdf (accessed March 20, 2019).
35. E. Mayoraz, E. Alpaydm, Support Vector Machines for Multi-class Classification, n.d.
36. L. Breiman, Random Forests, 2001.
37. K.K. Dobbin, R.M. Simon, Optimally splitting cases for training and testing high dimensional classifiers, BMC Med. Genomics. 4 (2011) 31. https://doi.org/10.1186/1755-8794-4-31.
38. A. Gerstmair, P. Daumke, K. Simon, M. Langer, E. Kotter, Intelligent image retrieval based on radiology reports, Eur. Radiol. 22 (2012) 2750–2758. https://doi.org/10.1007/s00330-012-2608-x.
39. R.M. Hogg, D. Denison, A history of the English language, Cambridge University Press, 2006.
40. D.D.B. Bates, J.N. Tkacz, C.A. LeBedis, N. Holalkere, Suboptimal CT pulmonary angiography in the emergency department: a retrospective analysis of outcomes in a large academic medical center, Emerg. Radiol. 23 (2016) 603–607. https://doi.org/10.1007/s10140-016-1425-y.

CHAPTER 3

Automated Detection of Pulmonary Embolism in CT Pulmonary Angiograms using an AI-powered Algorithm

BASED ON: Weikert T, Winkel DJ, Bremerich J, Stieltjes B, Parmar V, Sauter AW, Sommer G. Automated detection of pulmonary embolism in CT pulmonary angiograms using an AI-powered algorithm. Eur Radiol. 2020 Dec;30(12):6545-6553.

Abstract

Objectives: To evaluate the performance of an AI-powered algorithm for the automatic detection of pulmonary embolism (PE) on chest computed tomography pulmonary angiograms (CTPAs) on a large dataset.

Materials and Methods: We retrospectively identified all CTPAs conducted at our institution in 2017 (n = 1499). Exams with clinical questions other than PE were excluded from the analysis (n = 34). The remaining exams were classified into positive (n = 232) and negative (n = 1233) for PE based on the final written reports, which defined the reference standard. The fully anonymized 1-mm series in soft tissue reconstruction served as input for the PE detection prototype algorithm that was based on a deep convolutional neural network comprising a ResNet architecture. It was trained and validated on 28,000 CTPAs acquired at other institutions. The result series were reviewed using a web-based feedback platform. Measures of diagnostic performance were calculated on a per patient and a per finding level.

Results: The algorithm correctly identified 215 of 232 exams positive for pulmonary embolism (sensitivity 92.7%; 95% confidence interval [CI] 88.3–95.5%) and 1178 of 1233 exams negative for pulmonary embolism (specificity 95.5%; 95% CI 94.2– 96.6%). On a per finding level, 1174 of 1352 findings marked as embolus by the algorithm were true emboli. Most of the false positive findings were due to contrast agent–related flow artifacts, pulmonary veins, and lymph nodes.

Conclusion: The AI prototype algorithm we tested has a high degree of diagnostic accuracy for the detection of PE on CTPAs. Sensitivity and specificity are balanced, which is a prerequisite for its clinical usefulness.

Abbreviations

ADMIRE	Advanced Modeled Iterative Reconstruction CAD Computer-assisted detection
CTPA	CT pulmonary angiogram
DCNN	Deep convolutional neural network
DICOM	Digital imaging and communications in medicine
FN	False negative
FP	False positive
FTE	File transfer protocol
IR	Iterative reconstruction
NPV	Negative predictive value
PACS	Picture Archiving and Communication System
PE	Pulmonary embolism
PGY	Postgraduate year
PPV	Positive predictive value
SAFIRE	Sinogram Affirmed Iterative Reconstruction
SWCCE	Sample weighted categorical cross-entropy
TP	True positive

1. Introduction

In times of increasing hospital admission rates and numbers of computer tomography (CT) scans performed at emergency departments [1, 2], swift diagnosis and communication of critical findings is becoming one of the main challenges in radiology. To give a concrete example, at our department, the number of computed tomography pulmonary angiograms (CTPAs)—the standard diagnostic procedure for the evaluation of suspected pulmonary embolism (PE) [3]—increased by 32.7% from 2014 (n = 1130) to 2017 (n = 1499). In 2017, however, only 15.8% of CTPAs with the clinical question of PE were positive. At other centers, the share of CTPAs that contain critical findings is even lower [4]. Fast detection of these cases amidst a large number of unremarkable scans is crucial for patients faced with this potentially life-threatening condition, as an early onset of anticoagulation therapy is associated with better outcomes [5]. An important contribution to this end can be made by worklist prioritization.

This challenge can be addressed using algorithms for automated detection such as deep convolutional neural networks (DCNNs). One strength of these algorithms is pattern recognition [6]. A properly working neural network can assist radiologists by highlighting exams positive for PE in the worklist, thereby speeding up the diagnostic and communication workflow. Recent studies showed good results of DCNNs in the detection of critical findings in CT scans, among others, for intracranial hemorrhage [7], acute brain ischemia [8], and critical abdominal findings [9]. Good detection and segmentation performance of DCNNs was also reported for other findings and modalities [10–16]. Therefore, we hypothesized that the algorithm we tested is able to identify PE in CTPAs. Previously presented computer-assisted detection (CAD) algorithms for the automatic detection of PE on CTPAs had either relatively low sensitivity [17–21] or high rates of false positive findings [22–31] and mostly operated on small and/or unbalanced datasets [19,20,22–24,30,32–35].

The purpose of our study was to evaluate the performance of a trained and validated DCNN-based prototype algorithm for automated detection of PE in CTPAs on data reflecting clinical reality using all relevant CTPAs conducted at our department in one year.

2. Materials and Methods

This study was approved by the IRB (project number 2019-01050) and written informed consent was waived. The PE detection algorithm was a prototype algorithm provided by Aidoc Medical. There was no financial support for this study. The authors had control of the data and the information submitted for publication at all time.

2.1 Case selection

We retrospectively extracted all CTPAs conducted at our institution in 2017 using protocol name and time period (01 January 2017–31 December 2017) as criteria. Subsequently, we downloaded all reports and axial 1-mm slices in soft tissue reconstruction with an in-house-developed Radiology Information System/Picture Archiving and Communication System (PACS) search engine (n = 1499). In a next step, a radiology resident (T.W., third year of residency [PGY-3]) reviewed all reports. Reports with another clinical question than PE (n = 34) were excluded. Exams with suboptimal contrast filling as mentioned in the report (n = 28) were deliberately included into the testing dataset, as they reflect clinical realities. The remaining exams were manually categorized into positive (n = 232) and negative (n = 1233) for PE based on the reports. Furthermore, we determined the location of the most central embolus in each exam (central, segmental, subsegmental). All resulting 1465 exams of 2017 were then used for testing the performance of the algorithm. Figure 1 illustrates the study workflow.

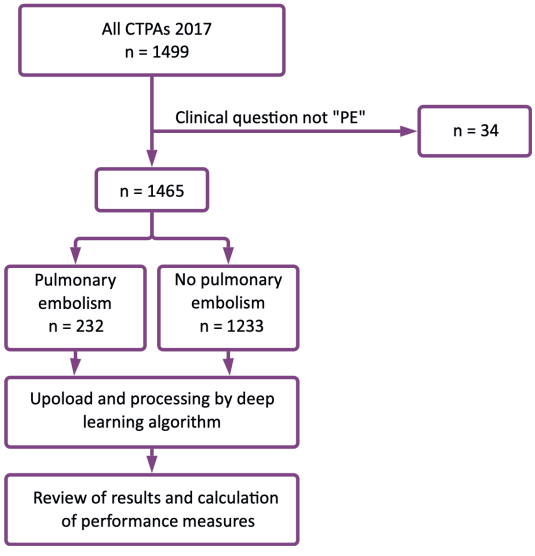


Fig. 1 Study workflow

2.2. Image acquisition

The CTPAs were acquired on three scanners: Somatom Definition AS+ (128 slices; iterative reconstruction algorithm (IR): Sinogram Affirmed Iterative Reconstruction [SAFIRE]), Somatom Definition Edge (128 slices; IR: Advanced Modeled Iterative Reconstruction [ADMIRE]) and Somatom Definition FLASH (2 × 128 slices; IR: SAFIRE; all scanners and IR: Siemens Healthineers AG). Pitch factor was 1.5; collimation was 0.6 mm. The dataset used for automated detection of PE were in soft tissue kernel (I26f) with a slice thickness of 1.0 mm. Mean peak X-ray tube voltage was 108.0 kVp (SD 12.3 kVp). As contrast agent, iopromide (Bayer AG) at a standard injection rate of 4.0 mL/s (mean injection rate 3.9 mL/s; SD 0.28 mL/s) and an amount of up to 70 mL (mean 66.8 mL; SD 12.0 mL; bolus tracking technique) was used.

2.3. AI algorithm training and validation

The cloud-based PE detection prototype algorithm, a fast region-based convolutional neural network, was trained and validated on 28,000 independent CTPAs from 9 other medical centers, acquired on 17 different scanner models from 4 vendors (Canon Medical Systems Corporation, GE Healthcare, Philips Healthcare, Siemens Healthineers AG). Training data was generated by board-certified radiologists. The algorithm consists of two stages: a region proposal stage and a false positive reduction stage. The first stage is a 3D DCNN. Its architecture is based on the Resnet architecture [36], which consists of repeated blocks of several convolutional layers with skip connections between them, followed by a pooling layer. This network is trained on segmented scans and produces a 3D segmentation map. The model was trained from scratch. From the segmentation map, region proposals are generated and passed as input to the second stage of the algorithm. The second stage classifies each region as positive or negative, based on features from the last layer of the first stage and traditional image processing methods. A designated loss function was developed to minimize false positive findings due to subsegmental location and suboptimal contrast timing (sample weighted categorical cross-entropy [SWCCE]). The loss function is a central concept in ML and evaluates how precise an algorithm models given data. Of the training cases, 43.4% contained pulmonary embolism. Training was performed on three kinds of AWS EC2 servers with up to 8NVIDIA GPUs (P3.2× large [1 GPU], P3.8× large [4 GPUs], and P3.16× large [8 GPUs]).

2.4. Image data processing

After anonymization, DICOM image data was uploaded to a cloud server through a secure FTP (File Transfer Protocol) connection. Uploading times were recorded. The 1-mm series in axial plane were preprocessed (normalization, z-axis resizing). They served as only input for the core of the PE detection algorithm. Finally, the original series with superimposed arrows indicating suspected findings were uploaded to a web-based platform for review.

2.5. AI algorithm testing

The performance of the prototype algorithm was tested on a per patient and per finding level. The written reports of the CTPAs, which had been approved by at least two physicians of the department of radiology, at least one of them being board-certified, served as reference standard for the analysis on the per-patient level. On a per finding level, the reference standard was established by the visual review of cases in synopsis with the written reports as described below. The results of the PE detection software were defined as index test in both cases. The visual review of the algorithms output was performed by a 3rd year radiology resident (T.W.; PGY-3) using a web-based validation platform displayed on a conventional medical imaging workstation (screen resolution: 1600 × 1200 pixels). Indeterminate findings were second read by a board-certified radiologist with 4 years of professional experience (GS), who could overrule the assessment of the resident. Exams negative for PE both according to the report and the algorithm were accepted to be true negatives and not checked visually. All images of other exams were reviewed on the web-based platform. Figure 2 shows how the results were displayed on the validation platform.

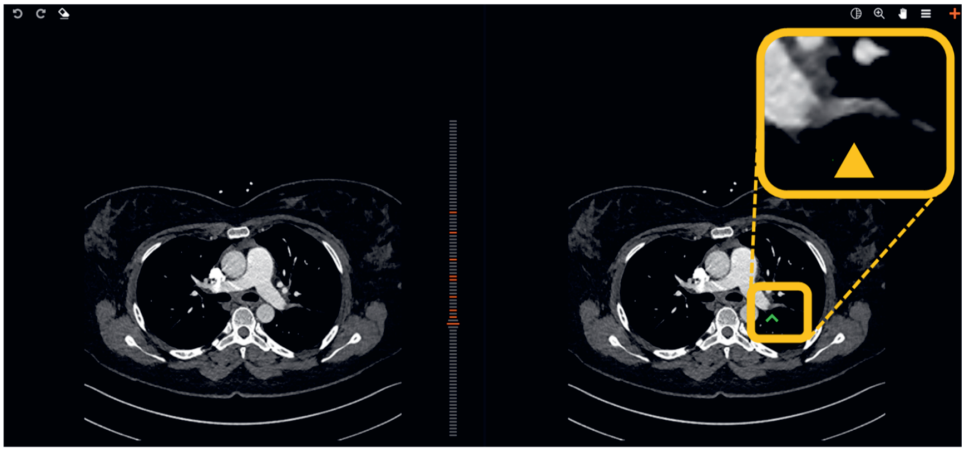


Fig. 2 Setup of the testing platform with correctly marked thrombus. A box with a magnified illustration of the finding has been added to improve visibility

During the reading of the complete image stack, findings of the algorithm were both compared with the clinically approved report and visually checked for plausibility. The findings were either confirmed (true positive finding, TP) or rejected (false positive finding, FP) by clicking on a check-mark or rejection-mark. For FPs, the reader described the underlying reason in an associated free text field according to a previously defined standard terminology (e.g., contrast agent-related flow artifact). Furthermore, bounding boxes were used to mark all emboli that had been missed by the algorithm (false negative findings, FN). The per finding analysis was conducted as "per clot" analysis meaning that continuous emboli over multiple segments were marked once. The processing times of the algorithm were recorded.

2.6. Statistical testing

On a per patient level, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), accuracy, and the F1 score for the detection of PE were calculated. For sensitivity and specificity, 95% CIs were calculated based on Wilson score intervals [37]. The F1 score is a harmonic average of PPV and sensitivity:

$$\mathrm{F1} = \left(\frac{PPV^{-1} + Sensitivity^{-1}}{2}\right)^{-1}$$

To further investigate the influence of embolus localization (central vs. segmental vs. subsegmental), the distinct detection rates of exams with (A) central emboli (and optionally also segmental and subsegmental emboli), (B) segmental emboli (and optionally also subsegmental emboli), and (C) subsegmental emboli only, according to the written report, were calculated.

On a per finding level ("per embolus"), we determined the number of true positives, false positives, and false negatives as well as calculated the false positive rate per patient and PPV. For calculations, Excel 2010 (Microsoft Corp.) and SPSS Statistics, Version 22.0 (IBM Corp.) were used. P-values < 0.05 were considered to be statistically significant.

3. Results

A total of 1465 exams of 2017 were used to evaluate the prototype algorithm. Mean age of the patients was 66.0 years (SD 16.9 years). There were 691/1465 male patients (47.2%) and

774/1465 female patients (52.8%). The rate of exams positive for PE was 15.8% (232/1465). The most proximal location of an embolus was central in 23/232 (= 9.9%) segmental in 163/232 (= 70.3%), and subsegmental in 46/232 (= 19.8%). Average uploading time per case was 25 s (SD 8 s), average processing time was 152 s (SD 17 s). There was no missing data.

3.1. Per patient

The performance measures of the algorithm for the detection of PE on a per patient level are displayed in Table 1. Of 232 cases being positive for PE according to the report, 215 were correctly flagged by the algorithm (TP). There were 55 exams incorrectly rated as positive (FP). Of 1233 cases being negative for PE according to the report, 1178 were correctly flagged (TN). There were 17 exams incorrectly rated as negative (FN). The subanalysis revealed that exams containing central emboli had the highest detection rates with 95.7% (95% CI 88.0–99.1%), followed by those containing segmental emboli with 93.3% (95% CI 87.3– 97.1%). Exams containing emboli in subsegmental location only had the lowest detection rate with 85.7% (95% CI 85.7–94.6%).

Table 1 Algorithm performance measures on a per patient level

	Specificity (95% CI)	PPV (95% CI)	NPV (95% CI)	F1 score
92.7% (215/232) (88.5–95.7%)	95.5% (1178/1233) (94.2–96.6%)	79.6% (215/270) (75.1–83.5%)	98.6% (1178/1195) (88.5–95.7%)	0.86

CI confidence interval, *PPV* positive predictive value, *NPV* negative predictive value

3.2. Per finding

Of 1352 findings marked by the algorithm, 1174 were true emboli (PPV 86.8%). There were 178 FPs in a total of 1465 exams, corresponding to a FP rate per case of 0.12. Table 2 specifies reasons for FPs, the three most frequent being the confusion of contrast agent-related flow artifacts, pulmonary veins, and lymph nodes with emboli. Figure 3 displays one example for each of the three most common reasons of FP findings. There were 203 FNs that were retrospectively marked by the reviewer. Figure 4 shows two emboli that were missed by the algorithm. Table 3 compares the results of our study with previous studies.

Table 2 Reasons for false positive findings

Finding confused with pulmonary embolus	*n*
Contrast agent-related flow artifact	46
Pulmonary vein	32
Intrapulmonary lymph node	29
Pulmonary infiltrate	20
Beam hardening artifact	12
Pulmonary metastasis	10
Breathing artifact	8
Bronchus	7
Hilar soft tissue	4
Other*	10
Sum	178

*Pleural effusion, atelectasis, fibrosis, vena azygos, granuloma, pulmonary artery bifurcation, and no obvious reason

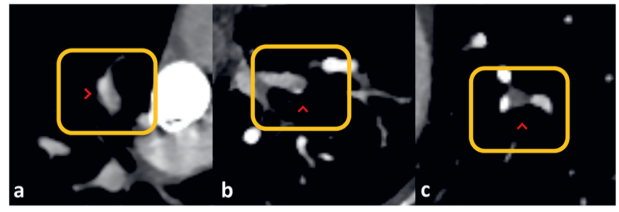


Fig. 3 Display of contrast-enhanced CT images of four FP findings indicated by an arrow and a box. These are due to a contrast agent-related flow artifact (**a**), detection of a pulmonary vein (**b**), and detection of an intrapulmonary lymph node (**c**)

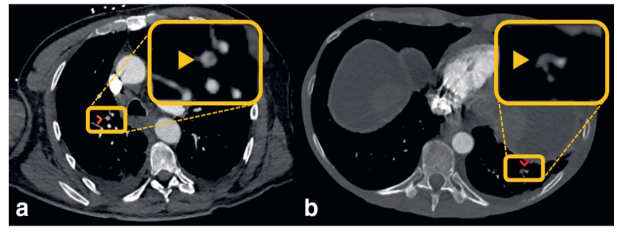


Fig. 4 Contrast-enhanced CT images of FN findings in segmental (**a**) and subsegmental (**b**) location. Small arrows and boxes indicate the embolus missed by the detection algorithm. Big boxes provide a magnified view of the finding.

Table 3 Performance of CAD systems for PE in literature with patient number and number of false positives per case in comparison to the prototype algorithm tested

Author	Year	Sensitivity in %	Level of analysis	False positives per case	Patients (*n*)
Masutani et al [22]	2002	100.0/85.0*	Case	7.7/2.6	19
Pichon et al [23]	2004	86	Finding	6.3	6
Digumarthi et al [27]	2006	88	Finding	4.0	39
Schoepf et al [28]	2007	90/92**	Finding	4.8	36
Liang and Bi [24]	2007	90.1	Finding	40.3	177
Maizlin et al [17]	2007	53.3	Case	1.0	104
Engelke et al [18]	2008	30.7	Finding	4.1	56
Das et al [33]	2008	83	Finding	4.0	43
Bouma et al [19]	2009	73/63/58*	Finding	15/4.9/4.0	19
Zhou et al [25]	2009	80/80***	Finding	18.9/22.6	59/69
Park et al [32]	2011	79.2	Finding	262.5	20
Wittenberg et al [29]	2010	94	Case	4.7	292
Lee et al [20]	2011	37.5	Finding	3.5	37
Blackmon et al [30]	2011	79.0/77.7/72.7/69.2****	Finding	3.3	79
Wittenberg et al [31]	2012	96.1	Case	4.9	209
Kligerman et al [21]	2014	91.7/65.5**	Case	3.8	53
Lahiji et al [35]	2014	97.5–55.4*****	Case	1.5–3.6	66
Özkan et al [26]	2014	95.1/90.0/61.0*	Finding	14.4/12.4/8.2	33
Tajbakhsh et al [40]	2015	83.4/34.6***	Finding	2.0/2.0	121/20
Tajbakhsh et al [41]	2019	33.0	Finding	2.0	20
Algorithm tested	**2020**	**92.7**	**Case and Finding**	**0.12**	**1465**

*Multiple thresholds

**Segmental/subsegmental

***Multiple training and/or testing data

****Central/lobar/segmental/subsegmental

*****Multiple reconstructions

4. Discussion

This study assessed the clinical performance of an algorithm capable of automatically detecting PE on CTPAs. It reached high diagnostic accuracy on the per examination level (sensitivity 92.7%; 215/232) at an FP rate of only 0.12 per case. As for human readers, the detection rates for exams containing central emboli (95.7%) were slightly superior to the one for exams with subsegmental emboli only (85.7%). Testing on all relevant CTPAs conducted at our institution within one year ensured the exact representation of clinical realities regarding the ratio of positive and negative exams and the distribution of emboli. This and the high number of cases included in the analysis (n = 1465) are unique features of this study.

Previous CAD solutions for the automatic detection of PE on CTPAs suffer from three shortcomings: first, many algorithms achieved rather low sensitivities [17,19,20,32,34]. In a clinical setting, this limits their usefulness, because immediate communication to referring physicians is needed in these cases. Second, studies reporting sensitivity levels > 85% accepted many FP findings of up to 292 per case [22–24,27–31,35,38]. While such algorithms increase the detection rate of PE, FP findings increase workload for radiologists

[39], thereby delaying diagnostic workup of other patients. In addition, extra workload decreases the acceptance among radiologists. Third, many studies tested on datasets that either strongly overweighted the number of PE-positive CTPAs compared to the ratio found in clinical practice [20,30,33–35] or even used solely PE-positive exams [19,24]. This is a potential cause of distortion to the measures of diagnostic accuracy. Furthermore, many of the evaluations operated on small sample size with less than 50 cases in the testing set [19,20,22,23,26,28,32,33].

While the previously mentioned CAD algorithms used traditional image processing techniques such as segmentation and thresholding [19,23–26], segmentation and feature analysis [18,21,22,29–31,33,35], or segmentation, thresholding, and feature analysis [17,27,28], only one group used a DCNN to detect PE [40]. Tajbakhsh and colleagues reported a sensitivity of 83% at 2 FP findings per case. However, when tested on an independent dataset, sensitivity dropped to 34.6% at 2 FP findings per case, rendering the solution unsuitable for direct clinical deployment. The same is true for an extended approach of the same group that used vessel-oriented representations of emboli candidates as input to serveral CNNs [41].

The F1 score of 0.86 on a per patient level indicates a balanced performance of the algorithm used in our study with respect to sensitivity and PPV. This is a prerequisite for the usefulness of the algorithm in clinical practice. Given this, it can serve as foundation for an automated worklist prioritization that is adding value by speeding up the diagnostic workflow without increasing the workload for the radiologist. Concretely, a notification can be sent to the radiologist in charge via email or a pop-up window whenever the algorithm prototype suspects the presence of pulmonary embolism in a CTPA and thereby initiate prompt reading of the case. Because time to therapy initiation is outcome relevant in patients presenting with PE [42], this can improve quality of care. The increasing number of exams performed and the use of teleradiology cause situations in which a radiologist is confronted with multiple potentially urgent exams at the same time. In these situations, an assisting tool can make sure that the communication of urgent findings to the referrers is not delayed. However, there is no general agreement on what level of performance can be considered sufficient for an algorithm to be deployed in a clinical setting, and further discussion among clinicians is warranted. Our subanalysis on the detection rates' dependence on the localization of the emboli revealed that CTPAs

containing central emboli were more likely to be detected (95.7%) compared to those exams with subsegmental emboli only (85.7%). This is in line with other studies on automated PE detection [18,20,25,28,30,33] as well as with reports on the performance of radiologists [20,33].

There are several limitations to our work. First, the evaluation was performed retrospectively on data acquired on scanners of one vendor. The performance on image data provided by systems of other vendors might differ. However, CTPA protocols are highly standardized, which makes generalizability probable. Second, the categorization of exams regarding the presence of PE and therefore the reference standard for the per patient evaluation was the clinically approved report. Initially wrongly classified exams regarding PE would therefore affect performance measures. Nevertheless, considering that identifying PE on a CTPA is straight-forward and all reports had been approved by at least two physicians of the department of radiology, at least one board-certified, the influence of this on the results is expected to be small. In addition, this approach of using clinically approved reports as standard of reference has also been applied in a range of other recent studies [7,43,44]. Third, the prevalence of CTPAs positive for PE at our institution in 2017 was 15.8%. It is well known that this ratio varies geographically. As prevalence has a strong influence on PPV and NPV, this might translate into an application site–dependent performance. However, because the pretest probability of 15.8% found at our institution is at the center of the spectrum of pretest probabilities worldwide [4,45], it is expected that the good performance of the prototype algorithm at the per-exam level will be replicable at other centers. Fourth, the fact that the emboli of the cases with PE were not marked prior to the processing by the algorithm, but afterwards during the check of the algorithms results might negatively affect the validity of the results on a per finding level. However, the whole image stack of all cases (TP, FP, FN) was subsequently thoroughly reviewed and all PE marks of the algorithm were checked for plausibility. Furthermore, all FN findings were marked.

In conclusion, the prototype algorithm tested on a large dataset exhibits a high diagnostic performance for the automatic detection of CTPAs containing PE. Of note, the performance is balanced in regard to sensitivity and specificity. As such, it constitutes a strong foundation for a clinical decision support tool that can speed up the diagnostic workup of critical cases by complementing traditional ways of worklist prioritization. To

what extent this contributes to a better quality of healthcare provision remains to be investigated by further prospective trials.

References

1. Brown J, Shesser R (2004) Computed tomography scan use is rising faster than other investigational modalities in the emergency department evaluation of patients. Ann Emerg Med 44:S33. https:// doi.org/10.1016/j.annemergmed.2004.07.109
2. Kocher KE, Meurer WJ, Fazel R, Scott PA, Krumholz HM, Nallamothu BK (2011) National trends in use of computed tomography in the emergency department. Ann Emerg Med 58:452– 462.e3. https://doi.org/10.1016/J.ANNEMERGMED.2011.05.020
3. Estrada-Y-Martin RM, Oldham SA (2011) CTPA as the gold standard for the diagnosis of pulmonary embolism. Int J Comput Assist Radiol Surg 6:557–563. https://doi.org/10.1007/s11548-010-05264
4. Pernod G, Caterino J, Maignan M et al (2017) D-dimer use and pulmonary embolism diagnosis in emergency units: why is there such a difference in pulmonary embolism prevalence between the United States of America and countries outside USA? PLoS One 12:e0169268. https://doi.org/10.1371/journal.pone.0169268
5. Smith SB, Geske JB, Maguire JM, Zane NA, Carter RE, Morgenthaler TI (2010) Early anticoagulation is associated with reduced mortality for acute pulmonary embolism. Chest 137: 1382–1390
6. Lecun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. Proc IEEE 86:2278– 2324. https://doi.org/10.1109/5.726791
7. Prevedello LM, Erdal BS, Ryu JL et al (2017) Automated critical test findings identification and online notification system using artificial intelligence in imaging. Radiology 285:923–931. https://doi. org/10.1148/radiol.2017162664
8. Nagel S, Sinha D, Day D et al (2017) e-ASPECTS software is noninferior to neuroradiologists in applying the ASPECT score to computed tomography scans of acute ischemic stroke patients. Int J Stroke 12:615–622. https://doi.org/10.1177/1747493016681020
9. Winkel DJ, Heye T, Weikert TJ, Boll DT, Stieltjes B (2018) Evaluation of an AI-based detection software for acute findings in abdominal computed tomography scans. Invest Radiol:1. https:// doi.org/10.1097/RLI.0000000000000509
10. Liu K, Li Q, Ma J et al (2019) Evaluating a fully automated pulmonary nodule detection approach and its impact on radiologist performance. Radiol Artif Intelli:e180084. https://doi.org/10. 1148/ryai.2019180084
11. Vorontsov E, Cerny M, Régnier P et al (2019) Deep learning for automated segmentation of liver lesions at CT in patients with colorectal cancer liver metastases. Radiol Artif Intell 1:180014. https://doi.org/10.1148/ryai.2019180014 DOI
12. Thian YL, Li Y, Jagmohan P, Sia D, Chan VE, Tan RT (2019) Convolutional neural networks for automated fracture detection and localization on wrist radiographs. Radiol Artif Intell 1:e180001. https://doi.org/10.1148/ryai.2019180001
13. Li L, Liu Z, Huang H, Lin M, Luo D (2018) Evaluating the performance of a deep learning-based computer-aided diagnosis (DL-CAD) system for detecting and characterizing lung nodules: comparison with the performance of double reading by radiologists. Thorac Cancer 10:1759–7714.12931. https://doi.org/10.1111/1759-7714.12931
14. Ekert T, Krois J, Meinhold L et al (2019) Deep learning for the radiographic detection of apical lesions. J Endod. https://doi.org/10.1016/j.joen.2019.03.016
15. Cheng C-T, Ho T-Y, Lee T-Y et al (2019) Application of a deep learning algorithm for detection and visualization of hip fractures on plain pelvic radiographs. Eur Radiol. https://doi.org/10.1007/s00330-019-06167-y
16. Ye W, Gu W, Guo X et al (2019) Detection of pulmonary ground-glass opacity based on deep learning computer artificial intelligence. Biomed Eng Online 18:6. https://doi.org/10.1186/s12938-019-0627-4
17. Maizlin ZV, Vos PM, Godoy MB, Cooperberg PL (2007) Computer-aided detection of pulmonary embolism on CT angiography. J Thorac Imaging 22:324–329. https://doi.org/10.1097/RTI.0b013e31815b89ca

18. Engelke C, Schmidt S, Bakai A, Auer F, Marten K (2008) Computer-assisted detection of pulmonary embolism: performance evaluation in consensus with experienced and inexperienced chest radiologists. Eur Radiol 18:298–307. https://doi.org/10.1007/s00330-007-0770-3
19. Bouma H, Sonnemans JJ, Vilanova A, Gerritsen FA (2009) Automatic detection of pulmonary embolism in CTA images. IEEE Trans Med Imaging 28:1223–1230. https://doi.org/10.1109/TMI.2009.2013618
20. Lee CW, Seo JB, Song J-W et al (2011) Evaluation of computer-aided detection and dual energy software in detection of peripheral pulmonary embolism on dual-energy pulmonary CT angiography. Eur Radiol 21:54–62. https://doi.org/10.1007/s00330-010-1903-7
21. Kligerman SJ, Lahiji K, Galvin JR, Stokum C, White CS (2014) Missed pulmonary emboli on CT angiography: assessment with pulmonary embolism–computer-aided detection. Am J Roentgenol 202:65–73. https://doi.org/10.2214/AJR.13.11049
22. Masutani Y, MacMahon H, Doi K (2002) Computerized detection of pulmonary embolism in spiral CT angiography based on volumetric image analysis. IEEE Trans Med Imaging 21:1517–1523. https://doi.org/10.1109/TMI.2002.806586
23. Pichon E, Novak CL, Kiraly AP, Naidich DP (2004) A novel method for pulmonary emboli visualization from high-resolution CT images. Proceedings of the SPIE, Volume 5367, p 161-170 (2004). p 161
24. Liang J, Bi J (2007) Computer aided detection of pulmonary embolism with tobogganing and mutiple instance classification in CT pulmonary angiography. Inf Process Med Imaging 20:630–641
25. Zhou C, Chan H-P, Sahiner B et al (2009) Computer-aided detection of pulmonary embolism in computed tomographic pulmonary angiography (CTPA): performance evaluation with independent data sets. Med Phys 36:3385–3396. https://doi.org/10.1118/1.3157102
26. Özkan H, Osman O, Şahin S, Boz AF (2014) A novel method for pulmonary embolism detection in CTA images. Comput Methods Programs Biomed 113:757–766. https://doi.org/10.1016/j.cmpb.2013.12.014
27. Digumarthy S, Kagay C, Legasto A, Muse V, Wittram C, Shepard J (2006) Computer-aided detection (CAD) of acute pulmonary emboli: evaluation in patients without significant pulmonary disease. Radiological Society of North America 2006 Scientific Assembly and Annual Meeting, November 26 December 1, 2006, Chicago IL
28. Schoepf UJ, Schneider AC, Das M, Wood SA, Cheema JI, Philip Costello P (2007) Pulmonary embolism: computer-aided detection at multidetector row spiral computed tomography. J Thorac Imaging 22:319–323. https://doi.org/10.1097/RTI.0b013e31815842a9
29. Wittenberg R, Peters JF, Sonnemans JJ, Prokop M, Schaefer-Prokop CM (2010) Computer-assisted detection of pulmonary embolism: evaluation of pulmonary CT angiograms performed in an on-call setting. Eur Radiol 20:801–806. https://doi.org/10.1007/s00330-009-1628-7
30. Blackmon KN, Florin C, Bogoni L et al (2011) Computer-aided detection of pulmonary embolism at CT pulmonary angiography: can it improve performance of inexperienced readers? Eur Radiol 21:1214–1223. https://doi.org/10.1007/s00330-010-2050-x DOI
31. Wittenberg R, Berger FH, Peters JF et al (2012) Acute pulmonary embolism: effect of a computer-assisted detection prototype on diagnosis—an observer study. Radiology 262:305–313. https://doi.org/10.1148/radiol.11110372
32. Sang Cheol Park SC, Chapman BE, Bin Zheng B (2011) A multistage approach to improve performance of computer-aided detection of pulmonary embolisms depicted on CT images: preliminary investigation. IEEE Trans Biomed Eng 58:1519–1527. https://doi.org/10.1109/TBME.2010.2063702
33. Das M, Mühlenbruch G, Helm A et al (2008) Computer-aided detection of pulmonary embolism: influence on radiologists' detection performance with respect to vessel segments. Eur Radiol 18:1350–1355. https://doi.org/10.1007/s00330-008-0889-x

34. Lee G, Lee HY, Park H et al (2017) Radiomics and its emerging role in lung cancer research, imaging biomarkers and clinical management: state of the art. Eur J Radiol 86:297–307. https://doi.org/10.1016/J.EJRAD.2016.09.005
35. Lahiji K, Kligerman S, Jeudy J, White C (2014) Improved accuracy of pulmonary embolism computer-aided detection using iterative reconstruction compared with filtered back projection. AJR Am J Roentgenol 203:763–771. https://doi.org/10.2214/AJR.13.11838
36. He K, Zhang X, Ren S, Sun J (2015) Deep residual learning for image recognition. Proceedings of the IEEE Conference on Cumputer Vision and Pattern Recognition: 770-778. https://doi.org/10.1109/CVPR.2016.90
37. Wilson EB (1927) Probable inference, the law of succession, and statistical inference. J Am Stat Assoc 22:209. https://doi.org/10.2307/2276774
38. Wittenberg R, Peters JF, van den Berk IAH et al (2013) Computed tomography pulmonary angiography in acute pulmonary embolism. J Thorac Imaging 28:315–321. https://doi.org/10.1097/RTI.0b013e3182870b97
39. Bhargavan M, Kaye AH, Forman HP, Sunshine JH (2009) Workload of radiologists in United States in 2006-2007 and trends since 1991-1992. Radiology 252:458–467. https://doi.org/10.1148/radiol.2522081895
40. Tajbakhsh N, Gotway MB, Liang J (2015) Computer-aided pulmonary embolism detection using a novel vessel-aligned multiplanar image representation and convolutional neural networks. In: Navab N, Hornegger J, Wells W, Frangi A (eds) Medical image computing and computer-assisted intervention --MICCAI 2015. MICCAI 2015. Lecture notes in computer science, vol 9350. Springer, Cham
41. Tajbakhsh N, Shin JY, Gotway MB, Liang J (2019) Computer-aided detection and visualization of pulmonary embolism using a novel, compact, and discriminative image representation. Med Image Anal 58:101541. https://doi.org/10.1016/J.MEDIA.2019.101541
42. Beydilli İ, Yılmaz F, Sönmez BM et al (2016) Thrombolytic therapy delay is independent predictor of mortality in acute pulmonary embolism at emergency service. Kaohsiung J Med Sci 32:572–578. https://doi.org/10.1016/j.kjms.2016.09.004
43. Annarumma M, Withey SJ, Bakewell RJ, Pesce E, Goh V, Montana G (2019) Automated triaging of adult chest radiographs with deep artificial neural networks. Radiology:291
44. Rayan JC, Reddy N, Kan JH, Zhang W, Annapragada A (2019) Binomial classification of pediatric elbow fractures using a deep learning multiview approach emulating radiologist decision making. Radiol Artif Intell 1:e180015. https://doi.org/10.1148/ryai.2019180015
45. Penaloza A, Kline J, Verschuren F et al (2012) European and American suspected and confirmed pulmonary embolism populations: comparison and analysis. J Thromb Haemost 10:375–381. https://doi.org/10.1111/j.1538-7836.2012.04631.x

CHAPTER 4

Assessment of a Deep Learning Algorithm for the Detection of Rib Fractures on Whole Body Trauma CT

BASED ON: Weikert T, Noordtzij LA, Bremerich J, Stieltjes B, Parmar V, Cyriac J, Sommer G, Sauter AW. Assessment of a Deep Learning Algorithm for the Detection of Rib Fractures on Whole-Body Trauma Computed Tomography. Korean J Radiol. 2020 Jul;21(7):891-899. doi: 10.3348/kjr.2019.0653.

Abstract

Objective: To assess the diagnostic performance of a deep learning-based algorithm for automated detection of acute and chronic rib fractures on whole-body trauma CT.

Materials and Methods: We retrospectively identified all whole-body trauma CT scans referred from the emergency department of our hospital from January to December 2018 (n = 511). Scans were categorized as positive (n = 159) or negative (n = 352) for rib fractures according to the clinically approved written CT reports, which served as the index test. The bone kernel series (1.5-mm slice thickness) served as an input for a detection prototype algorithm trained to detect both acute and chronic rib fractures based on a deep convolutional neural network. It had previously been trained on an independent sample from eight other institutions (n = 11455).

Results: All CTs except one were successfully processed (510/511). The algorithm achieved a sensitivity of 87.4% and specificity of 91.5% on a per-examination level [per CT scan: rib fracture(s): yes/no]. There were 0.16 false-positives per examination (= 81/510). On a per-finding level, there were 587 true-positive findings (sensitivity: 65.7%) and 307 false-negatives. Furthermore, 97 true rib fractures were detected that were not mentioned in the written CT reports. A major factor associated with correct detection was displacement.

Conclusion: We found good performance of a deep learning-based prototype algorithm detecting rib fractures on trauma CT on a per-examination level at a low rate of false-positives per case. A potential area for clinical application is its use as a screening tool to avoid false-negative radiology reports.

1. Introduction

Rib fractures are a common finding after thoracic trauma, occurring in approximately 40% of these patients (1). Efforts in the polytrauma setting are focused on the detection of life-threatening conditions, such as aortic dissection and organ laceration. This fact in combination with the lack of time (2), a noisy work environment (3), satisfaction of search (4), and the frequent co-occurrence of multiple traumas on whole-body computed tomography (CT) (5) lead to a significant number of missed rib fractures in this setting (6). While most rib fractures heal without surgical intervention (7), there are three reasons why it is nonetheless important to detect rib fractures: first, they are indicators of trauma-associated conditions that require immediate treatment, such as pneumothorax, and their onset can be delayed for several days (8). Second, often as a consequence of inadequate pain control, respiratory complications, such as posttraumatic pneumonia occur secondary to rib fractures (9, 10). Finally, the number and type of rib fractures can be a basis for further treatment strategies (11, 12). Thus, an accurate detection of rib fractures on CT scans contributes to appropriate patient care (13). To address the problem of missed rib fractures on trauma CT, some authors have proposed multiplanar (6) and rib unfolding reconstructions (14). A complementary approach comprises algorithms based on deep convolutional neural networks (DCNNs) (15) that successfully detect other findings on CT, such as myocardial infarction (16), intracranial hemorrhage (17), and acute abdominal findings (18). Given the efficiency of DCNNs in detection of findings on CT and other modalities (19-21), we hypothesized that they are also suited to detect rib fractures. While there are multiple studies on deep learning (DL)-based detection of fractures on plain radiographs (22-27), the number of algorithms detecting fractures on CT is limited. Studies on algorithms detecting vertebral body (28, 29) and skull fractures have been performed (30), but only one preliminary study dealt with the detection of rib fractures (31).

Therefore, the aim of this study was a comprehensive assessment of the diagnostic performance of a DL-based algorithm for automatic detection of rib fractures on trauma CT scans acquired within 1 year at a level-1 trauma center.

2. Materials and Methods

The local ethics committee approved the study protocol and waived the requirement of obtaining informed consent (Project ID: 2019-00510).

2.1. Case Selection

We retrospectively identified all whole-body trauma CT scans and the corresponding written reports acquired at our department in 2018 with an in-house developed radiology information system/picture archiving and communication system (PACS) search engine (n = 511). Selection criteria were the procedure code and time period (January to December 2018). Examinations were classified into positive (only acute, only chronic, or both acute and chronic) and negative for rib fractures according to the written CT reports (Fig. 1). Two radiology residents (1st and 3rd year of residency) performed this classification task.

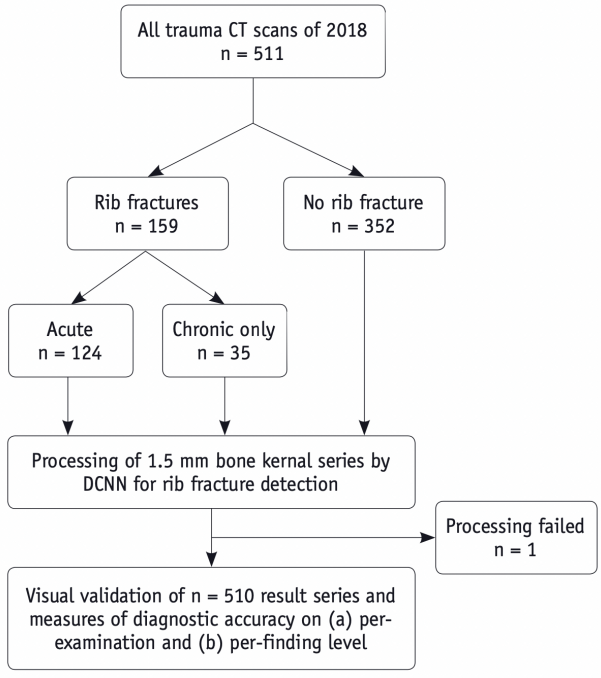


Fig. 1. Study flowchart. DCNN = deep convolutional neural network

2.2. Image Acquisition

Scans were acquired using three different CT scanners: Somatom Definition AS+ (n = 499; 128-slice system), Somatom Definition Edge (n = 5; 128-slice system), and Somatom Definition FLASH (n = 7; 2 x 128-slice system; all scanners: Siemens Healthineers, Erlangen, Germany). Scanning was performed following our standard protocol for whole-body trauma CT: patients were placed in the supine position with the scan ranging from the skull vertex to the upper thighs. Iopromide (Ultravist 370, Bayer Pharmaceuticals, Berlin, Germany) at a standard injection rate of 3.0 mL/s and a body weight-adapted volume of up to 120 mL was used as contrast agent. The peak kilovoltage was 120 kVp, and an automatic tube current modulation was performed. Transversal images in bone reconstruction kernel (70f) with a slice thickness of 1.5 mm served as the only input for the algorithm.

2.3. Index Test and Standard of Reference

The algorithm's output series with marked areas of suspected rib fractures was defined as the index test. The written CT reports established the standard of reference. These CT reports had been previously approved by a board-certified radiologist with at least 5 years of experience in emergency radiology at a level-1 trauma center. To determine the accuracy of the reports on acute fractures, we randomly selected a subset of 50 CT scans and performed a second reading without time constraints and without knowledge of the reports.

2.4. Algorithm Characteristics

The prototype algorithm used for rib fracture detection was trained to detect acute and chronic fractures and consisted of two stages: first, a region proposal stage, which was a three-dimensional convolutional deep neural network. Its architecture was based on ResNet (Aidoc Medical, Tel Aviv, Israel). ResNets enable the training of neural networks with many layers (32). The algorithm prototype had been trained using 11455 independent chest CT scans from eight other medical institutions, acquired on 15 different scanner models. These had been reviewed by two radiologists, one making annotations and another confirming those annotations. Hyperparameter optimization included approximately 30 experiments, performed with the parallelized stochastic gradient descent using the Horovod framework (33). Experiments were processed on different servers with one, two, four, and eight NVIDIA GPUs (NVIDA, Santa Clara, CA, USA). The DCNN provided suggestions for suspected rib fractures. Subsequently, a second stage

based on a Fast Region-based CNN disqualified some of the initial suggestions to reduce false-positives and selected locations for arrows indicating final findings. The output series is the original transversal series with overlaid arrows pointing at suspected rib fractures. On the internal test dataset, the performance was as follows: sensitivity 91.2% and specificity 90.7% on the per-examination level; and sensitivity 78.0% on the per-finding level.

2.5. Data Processing and Image Analysis

We performed full study data anonymization. The 1.5-mm transversal images in bone kernel (70f) were transferred to the detection algorithm. Processing the data comprised an automated cutting of the whole-body CT to the areas that displayed the ribs based on lung segmentation. These cropped series served as the only input for the core algorithm. The output series was reviewed on a validation platform using a conventional PACS monitor. A radiology resident initially reviewed all cases. Another resident and a board-certified radiologist discussed to reach a consensus on findings that were inconclusive to the first reader (e.g., does the finding display a true rib fracture or an artifact?) and all rib fractures that had been detected by the algorithm but had not been described in the written CT reports. Table 1 shows the detailed evaluation scheme for suspected findings.

Table 1. Evaluation Scheme

Feature	Subfeature (If Any)	Characterization
Location	Side	Left/right
	Section	Anterior/lateral/posterior
	Number of rib	1–12
Acuteness	-	Acute/chronic
Degree of displacement	-	No displacement (= nondisplaced acute fractures + chronic fractures)/ half-shaft/full-shaft/multifragmentary)
Mentioning in written report		Yes/no

Acute fractures were defined as fractures without any sign of healing, such as callus formation or complete or partial consolidation of the fracture gap. A non-displaced, acute fracture was defined as a fracture with cortical disruption but maintained alignment (34). Rib fractures missed by the algorithm were marked with a bounding box.

2.6. Statistical Testing

Statistical analyses were performed with SPSS Statistics, version 22 (IBM Corp., Armonk, NY, USA) and Microsoft Excel 2010 (Microsoft Corp., Redmond, WA, USA). P values

less than 0.05 were considered statistically significant. We performed descriptive statistics to describe patients' age and sex. To assess if there were statistically significant differences between patients with and without fractures, we performed the Chi-squared test for sex and the Mann-Whitney U test for age. On a per-examination level, an examination was defined as true positive when the algorithm correctly identified at least one fracture in a case with rib fractures according to the report. If the algorithm did not detect any fracture in an examination with at least one fracture according to the report, this examination was classified as false negative. Cases that were classified negative for rib fractures by both the report and the algorithm were rated as true negative. We calculated the sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), accuracy, and F1 score for the whole dataset and subsets.

On a per-finding level ("per-fracture"), we calculated the number of false-positives per examination. To analyze the correlation of location, displacement, and acuteness with the detection rate, a binomial logistic regression with location, acuteness and degree of displacement as independent variables and detection (yes/no) as the dependent variable was performed. In this model, the following categories were used to obtain dichotomy: left/right; acute/chronic; nondisplaced/displaced (nondisplaced vs. all other categories). Exp(B) is the exponentiation of the B coefficient, which is interpreted as the odds ratio within the model (35). To further investigate the association of detection with the localization of a fracture within a rib (anterior, lateral, or posterior) and level of the fractured rib (upper = rib 1–4; middle = rib 5–8; and lower = rib 9–12), we performed Chi-squared statistics.

3. Results

3.1. Examination Characteristics

The mean age of the patients was 58.4 ± 22.5 years. Patients with and without rib fractures did not statistically significantly differ in sex ratio ($\chi 2=0.23$; $p=0.63$) or age ($U=19830$; $p=0.44$). The rate of positive examinations for rib fractures (acute and/or chronic) according to the report was 31.2%. On a per-finding level, 894 rib fractures were described in the reports. Table 2 summarizes the characteristics of the fractures. A second reading was performed for the 49 findings that had been marked as inconclusive by the first reader. Our analysis of a subset of 50 randomly selected CT scans showed that 83.3% (ten of 12) of scans showing acute rib fractures were correctly described in the corresponding

reports. Of 511 trauma CT scans that were performed at our department in 2018, one scan that was negative for rib fractures according to the written CT report could not be processed because of failure in automated cropping.

Table 2. Characteristics of True-Positive Fractures That Were either Described in Written CT Reports (n = 894) or Additionally Detected by the Algorithm (n = 97)

Feature	Rib Fractures Described in Written CT Reports (n = 894)	Detection Rates for Subcategories in %	Rib Fractures Additionally Detected by Algorithm (n = 97)
Side			
Left	401	66.8 (268/401)	39
Right	493	64.7 (319/493)	58
Section			
Posterior	295	73.9 (218/295)	29
Lateral	383	68.7 (263/383)	41
Anterior	216	49.1 (106/216)	27
Height			
1–4	281	56.2 (158/281)	19
5–8	432	69.4 (300/432)	58
9–12	181	71.3 (129/181)	20
Acuteness			
Acute	688	67.7 (466/688)	65
Chronic	206	58.7 (121/206)	32
Degree of displacement			
No displacement	670	58.4 (391/670)	83
Half-shaft	118	88.1 (104/118)	10
Full-shaft	62	90.3 (56/62)	1
Multifragmentary	44	81.8 (36/44)	3

3.2. Per-Examination Level

On a per-examination level, the algorithm produced 139 true-positives, 30 false-positives, 321 true-negatives, and 20 false-negatives. This corresponded to a sensitivity of 87.4% (139 of 159 scans with rib fractures according to the report detected) and specificity of 91.5% (321 of 351 scans without rib fractures correctly classified as negative) on a per-examination level. Table 3 provides more details on the performance measures. Figure 2 shows a typical example of an acute fracture of the 9th right rib that was correctly identified by the algorithm and marked with an orange arrowhead. Our sub-analysis showed that the detection sensitivity of scans that contained acute fractures was significantly higher (91.9%) compared to that of scans that contained only chronic fractures (71.4%).

Table 3. Algorithm Performance on Per-Examination Level with 95% Confidence Intervals in Brackets

Sensitivity	Specificity	PPV	NPV	Accuracy	F1 Score
87.4% (81.2–92.1)	91.5% (88.0–94.2)	82.3% (76.6–86.8)	94.1% (91.4–96.0)	90.2% (87.3–92.6)	0.85

NPV = negative predictive value, PPV = positive predictive value

3.3. Per-Finding Level

On a per-finding level, there were 587 true-positives (sensitivity: 65.7%; 95% confidence interval: 68.8-92.4) and 307 false-negatives. Furthermore, 97 true rib fractures (65 acute and 32 chronic) were detected by the algorithm and confirmed by consensus reading, but not mentioned in the written CT reports. The binary logistic regression model set up to ascertain the effects of laterality, displacement, and acuteness of fractures on the likelihood of detection was statistically significant ($\chi 2=69.2$; $p<0.001$). While laterality had no impact on detection rates within the model, displaced rib fractures were 4.84 times (Exp(B)) more likely to be detected compared to nondisplaced fractures ($p<0.001$). Acute fractures were 4.60 times (Exp(B)) more likely to be detected compared

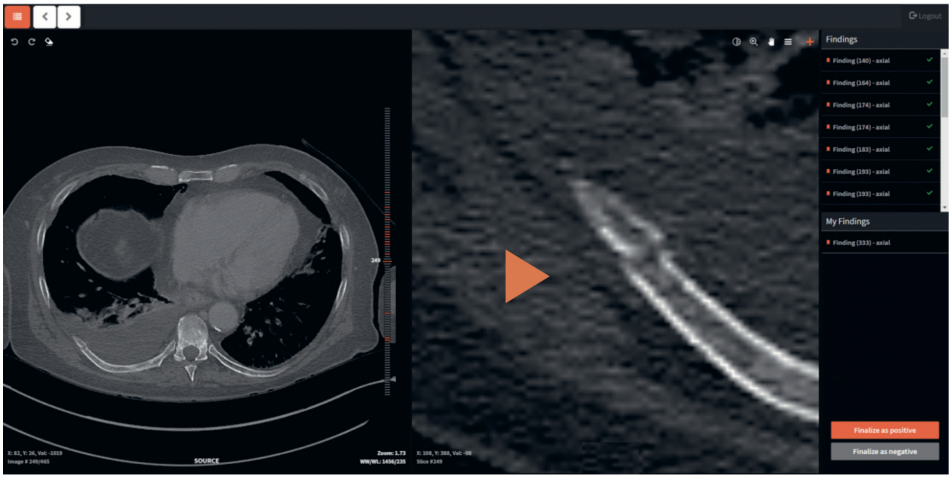


Fig. 2. Validation platform used for algorithm assessment.
Original 1.5-mm transversal series of trauma CT used as input on left side. Key-image of output series with enlarged finding suspected of representing acute rib fracture marked with orange arrowhead on right side.

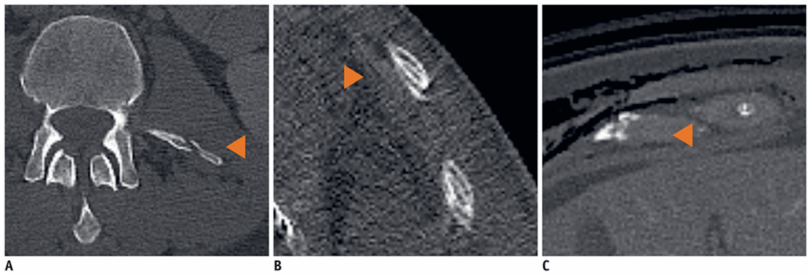
to chronic fractures ($p<0.001$). Furthermore, Chi-squared tests revealed significant associations between the fracture location and detection ($\chi 2=36.8$, $p<0.001$ for the position within a rib [anterior, lateral, posterior]; $\chi 2=16.4$, $p<0.001$ for the level of the rib [upper, middle, or lower]), with anteriorly and superiorly located fractures more likely to be missed. Table 2 provides detailed information on detection rates for all subcategories. The 81 false-positives translated to 0.16 false-positives per examination. Table 4 summarizes the number and reasons for false-positives. Additionally, we found 137 "double annotations," fracture was marked multiple times by the algorithm. Figure 3 displays

examples of false-positives. To further illustrate the clinical relevance of a rib fracture detection tool, Figure 4 shows an example of multiple, traumatic fractures in a 46-year-old woman after a car accident that required surgical stabilization. A fully anonymized basic study dataset containing information on the characteristics of the individual fractures and whether they were detected by the algorithm and mentioned in the radiology report can be found in Supplementary Table 1.

Table 4. Morphologic Correlates of False-Positives

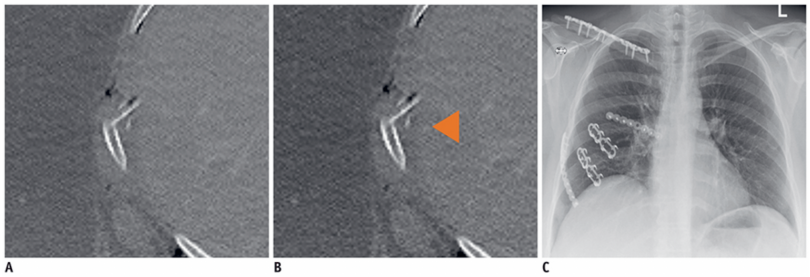
Anatomical Correlate	Number of Findings
Normal rib	18
Intercostal vessel	15
Breathing artifact	13
Out of bounds	11
Transition zone rib–costal cartilage	10
Fractures of other bones	10
Contrast agent artifact	3
Bone marrow calcification	1

Fractures of other bones = scapula, finger, processus transversus, Normal rib = intact rib misclassified as fracture, Out of bounds = fracture-mark with no anatomical correlation



A B C

Fig. 3. Three typical cases of false-positives marked with orange arrowheads, due to (A) fracture of transverse process, (B) breathing artifacts, and (C) physiological transition zone between rib and costal cartilage.



A B C

Fig. 4. Trauma CT scan of 46-year-old woman with multiple acute fractures of adjacent ribs after car accident (A) shows displaced, acute rib fracture (sixth lateral rib) that was (B) detected (orange arrowhead) and (C) subsequently required surgical stabilization.

4. Discussion

Our study assessed the diagnostic performance of a DL-based prototype algorithm for the automated detection of rib fractures on trauma CT. On a per-examination level, the algorithm reached a sensitivity of 87.4% and specificity of 91.5%. This is comparable to the accuracy of practicing radiologists (14). The F1 score was 0.85, and there were 0.16 false-positives per examination, mostly because of detections of intact ribs, normal intercostal vessels, and breathing artifacts. On a per-finding level, 587 of 894 fractures mentioned in the reports were detected (sensitivity: 65.7%). Main factors associated with the correct detection of fractures by the algorithm were displacement and acuteness. The superior performance of the algorithm on a per examination level compared to the per-finding level may be explained by the fact that in an emergency setting, multiple rib fractures are more frequent than isolated rib fractures. In our dataset, only 9.4% of scans with rib fractures contained only one rib fracture. The detection of only one of multiple fractures is sufficient to identify a positive case correctly on the per-examination level. Due to a high NPV of 94.1%, the algorithm prototype preserves its usability as a secondary reading tool on a per-examination level. Only one preliminary study evaluated an algorithm for the detection of rib fractures on CTs: Yan et al. (31) deployed a CNN, yielding a sensitivity of 95.0% and a significantly lower PPV of 55.7% for the detection of rib fractures, tested on a set of 244 fractures. The measure of false-positives and false-negatives per case was not reported. Our results are comparable to those of other researchers investigating the performance of algorithms for the detection of fractures of other bones on CT. In their study on the performance of a support vector machine for detection of vertebral body fractures, Burns et al. (28) found a sensitivity of 81.3% on a per-finding level and 2.7 false-positives per case. The number of false-positives per case that we found was much smaller (0.16), which translates to a better usability in the clinical workflows. Bar et al. (29) assessed an algorithm for detection of vertebral compression fractures based on a segmentation step and a patch-based CNN and reported a sensitivity of 83.9% and specificity 93.8%. While these results are similar to ours, the authors did not include information on false-positives. However, comparability is limited because each type of fracture has different characteristics and surrounding anatomical structures. We additionally found 137 rib fractures that were annotated multiple times by the algorithm. The consequences of this depend on the way the algorithm is used: if the algorithm is used to flag scans with fractures on the per-examination level, there is no consequence; if the algorithm is used to determine the exact number of fractures, and results are checked by a radiologist in a

second step, the workflow efficiency is reduced; and if the detailed results are adopted without checking, this results in a wrong assessment of the number of rib fractures. Altogether, the algorithm detected 97 acute fractures not mentioned in the written CT report. This is of interest to clinicians since Battle et al. (36) have demonstrated that a higher number of rib fractures is associated with an increased mortality, underpinning the importance of correct rib fracture detection. Moreover, we found that detection rates for fractures located anteriorly were lower than those for other locations. Interestingly, this is in line with the results of Ringl et al. (14) and might have resulted from the diagnostically challenging zone of transition from the rib to the cartilage.

Our study has several limitations. First, due to the retrospective design and limited availability of fracture-specific clinical data, the results could be linked neither to clinical symptoms nor to clinical outcomes. Second, the analysis was performed on data acquired on scanners of one vendor and at one center only. Therefore, the performance might differ across institutions and scanners. However, CT trauma protocols are highly standardized, as they are optimized to target specific clinical questions. Therefore, we do not expect a relevant bias. Third, the algorithm output was assessed by one radiologist only with a consensus reading of two radiologists in inconclusive cases. However, the fact that the assessment of patients with trauma is a task that residents learn early in their professional career supports our conclusion that it only slightly affects the validity of the study. Fourth, the reference standard was defined by the clinically approved written CT reports. We chose this definition because we consider these reports to be a valid basis for a reference standard. A complete reading of all 511 cases by multiple readers was not possible because of the substantial time required for this task.

Due to the continuum of bone healing and the resulting indefinite cut-off between acute and chronic fractures, we decided to include both fracture types. Since the algorithm provides good results on a per case level and has a high NPV, the algorithm is usable as a screening tool to flag scans with at least one suspected rib fracture. If no rib fractures were detected by the radiologist in charge during the first reading of a trauma CT scan flagged as suspicious by the algorithm, a quick second check for rib fractures might be appropriate to avoid false-negatives.

In conclusion, the algorithm we evaluated on a large dataset independent from its training data showed good diagnostic performance for the automated detection of rib fractures on

whole-body trauma CT on a per-examination level. Thus, despite lower sensitivity on a per-finding level, it constitutes a foundation for a clinical decision support tool for reading assistance.

Supplementary Materials

The Data Supplement is available at https://doi.org/10.3348/kjr.2019.0653.

Conflicts of Interest

The authors have no potential conflicts of interest to disclose.

Acknowledgments

We acknowledge the provision of the rib fracture detection algorithm prototype by Aidoc Medical (Tel Aviv, Israel). The company was neither involved in data analysis nor study design. No financial support was granted.

References

1. M, Türüt H, Topçu S, Gülhan E, Yazici U, Kaya S, et al. A comprehensive analysis of traumatic rib fractures: morbidity, mortality and management. Eur J Cardiothorac Surg. 2003;24:133–138.
2. Sokolovskaya E, Shinde T, Ruchman RB, Kwak AJ, Lu S, Shariff YK, et al. The effect of faster reporting speed for imaging studies on the number of misses and interpretation errors: a pilot study. J Am Coll Radiol. 2015;12:683–688.
3. Park SH, Song HH, Han JH, Park JM, Lee EJ, Park SM, et al. Effect of noise on the detection of rib fractures by residents. Invest Radiol. 1994;29:54–58.
4. Berbaum KS, Franken EA, Dorfman DD, Rooholamini SA, Coffman CE, Cornell SH, et al. Time course of satisfaction of search. Invest Radiol. 1991;26:640–648.
5. Banaste N, Caurier B, Bratan F, Bergerot JF, Thomson V, Millet I. Whole-body CT in patients with multiple traumas: factors leading to missed injury. Radiology. 2018;289:374–383.
6. Cho SH, Sung YM, Kim MS. Missed rib fractures on evaluation of initial chest CT for trauma patients: pattern analysis and diagnostic value of coronal multiplanar reconstruction images with multidetector row CT. Br J Radiol. 2012;85:e845–e850.
7. Mayberry JC, Schipper PH. Traumatic rib fracture: conservative therapy or surgical fixation? In: Ferguson M, editor. Difficult decisions in thoracic surgery. London: Springer; 2011. pp. 489–493.
8. Lu MS, Huang YK, Liu YH, Liu HP, Kao CL. Delayed pneumothorax complicating minor rib fracture after chest trauma. Am J Emerg Med. 2008;26:551–554.
9. Ho SW, Teng YH, Yang SF, Yeh HW, Wang YH, Chou MC, et al. Risk of pneumonia in patients with isolated minor rib fractures: a nationwide cohort study. BMJ Open. 2017;7:e013029
10. Tanaka H, Yukioka T, Yamaguti Y, Shimizu S, Goto H, Matsuda H, et al. Surgical stabilization of internal pneumatic stabilization? A prospective randomized study of management of severe flail chest patients. J Trauma. 2002;52:727–732. 732.
11. Bemelman M, de Kruijf MW, van Baal M, Leenen L. Rib fractures: to fix or not to fix? An evidence-based algorithm. Korean J Thorac Cardiovasc Surg. 2017;50:229–234.
12. de Jong MB, Kokke MC, Hietbrink F, Leenen LPH. Surgical management of rib fractures: strategies and literature review. Scand J Surg. 2014;103:120–125.
13. Murphy CE, Raja AS, Baumann BM, Medak AJ, Langdorf MI, Nishijima DK, et al. Rib fracture diagnosis in the Panscan era. Ann Emerg Med. 2017;70:904–909.
14. Ringl H, Lazar M, Töpker M, Woitek R, Prosch H, Asenbaum U, et al. The ribs unfolded-a CT visualization algorithm for fast detection of rib fractures: effect on sensitivity and specificity in trauma patients. Eur Radiol. 2015;25:1865–1874.
15. Lee JG, Jun S, Cho YW, Lee H, Kim GB, Seo JB, et al. Deep learning in medical imaging: general overview. Korean J Radiol. 2017;18:570–584.
16. Mannil M, von Spiczak J, Manka R, Alkadhi H. Texture analysis and machine learning for detecting myocardial infarction in noncontrast low-dose computed tomography: unveiling the invisible. Invest Radiol. 2018;53:338–343.
17. Prevedello LM, Erdal BS, Ryu JL, Little KJ, Demirer M, Qian S, et al. Automated critical test findings identification and online notification system using artificial intelligence in imaging. Radiology. 2017;285:923–931.
18. Winkel DJ, Heye T, Weikert TJ, Boll DT, Stieltjes B. Evaluation of an AI-based detection software for acute findings in abdominal computed tomography scans: toward an automated work list prioritization of routine CT examinations. Invest Radiol. 2019;54:55–59.
19. Alkadi R, Taher F, El-baz A, Werghi N. A deep learning-based approach for the detection and localization of prostate cancer in T2 magnetic resonance images. J Digit Imaging. 2019;32:793–807.
20. Kooi T, Litjens G, van Ginneken B, Gubern-Mérida A, Sánchez CI, Mann R, et al. Large scale deep learning for computer aided detection of mammographic lesions. Med Image Anal. 2017;35:303–312.
21. Cicero M, Bilbily A, Colak E, Dowdell T, Gray B, Perampaladas K, et al. Training and validating a deep convolutional neural network for computer-aided detection and classification of abnormalities on frontal chest radiographs. Invest Radiol. 2017;52:281–287.
22. Yahalomi E, Chernofsky M, Werman M. Detection of distal radius fractures trained by a small set of X-ray images and faster R-CNN. In: Arai K, Bhatia R, Kapoor S, editors. Intelligent computing. Cham: Springer; 2019. pp. 971–981.
23. Thian YL, Li Y, Jagmohan P, Sia D, Chan VEY, Tan RT. Convolutional neural networks for automated fracture detection and localization on wrist radiographs. Radiol Artif Intell. 2019;1:e180001
24. Kim DH, MacKinnon T. Artificial intelligence in fracture detection: transfer learning from deep convolutional neural networks. Clin Radiol. 2018;73:439–445.

25. Starosolski ZA, Kan H, Annapragada AV. CNN-based radiographic acute tibial fracture detection in the setting of open growth plates. [Accessed August 25, 2019];bioRxiv. 2019 doi: 10.1101/506154. Available at: DOI
26. Kitamura G, Chung CY, Moore BE. Ankle fracture detection utilizing a convolutional neural network ensemble implemented with a small sample, de novo training, and multiview incorporation. J Digit Imaging. 2019;32:672–677.
27. Lindsey R, Daluiski A, Chopra S, Lachapelle A, Mozer M, Sicular S, et al. Deep neural network improves fracture detection by clinicians. Proc Natl Acad Sci U S A. 2018;115:11591–11596.
28. Burns JE, Yao J, Muñoz H, Summers RM. Automated detection, localization, and classification of traumatic vertebral body fractures in the thoracic and lumbar spine at CT. Radiology. 2016;278:64–73.
29. Bar A, Wolf L, Amitai OB, Toledano E, Elnekave E. Compression fractures detection on CT. Medical Imaging 2017: Computer-Aided Diagnosi. 2017;10134:101344
30. Chilamkurthy S, Ghosh R, Tanamala S, Biviji M, Campeau NG, Venugopal VK, et al. Development and validation of deep learning algorithms for detection of critical findings in head CT scans. Cornell University; 2018. [updated April 2018]. [Accessed August 25, 2019]. Available at: https://arxiv.org/abs/1803.05854.
31. Yan L, Chuan X, Xia C, Wang S, Chen K. Deep learning for automatic detection of fractures on chest CT scans after blunt trauma (number: B-0566); ECR 2019 (European Congress of Radiology); 2019 February 27-March 3; Vienna, Austria.
32. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Cornell University; 2015. [Accessed August 25, 2019]. Available at: https://arxiv.org/abs/1512.03385.
33. Sergeev A, Del Balso M. Horovod: fast and easy distributed deep learning in TensorFlow. Cornell University; 2018. [updated February 2018]. [Accessed August 25, 2019]. Available at: https://arxiv.org/abs/1802.05799.
34. Talbot BS, Gange CP, Chaturvedi A, Klionsky N, Hobbs SK, Chaturvedi A. Traumatic rib injury: patterns, imaging pitfalls, complications, and treatment. Radiographics. 2017;37:628–651.
35. Park HA. An introduction to logistic regression: from basic concepts to interpretation with particular attention to nursing domain. J Korean Acad Nurs. 2013;43:154–164.
36. Battle CE, Hutchings H, Evans PA. Risk factors that predict mortality in patients with blunt chest wall trauma: a systematic review and meta-analysis. Injury. 2012;43:8–17.

CHAPTER 5

Evaluation of an AI-Powered Lung Nodule Algorithm for Detection and 3D Segmentation of Primary Lung Tumors

BASED ON: Weikert T, Akinci D'Antonoli T, Bremerich J, Stieltjes B, Sommer G, Sauter AW. Evaluation of an AI-Powered Lung Nodule Algorithm for Detection and 3D Segmentation of Primary Lung Tumors. Contrast Media Mol Imaging. 2019 Jul 1;2019:1545747. doi: 10.1155/2019/1545747.

Abstract

Purpose: Automated detection and segmentation are prerequisites for the deployment of image-based secondary analyses, especially for lung tumors. However, currently only applications for lung nodules ≤3 cm exist. Therefore, we tested the performance of a fully automated AI-based lung nodule algorithm for detection and 3D segmentation of primary lung tumors in the context of tumor staging using the CT component of FDG-PET/CT and including all T-categories (T1–T4).

Materials and Methods: FDG-PET/CTs of 320 patients with histologically confirmed lung cancer performed between 01/2010 and 06/2016 were selected. First, the main primary lung tumor within each scan was manually segmented using the CT component of the PET/CTs as reference. Second, the CT series were transferred to a platform with AI-based algorithms trained on chest CTs for detection and segmentation of lung nodules. Detection and segmentation performance were analyzed. Factors influencing detection rates were explored with binominal logistic regression and radiomic analysis. We also processed 94 PET/CTs negative for pulmonary nodules to investigate frequency and reasons of false-positive findings.

Results: The ratio of detected tumors was best in the T1-category (90.4%) and decreased continuously: T2 (70.8%), T3 (29.4%), and T4 (8.8%). Tumor contact with the pleura was a strong predictor of misdetection. Segmentation performance was excellent for T1 tumors ($r = 0.908$, $p<0.001$) and tumors without pleural contact ($r = 0.971$, $p<0.001$). Volumes of larger tumors were systematically underestimated. There were 0.41 false-positive findings per exam.

Conclusion: The algorithm tested facilitates a reliable detection and 3D segmentation of T1/T2 lung tumors on FDG-PET/CTs. The detection and segmentation of more advanced lung tumors is currently imprecise due to the conception of the algorithm for lung nodules <3 cm. Future efforts should therefore focus on this collective to facilitate segmentation of all tumor types and sizes to bridge the gap between CAD applications for screening and staging of lung cancer.

1. Introduction

Failure to detect lung cancer on imaging studies is a very common reason for malpractice suits [1]. The reasons for misdiagnosis are multilayered and include recognition error and satisfaction of search [2]. Strategies for the reduction of observer errors are therefore of great importance and computer-aided detection (CAD) of pulmonary nodules has gained increasing interest in this context [3]. Most recently, conventional CAD solutions that require visual confirmation to reduce false-positive calls [4] are being challenged by deep learning algorithms that have an inherent advantage of automatic feature exploitation [3].

The diagnostic task of imaging in lung cancer, however, does not end with tumor detection. Tumor staging using 18F-fluorodeoxyglucose(FDG-) PET/CT as the standard of care forms an integral part of the clinical diagnostic workup of patients with lung cancer [5]. The recent revision on the T-categories for the 8th edition of the TNM lung cancer classification emphasized that from 1 to 5 cm, each cm separates lesions of significantly different prognosis [6]. However, the implicit assumption that tumors are spherical and consequently proportional changes of tumor diameter and parallel changes in tumor volume is particularly disrupted for advanced tumors [7]. This clearly underlines the need for accurate tumor segmentation and precise tumor volumetry, particularly when it comes to therapy response monitoring [7], radiation treatment planning [8], radiomics [9], and other new developments in the framework of personalized medicine.

Sexauer et al. have shown that manual annotation and segmentation of lung tumors is feasible, but tumor stage and lesion size and count correlate significantly with segmentation time [10]. Algorithms for automatic pulmonary nodule detection and segmentation are currently under development but are commonly trained and validated based on intraparenchymal lesions which are less than 3 cm in size. Therefore, it is unclear how pulmonary masses beyond this diameter and with nonspherical shape will be treated by these algorithms. Moreover, the vast majority of CAD systems have been evaluated on chest CTs that have been acquired in deep-inspiration breath-hold technique [11–21]. So far, only few CAD applications were tested for PET/CT and that only for nodules smaller than 3 cm [22,23].

It was thus the aim of this study to evaluate the performance of a fully automated computer-assisted detection and 3D segmentation algorithm that was initially designed for lung nodule detection and segmentation in the context of tumor staging. This was done using the CT component of FDG-PET/CT studies of a patient cohort with histologically proven primary lung tumors from all T-categories.

2. Materials and Methods

This study was conducted under the provisions of the appropriate Swiss regional ethics committee (*Ethikkommission Nordwest-und Zentralschweiz*).

2.1. Case Selection

We compiled two datasets using an in-house-developed Radiology Information System/Picture Archiving and Communication System (RIS/PACS) search engine: First, we retrospectively identified 18F-fluorodeoxyglucose (FDG-) PET/CTs with histologically proven primary lung cancer that were acquired at our institution between 01/2010 and 06/2016. Selection criteria were protocol name, time period, and verified tumor histology according to our pathology archive. This resulted in 320 PET/CTs (lung tumor population). Second, for the creation of a dataset with exams not containing pulmonary nodules, appropriate PET/CTs were selected with the criteria protocol name, time period (01/2017–12/2018), and the presence of the text string “no pulmonary nodules” in the clinically approved reports. This resulted in 92 PET/CTs (nodule negative population). The study workflow is displayed in Figure 1.

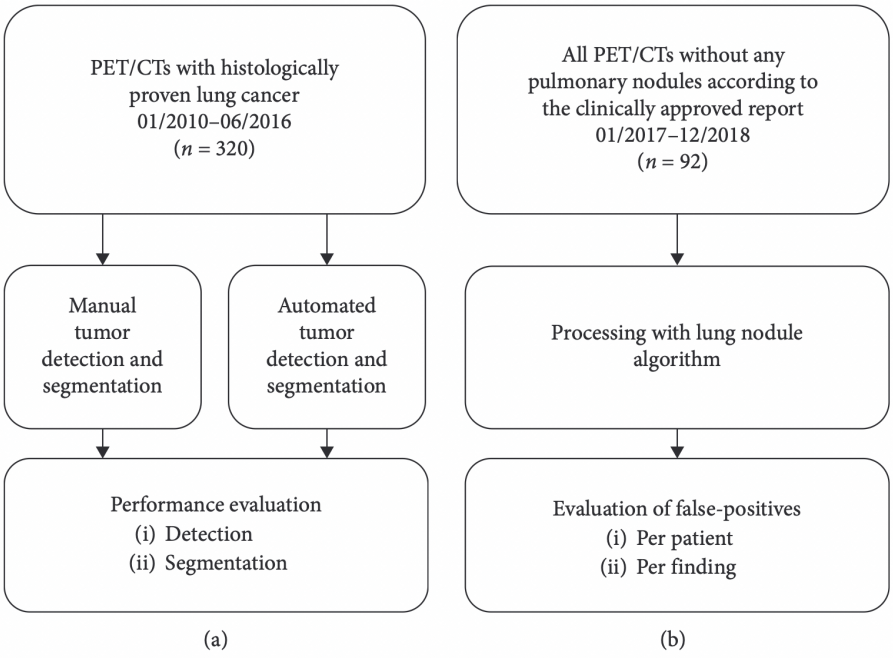


Figure 1: Study workflow for (a) lung tumor population and (b) nodule negative population.

2.2. Imaging Protocols

PET/CT examinations were performed on two integrated PET/CT systems: on a Discovery STE with 16-slice CT (GE Healthcare, Chalfont St Giles, UK) from 01/2008 to 11/2015 and on a Biograph mCT-X RT Pro Edition with 128-slice CT (Siemens Healthineers, Erlangen, Germany) from 12/2015 to 12/2016. Scans were obtained 1 hour after intravenous injection of 5 MBq FDG/kg body weight at glycemic levels below 10 mmol/L and previous fasting for at least 6 h. The CT component of the combined PET/CT examination was acquired with the following parameters: Discovery STE: slice thickness 3 mm, i50f kernel, X-ray tube voltage 120 kVp (SD: 0 kVp), exposure 80 mAs (SD: 15 mAs), CTDIvol 5.8 mGy (SD: 1.7 mGy), and DLP 536 mGy cm (SD: 100 mGy cm). Biograph mCT-X: slice thickness 3 mm, i50f kernel, X-ray tube voltage 120 kVp (SD: 0 kVp), 37 mAs (SD: 18 mAs), CTDIvol 3.1 mGy (SD: 1.5 mGy), and DLP 294 mGy cm (SD: 146 mGy cm). In 21 cases, Iopromide (Ultravist 370, Bayer Pharma, Germany, Berlin) was applied as contrast agent at a mean dose of 87.1 ml (SD: 24.9 ml). All other scans were acquired without contrast.

2.3. Ground Truth Segmentation

Manual tumor segmentations with reference to the clinically approved report were performed as previously described [10]. The PET/CT image dataset of each patient was segmented via a modified 3D-slicer-based segmentation tool (version 4.6.2, Slicer Python Interactor 2.7.11, Boston, USA). Segmentation of the data involved in this analysis was performed by a dual-board-certified radiologist and nuclear medicine physician with 10 years of experience in PET/CT reading (A. S., $n = 137$) as well as a radiology resident with 2 years of professional experience that was supervised by A. S. (T. W., $n = 183$). Tumors were segmented as a 3D volume defined by consecutive 2D regions of interest (ROIs) that were delineated on all transversal slices of the CT component showing a lesion. Fused PET information was used in addition whenever the tumor boundaries were not clearly definable on CT.

2.4. Algorithm Characteristics

The transversal 3 mm low-dose CT series of the PET/CTs with histologically proven primary lung tumor ($n = 320$) as well as the CT series of the PET/CTs negative for pulmonary nodules ($n = 94$) served as the only input for the in-house-deployed AI-based research algorithm for detection and segmentation of lung nodules. The image data were processed in three steps: First, lung and lung lobe segmentation were performed by a deep image-to-image network (DI2IN) that was trained on chest CTs acquired on scanners of multiple vendors. Its architecture has previously been described for liver segmentation by Yang et al. [24]. Second, nodule detection was performed by nodule candidate generation (NCG) and false-positive reduction (FPR). The NCG is a 3D region proposal network based on faster-RCNN [25] that outputs suspicious regions called "nodule candidates" and assigns probability scores. Then, for each nodule candidate, a small patch around it was sampled and sent to the FPR module consisting of several Res-Net units [26]. The FPR module further evaluated the likelihood for the nodule candidate to be a true nodule or a false positive by updating the scores generated by the NCG module. The final decision was made by taking the weighted sum of the scores generated by NCG and FPR modules. The training data for the nodule detection algorithm contained nodules up to a diameter of 3 cm. Third, nodules were segmented by an algorithm based on region growing. The principle of this method has been previously described by Hojjatoleslami and colleagues [27]. In the interest of improved readability, these three interlinked algorithms will be referred to as "algorithm" in this paper. None of the selected PET/CTs within the study was used to train the algorithm or to adapt hyperparameters.

2.5. Data Analysis

The output of the AI algorithm pipeline was the transversal chest CT component of the PET/CT with overlays for lung lobe boundaries and tumor boundaries of detected tumors. This output series also contained specifications of volume ($Volume_{AI}$), 2D diameter, and location (lung lobe) for every detected tumor and served as the index test. The reference standard was the CT component of the PET/CT for detection and the volumes that were calculated from the 3D tumor masks that resulted from the manual image segmentation process (ground truth volumes: $Volume_{GT}$). For each case, the segmented tumor was visually correlated with the output series of the algorithm and it was recorded whether the tumor was detected or not. The correctness of the indication of tumor location (lung lobe) was checked. We additionally established whether a lesion contacted parietal pleura or not by consensus reading (A. S. and T. W.). Finally, we reviewed the output series of the nodule negative population to describe numbers of and reasons for false-positive findings.

2.6. Statistical Analysis and Radiomics

Statistical analysis was performed using IBM SPSS Statistics for Windows, Version 22.0 (IBM Corp., Armonk, NY). Scatterplots and graphs were created with JMP, Version 14.2 (SAS Institute Inc., Cary, NC). For descriptive analyses of continuous data, we calculated the mean and standard deviations. To test for association between two or more categorical variables, we used the chi-squared test. To test for statistical differences among the means of two or more groups, we conducted a one-way analysis of variance. Normal distribution was assessed with the Shapiro–Wilk test, histograms, and Q-Q plots. To analyze the influence of histology, location, pleural contact, and maximal axial diameter on detection rates, we performed a binomial logistic regression with detection (yes/no) as the dependent variable. In this model, the largest histology subgroup and the most common location regarding the lung lobe (for location) were set as reference categories of the categorical variables. For the analysis of segmentation performance, all tumors with automatically calculated tumor volumes (VolumeAI) were considered (=all tumors detected). We used the Pearson correlation coefficient to assess the relationship between VolumeGT and VolumeAI. Values less than 0.05 were defined to indicate statistical significance.

To elucidate the influence of textual features on detection rates, we extracted 200 radiomic features with Pyradiomics version 2.1.0 [28]. Least absolute shrinkage selection operator (LASSO) regression and extended Bayesian information criterion (EBIC) were used for

feature selection in Stata Statistical Software Release 15 (StataCorp, College Station, TX). Selected features were then transferred into a logistic regression model and the predictive power was assessed. Youden cutoff values were generated for each selected feature [29].

3. Results

3.1. Lung Tumor Population

3.1.1. Population Characteristics

The mean patient age was 66.7 years (SD: 10.7 years). 70.3% of the patients were male (n = 225), and 29.7% were female (n = 95). The mean tumor volume was 68.2 cm3 (SD: 125.6 cm3; T1 = 3.0 cm3, T2 = 17.8 cm3, T3 = 56.7 cm3, and T4 = 210.0 cm3), and the mean axial tumor diameter was 5.0 cm (SD: 3.4 cm). Tumors were located in all lobes (right upper lobe: n = 101; middle lobe: n = 19; right lower lobe: n = 50; left upper lobe: n = 88; left lower lobe: n = 62). All T-categories were represented in the dataset with the following distribution: T1: n = 83; T2: n = 106; T3: n = 51; T4: n = 80. There were no statistically significant differences between the patients included in the T-categories regarding age and gender ($\chi 2 = 1.217$, $p = 0.749$). The distribution of tumor histology is shown in Table 1.

Table 1: Distribution of the lung tumor histology subtypes.

Tumor histology	*n*	%
Adenocarcinoma (AC)	174	54.2
Squamous cell carcinoma (SCC)	79	24.6
NSCLC not specified (NOS)	25	7.8
SCLC	15	4.7
Other*	28	8.7

*Large cell carcinoma, neuroendocrine tumor (NET), sarcomatoid carcinoma, spindle cell carcinoma, typical carcinoid, and combined carcinomas (NET + SCLC; SCLC + SCC; NET + SCC; NET + AC).

3.1.2. Detection

The attribution of a lesion to the corresponding lung lobe was correct in 100% of the detected lesions. Detection rates differed significantly across T-categories and declined towards advanced tumors: 90.4% for T1 (75 of 83), 70.8% for T2 (75 of 106), 29.4% for T3 (15

of 51), and 8.8% for T4 (7 of 80). This detection decline is also reflected in Figure 2(a) that shows the number of detected and missed tumors by T-category and Figure 2(b) that displays detection of tumors depending on the ground truth volume. Furthermore, mean VolumeGT was smaller for detected lesions (18.6 cm3; SD: 39.3 cm3) as compared to missed lesions (125.9 cm3; SD: 161.8 cm3).

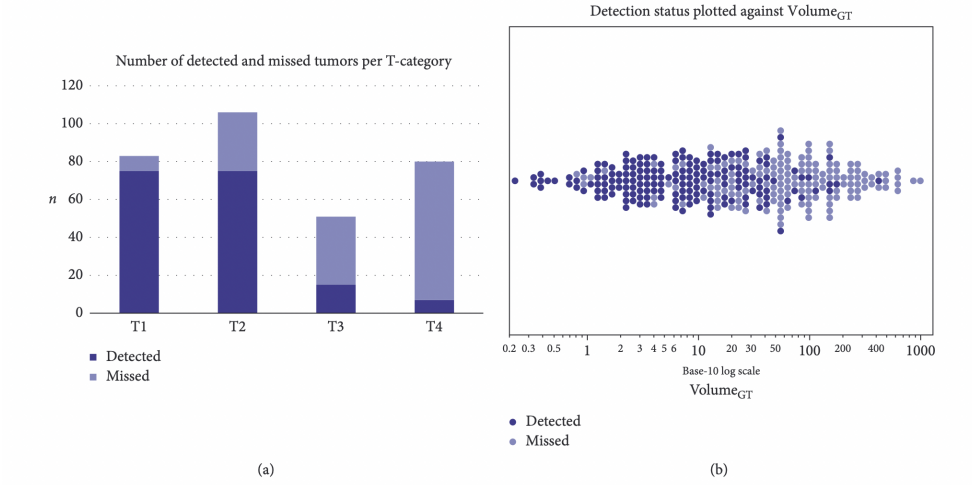


FIGURE 2: Tumors and their detection status. Tumors detected by the algorithm are visualized in dark blue and missed tumors in light blue. (a) Histogram per T-category. (b) Detection of tumors depending on the ground truth volumes. Every dot represents one tumor. *X*-axis with Volume$_{GT}$ in cm^3, in base-10 log scale.

Binominal logistic regression conducted to explore factors that influence detection rates showed that tumors with a larger maximal axial diameter and tumors with pleural contact were more likely to be missed by the detection algorithm (both $p < 0.001$). The results of this analysis are summarized in Table 2. Interestingly, squamous cell carcinomas and SCLC had a slightly higher likelihood to be missed compared to adenocarcinomas ($p < 0.001$ and $p = 0.015$, respectively). Location of a lesion in a specific lung lobe did not influence detection rates. With an Exp(B) of 74.4, pleural contact was by far the most relevant factor for nondetection in the model. This is also reflected by the fact that 94 of 95 lesions without pleural contact were detected (98.9%), while only 78 of 225 lesions with pleural contact were correctly identified (34.7%).

Table 2: Results of the binomial logistic regression.

Independent variables	p	Exp(B) with 95% CI
Histology subtype		
Reference: adenocarcinoma		
(1) Squamous cell carcinoma	**<0.001**	**0.209 (0.089–0.490)**
(2) NSCLC (NOS)	0.181	0.443 (0.134–1.461)
(3) SCLC	**0.015**	**0.093 (0.014–0.636)**
(4) Others	0.653	0.765 (0.237–2.464)
Location (lobes)		
Reference: right upper lobe		
(1) Middle lobe	0.350	0.499 (0.116–2.145)
(2) Right lower lobe	0.495	1.446 (0.502–4.167)
(3) Left upper lobe	0.905	1.054 (0.448–2.480)
(4) Left lower lobe	0.902	0.943 (0.369–2.408)
Pleural contact	**<0.001**	**74.400 (9.345–592.324)**
Maximal axial diameter	**<0.001**	**0.953 (0.938–0.969)**

Detection (yes/no) was set as dependent variable. Independent variables: histology (categorial), location (categorial), pleural contact (dichotomous), and maximal axial diameter (continuous). Exp(B) is the exponentiation of the B coefficient.

Table 3 summarizes the results of the radiomic analysis. It revealed that first order, shape, and texture features were significantly different in detected and missed tumors ($p < 0.001$). Tumors with finer, less heterogeneous texture (e.g., CT_glrlm_GrayLevelNonUniformityN: Lasso coefficient = −1.0776312, Youden cutoff = 0.1166608) and rounder shape (e.g., shape_Sphericity: Lasso coefficient = 0.2268932, Youden cutoff = 0.4293948) were more likely to be detected by the algorithm. Interestingly, three PET features (PET_firstorder_10Percentile, PET_firstorder_Maximum, PET_gldm_DependenceEntropy) indicated whether or not a tumor is detected on the CT component.

Table 3: Results of the radiomic analysis with features from Pyradiomics.

Selected feature	Lasso coefficient	Youden cutoff
CT_glrlm_GrayLevelNonUniformityN	−1.0776312	0.1166608
PET_firstorder_10Percentile	−0.0344698	1.7492108
PET_firstorder_Maximum	−0.0022762	6.9905767
PET_gldm_DependenceEntropy	0.0716689	2.2174546
shape_Maximum2DdiameterSlice	−0.0043233	32.866422
shape_Sphericity	0.2268932	0.4293948

3.1.3. Segmentation

All tumors detected by the algorithm were included in the second step of our analysis that investigated the segmentation performance (all: n = 172; T1: n = 75; T2: n = 75; T3: n = 15; T4: n = 7). We found a positive correlation between volumes calculated by the algorithm and ground truth volumes (Pearson correlation coefficient: $r = 0.634$, $p < 0.001$). As for detection rates, there were differences regarding T-categories: $r = 0.908$ for T1 ($p < 0.001$), $r = 0.797$ for T2 ($p < 0.001$), $r = 0.520$ for T3 ($p < 0.047$), and $r = 0.748$ for T4 ($p < 0.053$). This correlation is displayed in Figures 3(a)–3(d). It is worth mentioning that due to the low detection rate only seven T4 tumors were included and therefore the high Pearson correlation coefficient is likely related to random effects. Automatically calculated volumes of tumors that had no contact to pleura had a stronger correlation with ground truth volumes ($r = 0.971$, $p < 0.001$) as compared to tumors with pleural contact ($r = 0.586$, $p < 0.001$) for all T-categories. The volumes of larger tumors were systematically underestimated by the algorithm. Figure 4 displays a typical example of a T1 lesion without pleural contact that was manually segmented (a) as well as correctly segmented by the algorithm (b). Figure 4(c) shows an incompletely segmented T3 lesion with pleural attachment, and Figure 4(d) illustrates an invasive, completely missed T4 lesion.

3.2. Nodule Negative Population

Mean age of the patients was 63.2 years (SD: 16.6 years). There were 60.6% males (n = 57) and 39.4% females (n = 37). There were 39 false-positive findings (FP). This corresponds to 0.41 FP per patient. FPs were caused by dystelectases (n = 18), intrapulmonary vessels (n = 12), hilar calcified lymph nodes (n = 3), detection of ribs (n = 2), and a breathing artifact (n = 1).

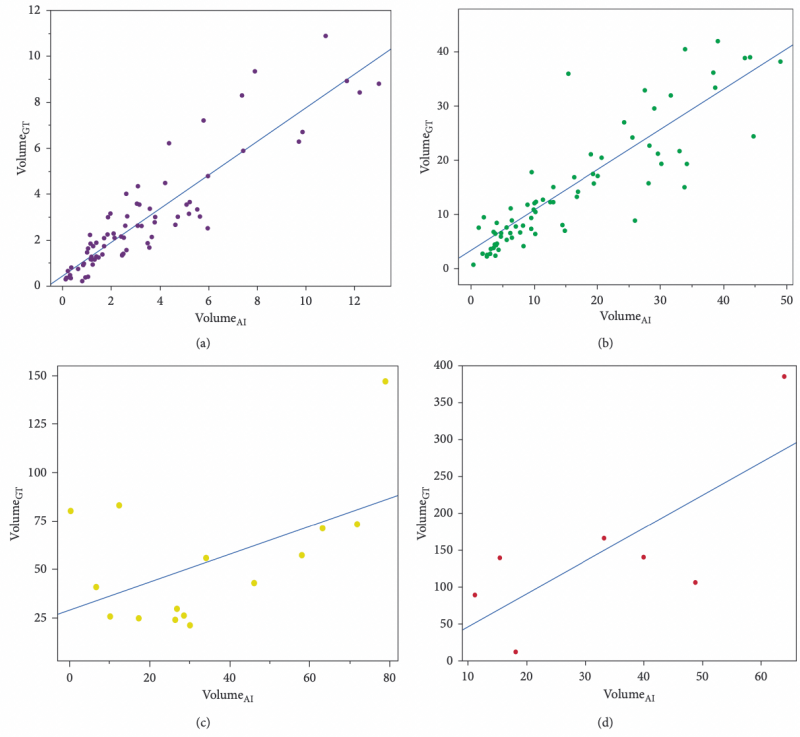


FIGURE 3: Segmented ground truth volumes (Volume$_{GT}$) in cm^3 (Y-axis) plotted against automatically calculated volumes (Volume$_{AI}$) in cm^3 (X axis) with linear regression line for (a) T1, (b) T2, (c) T3, and (d) T4.

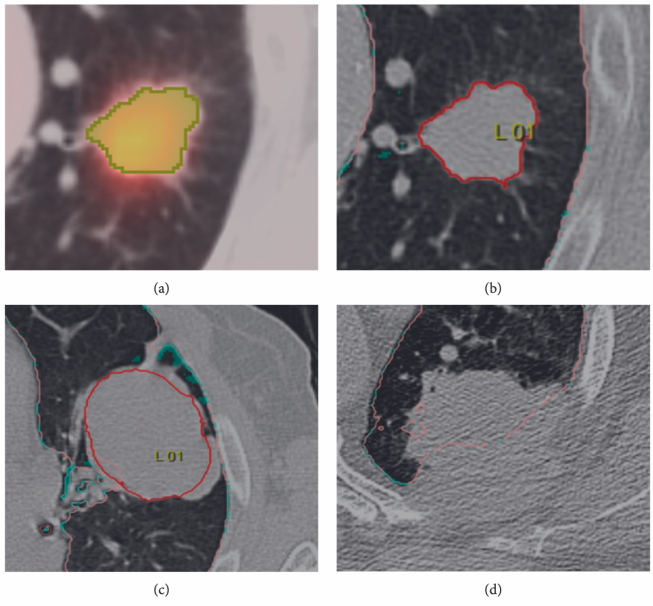


FIGURE 4: Examples for (a) manual segmentation of a T1 tumor without pleural contact with (b) corresponding excellent segmentation by the algorithm, (c) an incompletely segmented T3 lesion with pleural attachment, and (d) a completely missed T4 lesion with infiltration of the chest wall.

4. Discussion

The evaluated AI-driven algorithm allows for excellent detection and segmentation of pulmonary T1 lesions (detection rate: 90.4%; excellent correlation of $Volume_{AI}$ and $Volume_{GT}$: $r = 0.91$) and good detection and segmentation of T2 tumors (detection rate: 70.8%; correlation of $Volume_{AI}$ and $Volume_{GT}$: $r = 0.80$) on the CT component of PET/CTs. Given the fact that the algorithm is designed for the detection of lung nodules smaller than 3 cm, such good performance on tumors with a diameter of up to 5 cm is remarkable. This is even truer considering the fact that the CT series used as input for the algorithm had a slice thickness of 3 mm and were acquired in free breathing and mostly nonenhanced technique. In more advanced tumors (T3/T4), detection and segmentation are more challenging and subsequently detection rates are low. Furthermore, the segmentation mask volumes for T3/T4 tumors systematically underestimate ground truth volumes. It is therefore an important finding that the tested CAD system has conceptional limitations concerning the detection of advanced lung tumors, and human inspection is still necessary in these cases.

The first step of CAD systems is to detect the location of lesions in medical images [30]. Most previous studies used CT datasets from lung cancer screening trials (e.g., NLST) with nodule size between 3 and 30 mm [19]. As an exception, Dandil et al. analyzed 52 malignant and 76 benign lesions with a size range from 4 to 58 mm, but only 12.5% of these nodules were bigger than 20 mm in diameter [20]. They reported a sensitivity of 92.3%, which is in line with the detection performance we found for the comparable group of T1 tumors. Earlier this year, Vassallo et al. compared unassisted and cloud-based CAD of pulmonary nodules in patients with extrathoracic malignancy [13]. A total of 215 lung nodules with a diameter between 3 and 28 mm in 75 patients were used for evaluation. Stand-alone CAD sensitivity was 85%, and the mean false-positive rate per scan was 3.8. These performance measures are representative for recently published studies on lung nodule CAD software [12,14–18,21]. Our results show a sensitivity of 90.4% for small tumors with a diameter of up to 30 mm with a far superior rate of false-positive findings per exam of 0.41 on the nodule negative population. This low rate of false-positive findings is a prerequisite for integration into existing clinical workflows and acceptance by radiologists and nuclear medicine physicians. Liang et al. tested four CAD systems at two time points for the detection of nodules with a mean diameter of 4 mm and 11 mm, respectively, and found sensitivities ranging from 52% to 82% [11]. Again, false-positive

rates of 0.6–7.4 per exam ranged above the ones we found and —in line with our results— were often caused by detection of blood vessels and bone. They did not identify dystelectasis as a reason for FP findings—the most frequent cause we found. This can be explained by the fact that we tested on PET/CTs acquired in free breathing technique, while Liang and colleagues evaluated on chest CTs acquired in deep-inspiration breath-hold technique [11]. Of interest and with only one exception, they as well as some other authors [31,32] reported higher detection rates of the CADs for isolated cancers as compared to those attached to the pleura. This supports our finding that pleural contact negatively affects detection. It is important to understand that these features are not totally independent from each other. For example, advanced tumors more likely invade structures adjacent to the lung, which means that pleural contact exists. Of interest, we found no dependency of lesion detection on the location within the lung, whereas Liang and colleagues reported a higher probability of detection for nodules in lower lobes for three of the four evaluated CAD systems [11]. However, the effect was small and not statistically significant.

Our radiomics analysis revealed further features that influence the detection rates: a finer, less heterogeneous and rounder texture was associated with better detection. While the utility of texture analysis for the differentiation of benign vs. malign lung lesions [33,34], the differentiation of histologic subtypes [35,36] and the prediction of progression [37–39] is well established, more studies on its influence on detection rates are warranted. Regarding tumor histology, our analysis revealed slightly lower detection rates for SCLC and squamous cell carcinomas as compared to adenocarcinomas. Due to the low number of cases in the two groups, however, these results are likely to be influenced by random effects. Another explanation could be that no preliminary stages of adenocarcinoma were included in our patient population. It is well known that adenocarcinoma with lepidic growth pattern has lower detection rates by human readers [40].

After detection, segmentation of lung lesions is the subsequent step that, if done correctly, paves the way to a plethora of secondary analyses that are currently developed within the context of AI, radiomics, and personalized medicine. In this context, Owens et al. compared contours of 10 lung tumors ranging from 1.1 cm^3 to 10.5 cm^3 defined by human readers in consensus, corresponding to our categories T1 and T2, with 2 semiautomatic segmentation methods: Lesion Sizing Toolkit (LSTK) and GrowCut [41]. For these semiautomatic tools, the mean Dice similarity coefficients were 0.88 ± 0.06 and 0.88 ± 0.08

for LSTK and GrowCut, respectively, indicating very good segmentation quality. Our results which reveal an excellent correlation of $Volume_{GT}$ and $Volume_{AI}$ for T1 ($r = 0.90$) and a good correlation for T2 tumors ($r = 0.70$) are in line with these findings. Various other studies assessed automated segmentation methods for the segmentation of lung nodules on the Lung Image Database Consortium-Image Database Resource Initiative (LIDC-IDRI) dataset (diameters: 2 mm–38 mm, again corresponding to T1 and T2-category of our dataset) and reported overlaps of ground truth and automatically generated segmentation masks of 50.7% [42], 58% [43], 63% [31], 69%, and 71.2% [44], respectively. Furthermore, Hassani et al. mention in their review that difficulties of semi-automated and fully automated systems in segmenting subpleural nodules are due to masking of margins by adjacent normal structures [45]. Our results confirm this finding, showing a much better correlation of $Volume_{GT}$ and $Volume_{AI}$ for isolated lesions ($r = 0.97$) as compared to attached lesions ($r = 0.59$).

According to current guidelines, FDG-PET/CT is considered the standard imaging procedure of choice for noninvasive staging of lung cancer [5]. The CT component of this examination is often acquired in free breathing using thicker slices (3 mm) and a lower dose compared to diagnostic chest CTs. In opposition to Marten et al., who reported significantly dropping detection rates for increasing reconstruction slice thicknesses (0.75 mm: 73.9%, 2 mm: 59.0%, 4 mm: 4.4%) [46], we found detection rates for the comparable T1-category collective that are equal or superior to those reported by other authors for 1 mm slice thickness. This can be explained by the fact that detection rates of DCNN detection algorithms used in our study are superior compared to techniques based on histogram analysis and thresholding used years ago. Teramoto et al. evaluated a CAD system that used both the CT and PET component to generate candidate lesions with a subsequent reduction of false-positive findings through a convolutional neural network (slice thickness: 2 mm; 104 cases with 183 nodules) [22]. They report a sensitivity regarding detection of 91% that is very similar to the one we found but a higher rate of false-positive findings per case (4.9). An inclusion of the information contained in the PET-component of the FDG-PET/CT could be a direction of further development of the CAD we tested.

There are several limitations of our work. First, manual segmentation was performed by two readers in random order without consensus or double reading. Both, consensus and double reading are time-consuming tasks and therefore not practicable in this study with a total of 320 lesions. Second, the assessment of segmentation quality was based on

comparison of the automatically calculated tumor volumes with ground truth volumes. More advanced methods like Dice similarity coefficients or Hausdorff distances could not be applied since space coordinates were not accessible in the manually created tumor masks. Third, for the creation of manual tumor masks, the FDG-PET component was considered whenever tumor borders could not be well delineated on the CT component, while automated tumor detection was performed only on the CT component. Inclusion of the information contained in the PET components could possibly increase detection rates and segmentation quality. Fourth, the analysis was conducted in two steps: detection and segmentation. Due to lower detection rates for more advanced tumors, a selection bias in step two of the analysis could positively influence segmentation performance in this group.

In conclusion, the tested algorithm facilitates a fast and reliable detection and 3D segmentation of pulmonary T1 and T2 tumors that also works well on the CT component of PET/CTs acquired in free breathing and with a slice thickness of 3 mm. The detection and segmentation of more advanced lung tumors is currently imprecise due to the conception of the algorithm for lung nodules. Consequently, there is still an unmet need for CAD applications that also cope with the more complex segmentation tasks required in the context of lung cancer staging. Future efforts must therefore focus on this collective to facilitate segmentation of all tumor types and sizes and bridge the gap between CAD applications for screening and staging of lung cancer.

Acknowledgments

We want to thank Victor Parmar for proofreading the article. The manual segmentation masks were acquired during the project “LungStage—Computer Aided Staging of Non-Small Cell Lung Cancer (NSCLC),” funded by CTI (Commission for Technology and Innovation) (Project no. 25280.1).

References

1. S. R. Baker, R. H. Patel, L. Yang, V. M. Lelkes, and A. Castro, "Malpractice suits in chest radiology," Journal of Thoracic Imaging, vol. 28, no. 6, pp. 388–391, 2013.
2. A. D. Ciello, P. Franchi, A. Contegiacomo, G. Cicchetti, L. Bonomo, and A. R. Larici, "Missed lung cancer: when, where, and why?" Diagnostic and Interventional Radiology, vol. 23, no. 2, pp. 118–126, 2017.
3. A. Masood, B. Sheng, P. Li et al., "Computer-assisted decision support system in pulmonary cancer detection and stage classification on CT images," Journal of Biomedical Informatics, vol. 79, pp. 117–128, 2018.
4. M. Silva, C. M. Schaefer-Prokop, C. Jacobs et al., "Detection of subsolid nodules in lung cancer screening," Investigative Radiology, vol. 53, no. 8, pp. 441–449, 2018.
5. A. Kandathil, F. U. Kay, Y. M. Butt, J. W. Wachsmann, and R. M. Subramaniam, "Role of FDG PET/CT in the eighth edition of TNM staging of non-small cell lung cancer," RadioGraphics, vol. 38, no. 7, pp. 2134–2149, 2018.
6. R. Rami-Porta, V. Bolejack, D. J. Giroux et al., "The IASLC lung cancer staging project: the new database to inform the eighth edition of the TNM classification of lung cancer," Journal of Thoracic Oncology, vol. 9, no. 11, pp. 1618–1624, 2014.
7. V. Greenberg, I. Lazarev, Y. Frank, J. Dudnik, S. Ariad, and I. Shelef, "Semi-automatic volumetric measurement of response to chemotherapy in lung cancer patients: how wrong are we using RECIST?" Lung Cancer, vol. 108, pp. 90–95, 2017.
8. G. Della Gala, M. L. P. Dirkx, N. Hoekstra et al., "Fully automated VMAT treatment planning for advanced-stage NSCLC patients," Strahlentherapie und Onkologie, vol. 193, no. 5, pp. 402–409, 2017.
9. R. Thawani, M. McLane, N. Beig et al., "Radiomics and radiogenomics in lung cancer: a review for the clinician," Lung Cancer, vol. 115, pp. 34–41, 2018.
10. R. Sexauer, T. Weikert, K. Mader et al., "Towards more structure: comparing TNM staging completeness and processing time of text-based reports versus fully segmented and annotated PET/CT data of non-small-cell lung cancer," Contrast Media & Molecular Imaging, vol. 2018, Article ID 5693058, 10 pages, 2018.
11. M. Liang, W. Tang, D. M. Xu et al., "Low-dose CT screening for lung cancer: computer-aided detection of missed lung cancers," Radiology, vol. 281, no. 1, pp. 279–288, 2016.
12. Q. Wang, W. Zhu, and B. Wang, "Three-dimensional SVM with latent variable: application for detection of lung lesions in CT images," Journal of Medical Systems, vol. 39, no. 1, p. 171, 2015.
13. L. Vassallo, A. Traverso, M. Agnello et al., "A cloud-based computer-aided detection system improves identification of lung nodules on computed tomography scans of patients with extra-thoracic malignancies," European Radiology, vol. 29, no. 1, pp. 144–152, 2019.
14. C. Li, G. Zhu, X. Wu, and Y. Wang, "False-positive reduction on lung nodules detection in chest radiographs by ensemble of convolutional neural networks," IEEE Access, vol. 6, pp. 16060–16067, 2018.
15. J. Gong, J.-Y. Liu, L.-J. Wang, X.-W. Sun, B. Zheng, and S.-D. Nie, "Automatic detection of pulmonary nodules in CT images by incorporating 3D tensor filtering with local image feature analysis," Physica Medica, vol. 46, pp. 124–133, 2018.
16. A. Gupta, T. Saar, O. Martens, and Y. L. Moullec, "Automatic detection of multisize pulmonary nodules in CT images: large-scale validation of the false-positive reduction step," Medical Physics, vol. 45, no. 3, pp. 1135–1149, 2018.
17. A. A. A. Setio, F. Ciompi, G. Litjens et al., "Pulmonary nodule detection in CT images: false positive reduction using multi-view convolutional networks," IEEE Transactions on Medical Imaging, vol. 35, no. 5, pp. 1160–1169, 2016.
18. M. Javaid, M. Javid, M. Z. U. Rehman, and S. I. A. Shah, "A novel approach to CAD system for the detection of lung nodules in CT images," Computer Methods and Programs in Biomedicine, vol. 135, pp. 125–139, 2016.

19. G. Zhang, S. Jiang, Z. Yang et al., "Automatic nodule detection for lung cancer in CT images: a review," Computers in Biology and Medicine, vol. 103, pp. 287–300, 2018.
20. E. Dandil, M. Cakiroglu, Z. Eksi et al., "Artificial neural network-based classification system for lung nodules on computed tomography scans," in Proceedings of the 2014 6th International Conference of Soft Computing and Pattern Recognition (SoCPaR), pp. 382–386, IEEE, Tunis, Tunisia, August 2014.
21. S. Saien, H. A. Moghaddam, and M. Fathian, "A unified methodology based on sparse field level sets and boosting algorithms for false positives reduction in lung nodules detection," International Journal of Computer Assisted Radiology and Surgery, vol. 13, no. 3, pp. 397–409, 2018.
22. A. Teramoto, H. Fujita, O. Yamamuro, and T. Tamaki, "Automated detection of pulmonary nodules in PET/CT images: ensemble false-positive reduction using a convolutional neural network technique," Medical Physics, vol. 43, no. 6, pp. 2821–2827, 2016.
23. J. Zhao, G. Ji, Y. Qiang, X. Han, B. Pei, and Z. Shi, "A new method of detecting pulmonary nodules with PET/CT based on an improved watershed algorithm," PLoS One, vol. 10, no. 4, Article ID e0123694, 2015.
24. D. Yang, D. Xu, S. K. Zhou et al., "Automatic liver segmentation using an adversarial image-to-image network," 2017, https://arxiv.org/abs/1707.08037.
25. S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," 2015, https://arxiv.org/abs/1506.01497.
26. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, June 2015.
27. S. A. Hojjatoleslami and J. Kittler, "Region growing: a new approach," IEEE Transactions on Image Processing, vol. 7, no. 7, pp. 1079–1084, 1998.
28. J. J. M. van Griethuysen, A. Fedorov, C. Parmar et al., "Computational radiomics system to decode the radiographic phenotype," Cancer Research, vol. 77, no. 21, pp. e104–e107, 2017.
29. W. J. Youden, "Index for rating diagnostic tests," Cancer, vol. 3, no. 1, pp. 32–35, 1950.
30. J. Cai, D. Xu, S. Liu, and M. D. Cham, "The added value of computer-aided detection of small pulmonary nodules and missed lung cancers," Journal of Thoracic Imaging, vol. 33, p. 1, 2018.
31. T. Messay, R. C. Hardie, and S. K. Rogers, "A new computationally efficient CAD system for pulmonary nodule detection in CT imagery," Medical Image Analysis, vol. 14, no. 3, pp. 390–406, 2010.
32. J. Jiang, Y.-C. Hu, C.-J. Liu et al., "Multiple resolution residually connected feature streams for automatic lung tumor segmentation from CT images," IEEE Transactions on Medical Imaging, vol. 38, no. 1, pp. 134–144, 2019.
33. W. Choi, J. H. Oh, S. Riyahi et al., "Radiomics analysis of pulmonary nodules in low-dose CT for early detection of lung cancer," Medical Physics, vol. 45, no. 4, pp. 1537–1549, 2018.
34. C.-H. Chen, C.-K. Chang, C.-Y. Tu et al., "Radiomic features analysis in computed tomography images of lung nodule classification," PLoS One, vol. 13, no. 2, Article ID e0192002, 2018.
35. E. Linning, L. Lin, L. Li, H. Yang, L. H. Schwartz, and B. Zhao, "Radiomics for classifying histological subtypes of lung cancer based on multiphasic contrast-enhanced computed tomography," Journal of Computer Assisted Tomography, vol. 43, no. 2, pp. 300–306, 2019.
36. X. Zhu, D. Dong, Z. Chen et al., "Radiomic signature as a diagnostic factor for histologic subtype classification of non-small cell lung cancer," European Radiology, vol. 28, no. 7, pp. 2772–2778, 2018.
37. L. Shi, Y. He, Z. Yuan et al., "Radiomics for response and outcome assessment for non-small cell lung cancer," Technology in Cancer Research & Treatment, vol. 17, Article ID 153303381878278, 2018.

38. A. Chaddad, C. Desrosiers, M. Toews, and B. Abdulkarim, "Predicting survival time of lung cancer patients using radiomic analysis," Oncotarget, vol. 8, no. 61, pp. 104393–104407, 2017.
39. B. Ganeshan, E. Panayiotou, K. Burnand, S. Dizdarevic, and K. Miles, "Tumour heterogeneity in non-small cell lung carcinoma assessed by CT texture analysis: a potential marker of survival," European Radiology, vol. 22, no. 4, pp. 796–802, 2012.
40. Z.-G. Yang, S. Sone, F. Li et al., "Visibility of small peripheral lung cancers on chest radiographs: influence of densitometric parameters, CT values and tumour type," British Journal of Radiology, vol. 74, no. 877, pp. 32–41, 2001.
41. C. A. Owens, C. B. Peterson, C. Tang et al., "Lung tumor segmentation methods: impact on the uncertainty of radiomics features for non-small cell lung cancer," PLoS One, vol. 13, no. 10, Article ID e0205003, 2018.
42. R. Tachibana and S. Kido, "Automatic segmentation of pulmonary nodules on CT images by use of NCI lung image database consortium," in Proceedings of the Medical Imaging 2006: Image Processing, J. M. Reinhardt and J. P. W. Pluim, Eds., International Society for Optics and Photonics, San Diego, CA, USA, 2006.
43. Q. Wang, E. Song, R. Jin et al., "Segmentation of lung nodules in computed tomography images using dynamic programming and multidirection fusion Techniques," Academic Radiology, vol. 16, no. 6, pp. 678–688, 2009.
44. S. Wang, M. Zhou, Z. Liu et al., "Central focused convolutional neural networks: developing a data-driven model for lung nodule segmentation," Medical Image Analysis, vol. 40, pp. 172–183, 2017.
45. C. Hassani, B. A. Varghese, J. Nieva, and V. Duddalwar, "Radiomics in pulmonary lesion imaging," American Journal of Roentgenology, vol. 212, no. 3, pp. 497–504, 2019.
46. K. Marten, A. Grillhösl, T. Seyfarth, S. Obenauer, E. J. Rummeny, and C. Engelke, "Computer-assisted detection of pulmonary nodules: evaluation of diagnostic performance using an expert knowledge-based detection system with variable reconstruction slice thickness settings," European Radiology, vol. 15, no. 2, pp. 203–212, 2005.

CHAPTER 6

Prediction of Patient Management in COVID–19 Using Deep Learning-Based Fully Automated Extraction of Cardiothoracic CT Metrics and Laboratory Findings

BASED ON: Weikert T, Rapaka S, Grbic S, Re T, Chaganti S, Winkel DJ, Anastasopoulos C, Niemann T, Wiggli BJ, Bremerich J, Twerenbold R, Sommer G, Comaniciu D, Sauter AW. Prediction of Patient Management in COVID-19 Using Deep Learning-Based Fully Automated Extraction of Cardiothoracic CT Metrics and Laboratory Findings. Korean J Radiol. 2021 Jun;22(6):994-1004. doi: 10.3348/kjr.2020.0994.

Abstract

Objective: To extract pulmonary and cardiovascular metrics from chest CTs of patients with coronavirus disease 2019 (COVID-19) using a fully automated deep learning-based approach and assess their potential to predict patient management.

Materials and Methods: All initial chest CTs of patients who tested positive for severe acute respiratory syndrome coronavirus 2 at our emergency department between March 25 and April 25, 2020, were identified (n = 120). Three patient management groups were defined: group 1 (outpatient), group 2 (general ward), and group 3 (intensive care unit [ICU]). Multiple pulmonary and cardiovascular metrics were extracted from the chest CT images using deep learning. Additionally, six laboratory findings indicating inflammation and cellular damage were considered. Differences in CT metrics, laboratory findings, and demographics between the patient management groups were assessed. The potential of these parameters to predict patients' needs for intensive care (yes/no) was analyzed using logistic regression and receiver operating characteristic curves. Internal and external validity were assessed using 109 independent chest CT scans.

Results: While demographic parameters alone (sex and age) were not sufficient to predict ICU management status, both CT metrics alone (including both pulmonary and cardiovascular metrics; area under the curve [AUC] = 0.88; 95% confidence interval [CI] = 0.79–0.97) and laboratory findings alone (C-reactive protein, lactate dehydrogenase, white blood cell count, and albumin; AUC = 0.86; 95% CI = 0.77–0.94) were good classifiers. Excellent performance was achieved by a combination of demographic parameters, CT metrics, and laboratory findings (AUC = 0.91; 95% CI = 0.85–0.98). Application of a model that combined both pulmonary CT metrics and demographic parameters on a dataset from another hospital indicated external validity (AUC = 0.77; 95% CI = 0.66–0.88).

Conclusion: Chest CTs of patients with COVID-19 contain valuable information that can be accessed using automated image analysis. These metrics are useful for the prediction of patient management.

1. Introduction

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has caused a global pandemic with over 1.28 million deaths worldwide as of November 12, 2020 [1]. The associated infectious disease, named coronavirus disease 2019 (COVID-19), progresses mildly in most cases [2]. However, severe and critical courses of the disease occur in approximately 20% of patients [3], mostly demonstrating atypical pneumonia [4]. These patients require hospitalization or even intensive care unit (ICU) treatment. There are regional differences in utilization of these scarce resources during a pandemic with temporary shortages. Therefore, criteria for early prediction of patient management, especially whether ICU care is needed or not, are important.

While viral testing remains the only specific method of diagnosis [5], CT plays a role in the workup of suspected pulmonary manifestations of COVID-19 and associated complications. There is growing evidence that radiographic [6] and chest CT [7,8,9,10,11,12,13,14,15] features are associated with disease severity in COVID-19 based on (semi)-manual assessment and visual scoring of pulmonary parameters. This study intends to build on these approaches and expand them in three aspects: First, by introducing a fully automated and user-independent evaluation method, which is especially relevant in a pandemic with heavy workloads on healthcare providers. Second, this study explicitly focusses on the ultimate patient management status defined by a patient's clinical pathway established with sufficient temporal distance. Third, the inclusion of five cardiovascular metrics that are derivable from all chest CTs has rarely been reported systematically. Notably, preexisting cardiovascular disease is a major risk factor for adverse outcomes in COVID-19 [16]. Laboratory findings were included to assess the value added by the CT metrics.

We hypothesized that pulmonary and cardiovascular CT metrics are associated with ultimate patient management in patients with COVID-19. It is the goal of this study to extract these CT metrics using a fully automated deep learning-based approach and assess their potential, alone and in combination with laboratory findings and demographic data, for the prediction of patient management.

2. Materials and Methods

This study was approved by the local ethics committee (Ethikkommission Nordwest-und Zentralschweiz; IRB approval number: 2020-00566). It is part of a research project registered on ClinicalTrials.gov on, April 04/29/2020 (Identifier: NCT04366765).

2.1. Study population

All reverse-transcription polymerase chain reaction (RT-PCR) results for SARS-CoV-2 performed at the emergency department (ED) of our institution between March 25 and April 25, 2020, were downloaded from our laboratory database (n = 6080 RT-PCR results in 5120 patients). RT-PCR for SARS-CoV-2 was performed using specimens from nasopharyngeal and oropharyngeal swabs. All patients RT-PCR positive for COVID-19 were identified (n = 438). In cases with multiple RT-PCRs, a patient was rated positive if at least one of the specimens was positive. For the 438 patients, we searched our RIS/PACS system for the chest CTs performed during the study period, which resulted in 169 chest CTs. At our institution, chest CT is the imaging standard for verifying suspected pulmonary involvement in patients with SARS-CoV-2. For ensuring the independence of observations, all follow-up CTs from a given patient were excluded from the analysis (n = 49). This resulted in 120 CT scans in 120 patients. The time interval between the presentation at the ED and CT acquisition was determined. Figure 1 illustrates the search strategy.

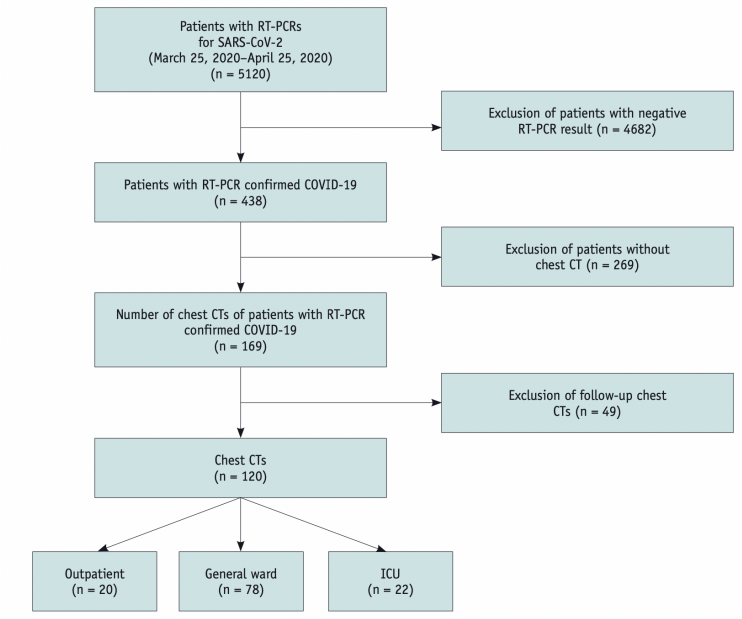


Fig. 1. Search strategy for the main analysis dataset. COVID-19 = coronavirus disease 2019, ICU = intensive care unit, RT-PCR = reverse-transcription polymerase chain reaction, SARS-CoV-2 = severe acute respiratory syndrome coronavirus 2

2.1. Definition of patient management

Information on the ultimate clinical pathway of a patient was retrieved from our hospital information system 12 weeks after completion of CT data collection (date of determination of ultimate patient management: July 20, 2020). Based on this information, the following three groups were defined: group 1 (outpatient treatment), group 2 (inpatient treatment, general ward), and group 3 (admission to ICU). Each patient was assigned to the highest category individually reached (for instance, a patient that had initially been treated on the general ward and eventually needed ICU care was assigned to group 3).

2.2. CT acquisition parameters

Chest CT scans were acquired in supine position using two 128-slice scanners: SOMATOM Definition AS+ (n = 119) and SOMATOM Force (n = 1) (both Siemens Healthineers). Mean tube voltage was 105.0 kVp (standard deviation [SD]: 10.1), mean tube current-time product 81.1 mAs (SD: 19.2), and pitch factor 1.05 in all cases. Most of the scans were performed without a contrast agent (n = 99), whereas 21 CTs were performed with a mean of 71.8 mL (SD: 17.2) of contrast agent (Iopromide, Bayer AG) at an injection rate of 4 mL/s

for excluding pulmonary embolism. The 1-mm soft-tissue kernel series served as input to the algorithms.

2.3. Laboratory Findings

For all patients, the results of six standard laboratory parameters of inflammation and tissue damage were retrieved from our laboratory system (blood sample type in parentheses): C-reactive protein (CRP; heparin plasma), lactate dehydrogenase (heparin plasma), white blood cell count (EDTA), procalcitonin (heparin plasma), albumin (heparin plasma), and D-dimers (citrate plasma). Laboratory results were obtained on the day of chest CT acquisition.

2.4. Technical Details of the Algorithms

Multiple deep convolutional neural networks (DCNNs) were locally deployed on an imaging post-processing platform (Siemens Healthineers, Corporate Technology).

2.5. Pulmonary Metrics

The 1-mm series in soft kernel reconstruction served as input to an algorithm prototype based on a deep image-to-image network for lung and lung lobe segmentation and a subsequent DenseUNet for segmentation of opacities. They were trained on chest CTs of n = 9549 (Deep-Image-to-Image Network) and 901 (DenseUNet) patients, completely independent of the testing dataset used in this study. DenseUNet defined all voxels with ground-glass opacity (GGO) or consolidation as positive/foreground and all other areas of the lung as negative/background. Subsequently, a Hounsfield unit (HU) threshold of -200 was applied to the prediction mask for differentiating GGO from consolidations. Table 1 provides details for all metrics. Further technical details and high diagnostic performance of the algorithms have been reported previously [17].

Table 1. Features Analyzed by Algorithms with Units and Definitions

Metric Name	Unit	Definition
Pulmonary metrics		
Lung volume	mL	Total volume of the lung
PO	%	Percentage of volume of the lung affected by opacities, equivalent to GGO and consolidation
PHO	%	Percentage of volume of the lung affected by high HU opacities (equal to or above -200 HU), equivalent to consolidations
%LowHU	%	Percentage of volume of the lung with low attenuation, defined as < -950 HU
LL#	N	Number of lung lesions detected (GGO or consolidations)
LSS	N	Sum of severity scores for each of the five lung lobes*
LHOS	N	Sum of severity scores for each of the five lung lobes, consolidation only*
Cardiovascular metrics		
TPV	mL	Total pericardial volume. This includes the heart as well as pericardial structures like fat and/or effusion
QCC	mm^3	Total volume of coronary calcifications. Only available for CT without contrast
D_AAsc	mm	Diameter of ascending aorta at the level of the right pulmonary artery
D_Arch	mm	Diameter of aorta, midway through aortic arch region
D_ADsc	mm	Diameter of descending aorta, midway between the branching of the subclavian artery and celiac artery

*Severity scores calculated as follows. 0: lobe not affected by GGO/consolidation, 1: 1–25% of lobe volume affected, 2: 25–50%, 3: 50–75%, 4: 75–100%. GGO = ground-glass opacity, HU = Hounsfield unit, QCC = quantification of coronary calcification

2.6. Cardiovascular Metrics

The non-electrocardiogram-gated 1-mm series served as the only input to a DCNN based on U-Net architecture for segmentation of the thoracic aorta and the total pericardial volume (TPV). TPV segmentation, which includes the heart and pericardial structures, such as fat and (if present) pericardial effusion, was used to identify candidate coronary calcification voxels by applying a threshold of > 130 HU. The calcium detection model based on ResNet subsequently predicts true coronary calcifications. The diameters of the aorta were computed at key anatomical landmarks. The cardiovascular algorithms were trained using 3550 CT scans. Detailed information has been provided elsewhere [18,19]. The quantification of coronary calcifications (QCCs) could only be calculated for series without contrast (n = 99/120). Table 1 lists all the cardiovascular metrics analyzed in this study.

2.7. Statistical Analysis

Categorical variables were expressed as counts and percentages. For continuous variables, means with corresponding SDs are provided as measures of variance. For comparing the differences between groups, one-way analyses of variance for normally distributed continuous variables, the Kruskal-Wallis H tests for non-normally distributed continuous variables, and the chi-square tests for categorical variables were performed. The statistical analysis comprised the following three steps:

Step 1: A series of univariable analyses with appropriate post hoc tests to assess the association of CT metrics, laboratory findings, and patient characteristics with patient management. Studies with contrast were excluded during the analysis of QCC, as this measure was only calculated on non-contrast series.

Step 2: A series of multivariable binary logistic regressions to assess the potential of CT metrics, laboratory findings, and patient demographics as well as their combinations, to classify patients who needed ICU care from those who did not. The ICU status of a patient (0 = no ICU; 1 = ICU) served as the dependent variable. CT metrics, laboratory findings, and patient demographics (age and sex) served as independent variables. Inclusion criteria for CT metrics and laboratory findings were p values ≤ 0.05 in the subgroup comparisons between groups 1 and 3 (outpatient vs. ICU) or group 2 vs. 3 (general ward vs. ICU) in Step 1 of the analysis. Furthermore, the parameters had to be available for all patients. This resulted in the following five models:

- D: Patients' demographics only
- L: Laboratory findings only
- PC: CT metrics only
- PD: Pulmonary CT metrics and demographics
- PCLD: All parameters (CT metrics, laboratory findings, demographics)

For all approaches, area under the curve (AUC) with 95% confidence intervals (CIs) were calculated using prediction probabilities obtained from the binary logistic regression analyses. Furthermore, we analyzed the differences in the AUCs between the models according to the method proposed by DeLong et al. [20].

Step 3: Internal and external validation to assess generalizability. For internal validation, we processed all chest CTs of patients with positive RT-PCR for SARS-CoV-2 acquired at our institution during a later period (April 26, 2020–May 20, 2020; information on ultimate patient management retrieved on August 12, 2020) with all algorithms. For testing external validity, we used data from another hospital (Supplementary Materials, Supplementary Table 1). ICU status was predicted using regression equations obtained from Step 2 of the analysis.

Statistical analyses were performed with IBM SPSS Statistics for Windows, Version 22.0 (IBM Corp.), using default settings. *P* values ≤ 0.05 were defined to indicate statistical significance.

3. Results

3.1. Patient characteristics

The main analysis dataset of this study included 120 patients with a mean age of 60.8 years (SD: 17.5; range: 18–92 years; 47 [39.2%] females). Table 2 summarizes the patient characteristics of the three patient management groups.

Table 2. Patient Characteristics in Patient Management Groups with Intergroup Comparison

	Group 1 (Outpatient)	Group 2 (General Ward)	Group 3 (ICU)	*P* (Intergroup Comparison)
Number of patients	20	78	22	
Mean age ± SD, year	58.2 ± 19.3	62.1 ± 17.4	63.1 ± 14.6	0.100
Sex, % female	35.0 (7/20)	41.0 (32/78)	36.4 (8/22)	0.847

ICU = intensive care unit, SD = standard deviation

3.2. Automated analysis of CT metrics

All datasets were successfully processed using the algorithm. The mean time interval between presentation at ED and CT acquisition was 0.98 days (SD: 2.32 days).

Step 1: Association of Metrics with Patient Management

Table 3 summarizes the results of the univariable analyses of CT metrics and laboratory findings.

Pulmonary CT metrics

PO, PHO, LSS, and LHOS increased continuously from group 1 to group 3, while lung volume and %LowHU decreased from group 1 to group 3. All these differences were statistically significant. Post hoc testing revealed that differences in PO, PHO, LSS, and LHOS were statistically significant at *p* values < 0.01 between all three subgroups. Regarding lung volume and %LowHU, group comparisons 1 vs. 3 and 2 vs. 3 differed statistically significantly. Figure 2 displays an image example for each group.

Cardiovascular CT metrics

TPV and D_AAsc differed significantly among the three groups. Post hoc analysis revealed that differences in both TPV and D_AAsc were statistically significant only for the comparison of groups 1 and 3 (TPV: $p = 0.041$; D_AAsc: $p = 0.033$). QCC, D_Arch, and D_ADsc did not differ significantly. Figure 3 illustrates the outputs of the cardiovascular algorithms.

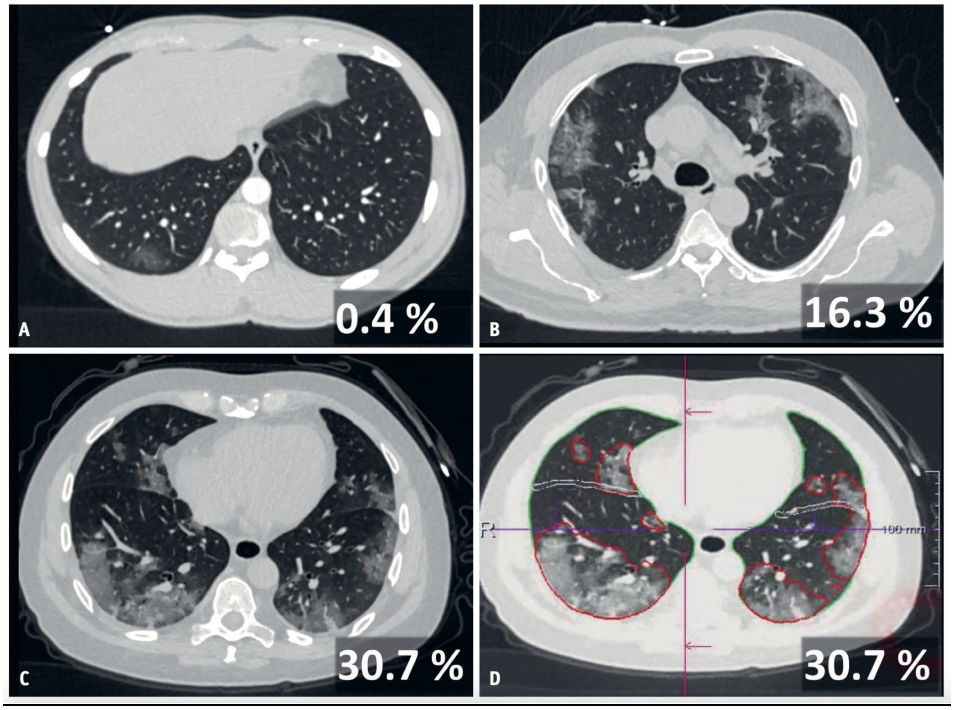


Fig. 2. Exemplary axial chest CT images for groups 1 (A), 2 (B), and 3 (C) with PO in percent in the lower right corner. (D) shows the output of pulmonary opacity segmentation algorithm for case (C), with fine delineation of areas with pulmonary opacities in red color.

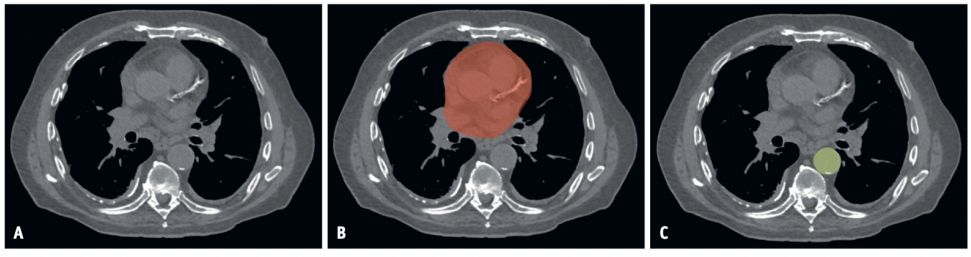


Fig. 3. Exemplary axial chest CT images visualizing the output of cardiovascular algorithms: original input image (A), segmentation of the total pericardial volume as a red overlay (B), and segmentation of the descending aorta as a yellow overlay (C) as the basis for diameter calculations.

Laboratory findings

Laboratory analysis for D-dimers and procalcitonin was not performed in some cases (11/120 and 36/120, respectively). All laboratory parameters differed significantly among the three groups. CRP levels increased steadily from groups 1 to 3, while albumin decreased. Subgroup comparisons were statistically significant for all group comparisons (CRP, albumin), group comparisons 1 vs. 3, and 2 vs. 3 (lactate dehydrogenase and procalcitonin), group comparison 2 vs. 3 only (white blood cell count), and group comparison 1 vs. 3 only (D-dimer).

Table 3. Results of Univariable Analyses of CT Metrics and Laboratory Findings

Parameter	Unit	Group 1	Group 2	Group 3	Statistic	*P*
Pulmonary metrics						
Lung volume	mL	4917.4 (1286.2)	4336.5 (1205.5)	3611.3 (1059.5)	ANOVA	0.002*
PO	%	3.7 (6.1)	15.6 (15.9)	39.4 (24.1)	Kruskal-Wallis	< 0.001*
PHO	%	0.6 (1.2)	3.3 (4.5)	10.1 (8.6)	Kruskal-Wallis	< 0.001*
%LowHU	%	7.8 (6.8)	5.4 (4.3)	2.7 (2.4)	Kruskal-Wallis	< 0.001*
LL#	N	17.3 (10.9)	28.7 (17.3)	21.6 (10.7)	ANOVA	0.061
LSS	N	2.3 (2.2)	5.7 (3.5)	10.8 (4.8)	Kruskal-Wallis	< 0.001*
LHOS	N	0.9 (1.3)	2.8 (2.0)	5.1 (2.2)	Kruskal-Wallis	< 0.001*
Cardiovascular metrics						
TPV	mL	762.1 (194.5)	881.3 (236.6)	908.8 (166.1)	Kruskal-Wallis	0.043*
QCC	mm^2	321.4 (700.3)	335.8 (633.6)	461.9 (915.6)	ANOVA	0.552
D_AAsc	mm	32.9 (3.9)	35.5 (4.4)	36.6 (4.1)	Kruskal-Wallis	0.038*
D_Arch	mm	28.8 (3.5)	29.6 (2.9)	30.4 (2.7)	Kruskal-Wallis	0.184
D_ADsc	mm	25.0 (4.3)	27.2 (2.6)	26.0 (4.3)	Kruskal-Wallis	0.215
Laboratory findings						
CRP	mg/L	32.5 (59.9)	62.0 (73.5)	104.5 (81.9)	Kruskal-Wallis	< 0.001*
Lactate de-hydrogenase	U/L	232.7 (84.6)	283.0 (159.7)	406.8 (144.0)	Kruskal-Wallis	< 0.001*
White blood cell count	10^9/L	6.9 (2.9)	6.9 (3.4)	9.4 (4.3)	Kruskal-Wallis	0.035*
Procalcitonin	µ/L	0.099 (0.141)	0.205 (0.531)	0.399 (0.427)	Kruskal-Wallis	< 0.001*
Albumin	g/L	35.8 (4.5)	30.8 (6.3)	25.32 (6.4)	Kruskal-Wallis	< 0.001*
D-Dimer	µ/mL	0.472 (0.302)	1.113 (1.438)	1.397 (1.401)	Kruskal-Wallis	0.016*

Data are mean values and standard deviations (in brackets). *$p \leq 0.05$. CRP = C-reactive protein, QCC = quantification of coronary calcification, TPV = total pericardial volume

Step 2: Prediction of ICU Status

Table 4 specifies metrics and parameters included in the five multivariable models for classification of ICU status (yes/no) according to the criteria mentioned in the methods section. The best performing model was the PCLD model combining CT-derived, laboratory, and demographic parameters (AUC = 0.91). Demographic parameters alone could not distinguish ICU patients from non-ICU patients (AUC = 0.55). CT-derived metrics (including both pulmonary and cardiovascular metrics) alone, laboratory metrics alone, and pulmonary CT metrics combined with demographic parameters were all good classifiers with AUCs ≥ 0.84. Table 5 provides detailed information. The AUC of the D model differed significantly from that of all other models ($p < 0.001$). The difference in the

AUCs of models with CT-derived parameters alone vs. laboratory parameters alone was not statistically significant (p = 0.462). Figure 4 displays the receiver operating characteristic curves of the PCLD, PC, and L models. As D-dimers and procalcitonin were not available in all cases, these two parameters were excluded from the analysis.

Table 4. Metrics Included in Five Multivariable Models for Prediction of Patient Management Regarding Need of ICU-Care (Yes/No)

Model	Metrics and Parameters Included
D	Age, sex
L	CRP, lactate dehydrogenase, white blood cell count, albumin
PC	Lung volume, PO, PHO, %LowHU, LSS, LHOS, TPV, D_AAsc
PD	Lung volume, PO, PHO, age, sex
PCLD	Lung volume, PO, PHO, %LowHU, LSS, LHOS, TPV, D_AAsc, CRP, lactate dehydrogenase, white blood cell count, albumin, age, sex

Name of models indicates which metric/parameter groups were included as independent variables. C = cardiovascular metrics, CRP = C-reactive protein, D = demographic paramters, ICU = intensive care unit, L = laboratory findings, P = pulmonary CT metrics

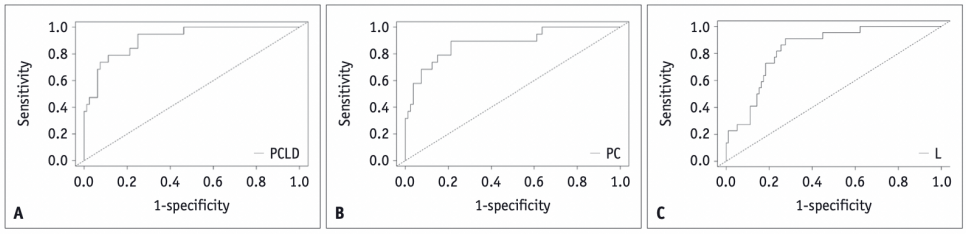


Fig. 4. Receiver operating characteristic curves for the PCLD model (A), PC model (B), and L model (C) to assess discriminatory power regarding ICU-care on main analysis dataset (whether a patient finally required ICU care or not). A combination CT metrics, laboratory findings, and patient demographics was the best classifier (**A;** area under the curve = 0.91; 95% confidence interval = 0.85–0.98). ICU = intensive care unit

Step 3: Internal and External Validation

The internal validation comprising 16 new cases of patients with positive RT-PCR results for SARS-CoV-2 resulted in a sensitivity of 80.0% (4 of 5 patients admitted to ICU correctly classified) and a specificity of 81.8% (9 of 11 patients not admitted to ICU correctly classified) for the PCLD model. Table 5 presents the results. In general, the performance measures on the internal validation dataset were slightly worse than those on the main analysis dataset but still acceptable. The mean age of the internal validation dataset was 63.0 years (SD: 16.0) and not statistically significantly different from the dataset used for the main analysis (p = 0.635). We also found evidence for external validity using the PD model (detailed information in Supplement). Table 6 demonstrates that demographic information did not differ significantly among the three datasets.

Table 5. Performance Measures of Five Multivariable Models for Prediction of Patient Management Regarding Need of ICU-Care (Yes/No)

Model	Statistic	Main Analysis Dataset	Internal Validation Dataset	External Validation Dataset
D	Sensitivity (%)	63.6 (14/22)	40 (2/5)	NA
	Specificity (%)	50.0 (49/98)	36.4 (4/11)	NA
	AUC (95% CI)	0.55 (0.42-0.67)	0.42 (0.15–0.79)	NA
L	Sensitivity (%)	90.9 (20/22)	80.0 (4/5)	NA
	Specificity (%)	73.5 (72/98)	45.5 (5/11)	NA
	AUC (95% CI)	0.86 (0.77–0.94)	0.47 (0.15–0.79)	NA
PC	Sensitivity (%)	90.9 (20/22)	80.0 (4/5)	NA
	Specificity (%)	78.6 (77/98)	72.7 (8/11)	NA
	AUC (95% CI)	0.88 (0.79–0.97)	0.75 (0.47–1.00)	NA
PD	Sensitivity (%)	81.8 (18/22)	60.0 (3/5)	74.2 (23/31)
	Specificity (%)	70.4 (69/98)	72.7 (8/11)	75.8 (47/62)
	AUC (95% CI)	0.84 (0.75–0.94)	0.71 (0.39–1.00)	0.77 (0.66–0.88)
PCLD	Sensitivity (%)	95.5 (21/22)	80.0 (4/5)	NA
	Specificity (%)	75.5 (74/98)	81.8 (9/11)	NA
	AUC (95% CI)	0.91 (0.85–0.98)	0.75 (0.48–1.00)	NA

Sensitivity and specificity at optimal cutoff according to Youden with corresponding number of patients in brackets. AUC = area under the curve, CI = confidence interval, ICU = intensive care unit, NA = not available

4. Discussion

This study demonstrated that it is feasible to automatically extract pulmonary and cardiovascular metrics from chest CT scans of patients with RT-PCR-confirmed COVID-19. Those metrics are useful for the prediction of patient management. Multiple CT metrics continuously and significantly increased or decreased with intensified patient management. The same was true for laboratory parameters reflecting inflammation and cell damage. The best prediction regarding ICU status was achieved by combining CT metrics, laboratory findings, and demographic information, while the latter alone could not differentiate the two classes. The CT metrics and laboratory findings were good classifiers on their own. Internal and external validation demonstrated marginally inferior performance.

Our results regarding the relevance of pulmonary CT metrics in COVID-19 and their association with patient management are in line with previous studies and expected, as they reflect pathologic changes, concretely inflammatory GGO, and consolidations. Li et al. [8] reported an increasing extent of inflammatory pulmonary lesions from light to common to severe/critical clinical manifestations. Sun et al. [7] and Tan et al. [21] confirmed that quantitative CT parameters strongly correlate with laboratory inflammation markers. Lyu et al. [22] showed that the number of lung segments and lobes affected by consolidations increased with case severity, which is in line with the increase in LSS and LHOS with higher admission status [22]. Similarly, Liu et al. [10] reported an

association between a higher lung severity score and extended hospitalization time. A significant number of additional studies have successfully applied lung volume assessment with or without a combination of clinical and laboratory tests for predicting disease severity, treatment intensity, outcome, and mortality [7,10,15,23,24,25,26]. Notably, the analyses in these studies required substantial manual interaction and visual assessment.

However, in a pandemic with limited human resources, fully automated approaches are preferred. In this respect, Huang et al. [27] applied CT-derived opacification measures using deep learning to stratify four clinical subtypes according to their baseline clinical, laboratory, and CT findings. They provided further evidence of CT as an important tool for risk stratification in patients with COVID-19 and reported percentages of lung areas with opacities ranging from 0% (mild disease) to 49.6% (critical disease), which is in line with the results of the analysis at hand. However, radiological findings used to predict the outcomes were at the same time part of the outcome definition criteria of this study [28]. As previously shown, preexisting cardiovascular disease is a risk factor for adverse outcomes in COVID-19 [24] and, COVID-19 simultaneously affects the cardiovascular system [29]. However, the aforementioned approaches did not include quantitative measurements of cardiovascular CT metrics. This study included cardiovascular metrics, such as TPV, as an estimate of heart size. Indeed, a higher TPV was associated with a higher risk of intensified patient management. As age and sex did not differ significantly between groups, differences were caused probably by increased heart size or increased amount of pericardial fat. The AUCs of the models considering CT-derived metrics only vs. laboratory parameters only were both high and did not differ statistically significantly. This is probably due to the fact that both reflect inflammation of the lungs and are highly correlated. Internal validation indicated good internal generalizability, as did the external data for the PD model.

This study had several limitations. First, the internal validation dataset was small, resulting in wide CIs; therefore, the results should be interpreted cautiously. However, high standardization of chest CT and the fact that other studies on the topic reported similar effect sizes provide confidence that the results are generalizable. Second, while the main investigator site had access to the algorithms with all pulmonary and cardiovascular metrics, the remote site had access to pulmonary metrics only. Third, this study included only patients with COVID-19 who underwent a CT scan, the diagnostic standard for

patients with suspected pulmonary manifestations of COVID-19 at our center. Therefore, the presented approach might be less relevant in medical centers that rarely perform chest CT in this context. Fourth, other features, such as the initial severity of symptoms, might be useful to classify patient management. Besides the focus on automatically retrieved CT metrics, this study also considered demographic parameters and laboratory findings.

To conclude, this study provides evidence that chest CT of patients with COVID-19 contains valuable information for the prediction of ultimate patient management. Furthermore, this information is accessible using a deep learning-based, fully automated image analysis workflow, which is especially helpful during the COVID-19 pandemic.

Acknowledgments

We want to thank Ullaskrishnan Poikavilla from Siemens Healthineers, USA, for installing the algorithm prototype at our medical center. Additionally, we appreciate the great support of our research team, namely Rita Achermann, Ivan Nesic, Joshy Cyriac, and Bram Stieltjes.

References

1. Johns Hopkins University. COVID-19 Map Johns Hopkins Coronavirus Resource Center. Coronavirus.jhu.edu Web site. [Accessed May 26, 2020]. https://coronavirus.jhu.edu/map.html.
2. Epidemiology Working Group for NCIP Epidemic Response, Chinese Center for Disease Control and Prevention. The epidemiological characteristics of an outbreak of 2019 novel coronavirus diseases (COVID-19) in China. Zhonghua Liu Xing Bing Xue Za Zhi. 2020;41:145–151. [PubMed] [Google Scholar]
3. Verity R, Okell LC, Dorigatti I, Winskill P, Whittaker C, Imai N, et al. Estimates of the severity of coronavirus disease 2019: a model-based analysis. Lancet Infect Dis. 2020;20:669–677. [PMC free article][PubMed] [Google Scholar]
4. Mehta P, McAuley DF, Brown M, Sanchez E, Tattersall RS, Manson JJ HLH Across Speciality Collaboration, UK. COVID-19: consider cytokine storm syndromes and immunosuppression. Lancet. 2020;395:1033–1034. [PMC free article] [PubMed] [Google Scholar]
5. American College of Radiology. ACR recommendations for the use of chest radiography and computed tomography (CT) for suspected COVID-19 infection. Acr.org Web site. Published 2020. [Accessed August 6, 2020]. https://www.acr.org/Advocacy-and-Economics/ACR-Position-Statements/Recommendations-for-Chest-Radiography-and-CT-for-Suspected-COVID19-Infection.
6. Hwang EJ, Kim H, Yoon SH, Goo JM, Park CM. Implementation of a deep learning-based computer-aided detection system for the interpretation of chest radiographs in patients suspected for COVID-19. Korean J Radiol. 2020;21:1150–1160. [PMC free article] [PubMed] [Google Scholar]
7. Sun D, Li X, Guo D, Wu L, Chen T, Fang Z, et al. CT quantitative analysis and its relationship with clinical features for assessing the severity of patients with COVID-19. Korean J Radiol. 2020;21:859–868.[PMC free article] [PubMed] [Google Scholar]
8. Li K, Fang Y, Li W, Pan C, Qin P, Zhong Y, et al. CT image visual quantitative evaluation and clinical classification of coronavirus disease (COVID-19) Eur Radiol. 2020;30:4407–4416. [PMC free article][PubMed] [Google Scholar]
9. Zhang R, Ouyang H, Fu L, Wang S, Han J, Huang K, et al. CT features of SARS-CoV-2 pneumonia according to clinical presentation: a retrospective analysis of 120 consecutive patients from Wuhan city. Eur Radiol. 2020;30:4417–4426. [PMC free article] [PubMed] [Google Scholar]
10. Liu Z, Jin C, Wu CC, Liang T, Zhao H, Wang Y, et al. Association between initial chest CT or clinical features and clinical course in patients with coronavirus disease 2019 pneumonia. Korean J Radiol. 2020;21:736–745. [PMC free article] [PubMed] [Google Scholar]
11. Wan S, Li M, Ye Z, Yang C, Cai Q, Duan S, et al. CT manifestations and clinical characteristics of 1115 patients with coronavirus disease 2019 (COVID-19): a systematic review and meta-analysis. Acad Radiol. 2020;27:910–921. [PMC free article] [PubMed] [Google Scholar]
12. Li M, Lei P, Zeng B, Li Z, Yu P, Fan B, et al. Coronavirus disease (COVID-19): spectrum of CT findings and temporal progression of the disease. Acad Radiol. 2020;27:603–608. [PMC free article][PubMed] [Google Scholar]
13. Zheng Y, Xiao A, Yu X, Zhao Y, Lu Y, Li X, et al. Development and validation of a prognostic nomogram based on clinical and CT features for adverse outcome prediction in patients with COVID-19. Korean J Radiol. 2020;21:1007–1017. [PMC free article] [PubMed] [Google Scholar]
14. Yin X, Min X, Nan Y, Feng Z, Li B, Cai W, et al. Assessment of the severity of coronavirus disease: quantitative computed tomography parameters versus semiquantitative visual score. Korean J Radiol. 2020;21:998–1006. [PMC free article] [PubMed] [Google Scholar]
15. Park B, Park J, Lim JK, Shin KM, Lee J, Seo H, et al. Prognostic implication of volumetric quantitative ct analysis in patients with COVID-19: a multicenter study in Daegu,

Korea. Korean J Radiol. 2020;21:1256–1264. [PMC free article] [PubMed] [Google Scholar]
16. Driggin E, Madhavan MV, Bikdeli B, Chuich T, Laracy J, Biondi-Zoccai G, et al. Cardiovascular considerations for patients, health care workers, and health systems during the COVID-19 pandemic. J Am Coll Cardiol. 2020;75:2352–2371. [PMC free article] [PubMed] [Google Scholar]
17. Chaganti S, Grenier P, Balachandran A, Chabin G, Cohen S, Flohr T, et al. Automated quantification of CT patterns associated with COVID-19 from chest CT. Radiology: Artificial Intelligence. 2020;2:e200048. [PMC free article] [PubMed] [Google Scholar]
18. Ali A, Balachandran A, Vishwanath RS, Barthur A, Wichmann JL, Cimen S, et al. Evaluation of a deep learning based aortic diameter quantification system against multi-reader consensus measurement; Proceedings 2020 European Congress of Radiology (ECR); 2020 Jul 15–19; Vienna, Austria. European Society of Radiology: [Google Scholar]
19. Martin SS, van Assen M, Rapaka S, Hudson HT, Jr, Fischer AM, Varga-Szemes A, et al. Evaluation of a deep learning-based automated CT coronary artery calcium scoring algorithm. JACC Cardiovasc Imaging. 2020;13:524–526. [PubMed] [Google Scholar]
20. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988;44:837–845.[PubMed] [Google Scholar]
21. Tan C, Huang Y, Shi F, Tan K, Ma Q, Chen Y, et al. C-reactive protein correlates with computed tomographic findings and predicts severe COVID-19 early. J Med Virol. 2020;92:856–862.[PMC free article] [PubMed] [Google Scholar]
22. Lyu P, Liu X, Zhang R, Shi L, Gao J. The performance of chest CT in evaluating the clinical severity of COVID-19 pneumonia: identifying critical cases based on CT characteristics. Invest Radiol. 2020;55:412–421. [PMC free article] [PubMed] [Google Scholar]
23. Francone M, Iafrate F, Masci GM, Coco S, Cilia F, Manganaro L, et al. Chest CT score in COVID-19 patients: correlation with disease severity and short-term prognosis. Eur Radiol. 2020;30:6808–6817.[PMC free article] [PubMed] [Google Scholar]
24. Li B, Yang J, Zhao F, Zhi L, Wang X, Liu L, et al. Prevalence and impact of cardiovascular metabolic diseases on COVID-19 in China. Clin Res Cardiol. 2020;109:531–538. [PMC free article] [PubMed] [Google Scholar]
25. Ruch Y, Kaeuffer C, Ohana M, Labani A, Fabacher T, Bilbault P, et al. CT lung lesions as predictors of early death or ICU admission in COVID-19 patients. Clin Microbiol Infect. 2020;26:1417.e5–1417.e8.[PMC free article] [PubMed] [Google Scholar]
26. Kim YC, Chung Y, Choe YH. Automatic localization of anatomical landmarks in cardiac MR perfusion using random forests. Biomed Signal Proces. 2017;38:370–378. [Google Scholar]
27. Huang L, Han R, Ai T, Yu P, Kang H, Tao Q, et al. Serial quantitative chest CT assessment of COVID-19: a deep learning approach. Radiology: Cardiothoracic Imaging. 2020;2:e200075. [PMC free article][PubMed] [Google Scholar]
28. National Health Commission & National Administration of Traditional Chinese Medicine. Diagnosis and treatment protocol for novel coronavirus pneumonia. Chin Med J (Engl) 2020;133:1087–1095.[PMC free article] [PubMed] [Google Scholar]
29. Zheng YY, Ma YT, Zhang JY, Xie X. COVID-19 and the cardiovascular system. Nat Rev Cardiol. 2020;17:259–260. [PMC free article] [PubMed] [Google Scholar]

CHAPTER 7

Machine Learning in Cardiovascular Radiology: ESCR Position Statement on Design Requirements, Quality Assessment, Current Applications, Opportunities, and Challenges

BASED ON: Weikert T, Francone M, Abbara S, Baessler B, Choi BW, Gutberlet M, Hecht EM, Loewe C, Mousseaux E, Natale L, Nikolaou K, Ordovas KG, Peebles C, Prieto C, Salgado R, Velthuis B, Vliegenhart R, Bremerich J, Leiner T. Machine learning in cardiovascular radiology: ESCR position statement on design requirements, quality assessment, current applications, opportunities, and challenges. Eur Radiol. 2021 Jun;31(6):3909-3922. doi: 10.1007/s00330-020-07417-0.

Abstract

Machine learning offers great opportunities to streamline and improve clinical care from the perspective of cardiac imagers, patients, and the industry and is a very active scientific research field. In light of these advances, the European Society of Cardiovascular Radiology (ESCR), a non-profit medical society dedicated to advancing cardiovascular radiology, has assembled a position statement regarding the use of machine learning (ML) in cardiovascular imaging. The purpose of this statement is to provide guidance on requirements for successful development and implementation of ML applications in cardiovascular imaging. In particular, recommendations on how to adequately design ML studies and how to report and interpret their results are provided. Finally, we identify opportunities and challenges ahead. While the focus of this position statement is ML development in cardiovascular imaging, most considerations are relevant to ML in radiology in general.

Abbreviations

AI:	Artificial intelligence
ASCI:	Asian Society of Cardiac Imaging
AUC:	Area under the curve
CONSORT:	Consolidated Standards of Reporting Trials
CPU:	Central processing unit
CT:	Computed tomography
DCNN:	Deep convolutional neural network
DL:	Deep learning
DSC:	Dice similarity coefficient
ECG:	Electrocardiography
ESCR:	European Society of Cardiovascular Radiology
EUSOMII:	European Society of Medical Imaging Informatics
FDA:	Food and Drug Administration
FFR:	Fractional flow reserve
FN:	False negative
FP:	False positive
GPU:	Graphics processing unit
HU:	Hounsfield unit
IoU:	Intersection-Over-Union
LV:	Left ventricular/left ventricle
LVEF:	Left ventricular ejection fraction
ML:	Machine learning
MRI:	Magnetic resonance imaging
NASCI:	North American Society for Cardiovascular Imaging
NIfTI:	Neuroimaging Informatics Technology Initiative
PACS:	Picture archiving and communication system
RIS:	Radiology information system
RVEF:	Right ventricular ejection fraction
SCMR:	Society of Cardiovascular Magnetic Resonance
SPIRIT:	Standard Protocol Items: Recommendations for Interventional Trials
SSCT:	Society of Cardiovascular Computed Tomography
STARD:	Standards for Reporting of Diagnostic Accuracy Studies

TN:	True negative
TP:	True positive
TRIPOD:	Transparent Reporting of Multivariable Prediction Model for Individual Prognosis or Diagnosis
US:	Ultrasound

Introduction

Artificial intelligence (AI), machine learning (ML), and deep learning (DL) are currently getting a lot of attention in the public arena and in science [1]. Their relation is hierarchically nested as shown in Fig. 1. AI is an umbrella term encompassing all techniques that mimic human intelligence, which have been studied and applied for decades [2], whereas ML describes a subset of AI algorithms that learn to map input parameters to output from training data (supervised ML) or identify previously undetected patterns (unsupervised ML). DL comprises a subset of ML algorithms that use multiple, connected calculation layers [3].

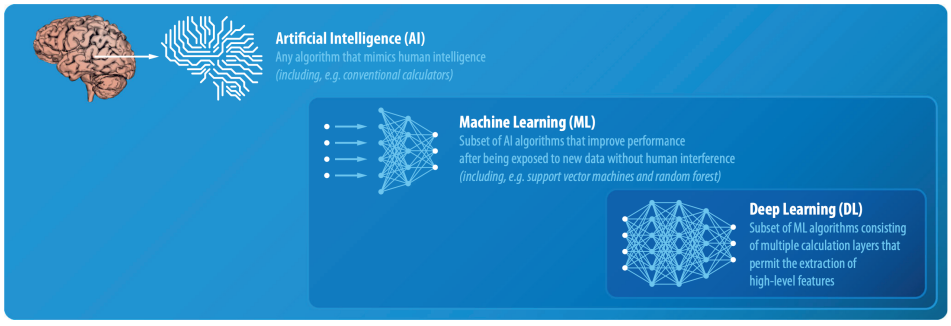


Fig. 1. Relation and definition of artificial intelligence (AI), machine learning (ML), and deep learning (DL)

Open-source programming tools, as well as greater computational power and easier data transfer facilitate the current boost in availability and productivity of AI algorithms. Concretely, in the most prevalent case of supervised ML, multiple pairs of input (e.g., MR image data of the heart) and output (e.g., ground truth segmentation of the left ventricle [LV]) are used for training. Subsequently, the trained algorithm can be used to automatically solve the learned task upon presentation of new, unseen input data. The fundamentals of ML have been described extensively elsewhere [4,5,6].

ML algorithms are of special interest to radiologists, because main areas of application are image processing, image analysis, and detection of findings — all core components of the radiological workflow before interpretation. One of their strengths is image segmentation, which is a prerequisite for analyses such as calculation of cardiac stroke volumes. Consequently, an increasing number of ML algorithms have been designed and evaluated in the field of cardiovascular radiology. Segmentation tasks are predominantly solved with

DL algorithms, which have shown performances superior to traditional image processing methods. However, applications of ML extend beyond image analysis and can support many other tasks within the field of radiology such as triage of exams according to urgency or provision of a second reading to avoid missing relevant findings. They can also help with predicting outcomes and extending the diagnostic capabilities of CT and MRI, e.g., by assessing the fractional flow reserve from cardiac CT angiography.

In light of these advances, the European Society of Cardiovascular Radiology (ESCR), a non-profit medical society dedicated to advancing cardiovascular radiology, has assembled a position statement regarding the use of ML in cardiovascular imaging in close cooperation with other leading societies in the field. While the focus of this position statement is ML development in cardiovascular imaging, most considerations are relevant to ML in radiology in general.

Requirements for successful development and implementation of ML algorithms in radiology

1. Human resources and expertise

Consensus statement

•Machine learning projects in cardiovascular imaging should involve experts with different professional backgrounds, mainly medical and ML experts, and in later stages also experts in user interface design and regulatory matters.

•The research and business community should agree on common data format standards and easy export of segmentation masks from clinically used post-processing software is needed to foster data interchangeability and reusability of data.

•Integration of ML algorithms into existing clinical workflows should be smooth, preferably into primary systems (RIS/PACS), to assure utility and acceptance by users.

A successful ML project in radiology is almost always multidisciplinary. It can be seen as a four-step process, each step requiring special expertise and ideally a close cooperation between multiple professionals, mainly medical and ML experts. This section describes human resources required for a successful ML project as well as key activities at each of the four steps.

The initial, crucial question is what problem to solve. For this, one needs clinical domain knowledge to identify a relevant problem. On the other hand, knowledge in programming is needed to assess whether the identified clinical problem is technically solvable by means of ML. Furthermore, one should always ask whether ML is the best solution for the problem or if there are other, less complex solutions, such as a manual workflow in the case of rarely occurring tasks. Apart from ML, there are many traditional imaging processing techniques like region growing that are very effective for e.g., tracing the coronary arteries [7]. To clarify these issues, a close cooperation between clinical experts and computer scientists is necessary. Furthermore, patients' interests should be considered at this stage.

Once a relevant clinical problem best solved by means of ML is identified, the second step is algorithm development. In the predominant case of *supervised learning*, this starts with data selection and establishment of a ground truth. Obtaining a sufficient amount of high-quality ground truth data, which can be thought of as "gold standard" used for both ML training and evaluation (e.g., segmentation masks of the left ventricle) is a necessary requirement for creating an algorithm. The amount of data needed depends on the complexity of the problem, the ML algorithm used, and the ratio between the finding of interest and the whole dataset. As a rule of thumb and providing an example from the field of object detection, tasks that are easy to solve for a human reader (e.g., detection and segmentation of a healthy lung within a chest CT scan) will require less training data than the detection of subtle, small changes in a whole-body CT scan. The less training data is available, the better the quality of the data should be. Large training data sets allow for some inaccuracies. It is important to keep in mind that training data quality is a limiting factor for an algorithm's performance. Therefore, the required data quality also depends on the envisaged performance of the algorithm. The decision how to obtain ground truth again requires both clinical and technical expertise. It revolves around questions such as how data can be extracted from hospital information systems, whether or not to use public databases, and how and by how many experts the ground truth should be established. Whenever patient data leaves primary clinical systems, it is of upmost importance to ensure complete de-identification. In radiology, this includes erasing or overwriting all DICOM tags that contain data privacy relevant information. It is highly recommended to double-check the success of this de-identification process by reviewing DICOM metadata before sending image data. An approach that allows for collaborative

training of an ML model without exchanging data samples is "Federated learning" [8]: a copy of an algorithm is downloaded and local data used to further improve it. The resulting changes to the model are summarized in an update that is uploaded and merged with the central consensus model. Preferably, ground truth data is stored in interchangeable formats (e.g., the Neuroimaging Informatics Technology Initiative (NIfTI) data format for segmentations) to assure usability for other projects. This is also in accordance with the FAIR guiding principles for scientific data management and stewardship [9]. Unfortunately, so far, no standards have been established for ML and most clinically used post-processing software is not capable of exporting segmentation masks, which limits the reusability of data and impedes reproducibility of studies. There are many other types of ground truth labels, comprising labels on the level of a whole dataset (fracture on radiograph: yes/no) and outcome labels for prediction modeling (patient death: yes/no).

In cardiovascular radiology, most algorithms solve segmentation tasks; therefore, segmentation masks are the predominant type of ground truth in this field. However, also tissue characterization (e.g., T1 and T2 mapping, late gadolinium enhancement) is an increasing part of cardiovascular radiology and needs adequate ground truth labels, e.g., histological results from endomyocardial biopsies in myocarditis. Awareness for potential biases introduced by the composition of datasets is important: an algorithm for clinical outcome prediction developed on a training dataset containing 80% males from country A might not work well on data from female patients in country B. In general, the dataset on which an algorithm is developed should reflect the population on which the algorithm is later applied as good as possible. Either one is aware of these limitations or overcomes the challenge by using large, heterogeneous datasets. The gold standard is to evaluate an algorithm's performance and influence on clinical workflows in clinical practice. Finally, a suitable ML technique has to be selected (e.g., random forest or deep convolutional neural networks). This task demands the expertise of the ML expert. This, as other steps, also involves ethical considerations, e.g., whether it is legitimate to develop algorithms on data of highly developed countries only (resulting in better performance in these patient collectives). For a detailed discussion of ethical implications of the use of ML in radiology, we refer to a recently published multi-society statement [10].

The third step is performance evaluation. There is a wide range of statistical tests that can be used, beginning with simpler concepts like sensitivity and specificity for detection tasks

to more complex evaluations like the Dice similarity score that ranges from 0 to 1 and quantifies the overlap of two regions of interest [11]. The evaluation method should be defined in advance to avoid method selection bias and involvement of a statistician is highly recommended. It is the responsibility of the radiologist to make sure the selected evaluation method reflects the clinically relevant endpoint. This demands functionality to visually check the validity of the data. Furthermore, the evaluation has to reflect the intended clinical use in the specific patient population the algorithm was designed for. It is also important to consider multicenter testing on different scanner models and patient populations should the algorithm later be used at other clinical centers and in different patient populations.

Finally, if assessed as an effective solution to the clinical problem, translation into clinical practice follows. This last step is at least as challenging as all previous steps and requires expertise in fields that are rarely covered by medical and ML experts, namely in user interface design, graphic design, regulatory matters, and in assuring compatibility with existing hospital IT environments that are subject to changes over time and location. While the creation of a dedicated software package is the most common option, the gold standard is the direct integration of an algorithm and its output into existing systems, preferably into Radiology Information Systems/Picture Archiving and Communication (RIS/PACS) systems and radiology reports. This results in a smooth workflow, thereby ensuring acceptance and engagement by users.

In our experience, the best results come from a close cooperation between experts from different disciplines. Figure 2 summarizes the four-step process.

2. Hardware and software requirements

Consensus statement

•While some less computationally intensive ML applications can be run on central processing units (CPUs), most currently applied ML algorithms require hardware with dedicated graphics processing units (GPUs).

•Experts involved in ML development should make use of online resources for creating, sharing, and discussing ML algorithms.

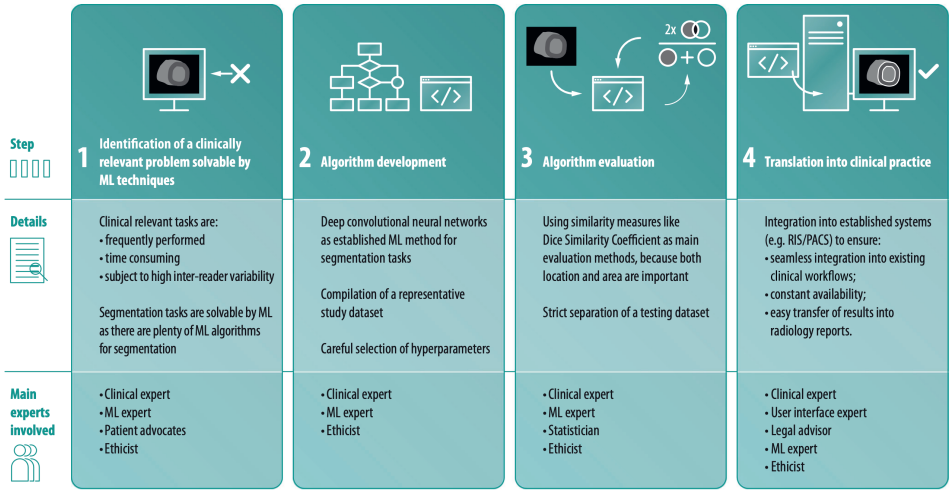


Fig. 2. Expertise needed during ML algorithm development and implementation, illustrated with the example of a segmentation task.

Besides human expertise, ML projects have requirements with regard to hardware and software.

Hardware

Standard CPUs are sufficient to run most non-DL ML algorithms and even DL approaches like deep convolutional neural networks (DCNNs) with few layers. However, DCNNs with multiple layers, which constitute the majority of currently developed ML algorithms of interest for cardiovascular imaging, are more computationally intensive. These algorithms need dedicated hardware with GPUs. Commercially available consumer GPUs with 8 GB or more system memory currently suffice for many applications. A detailed overview on hardware for ML, its performance, and pricing is provided by Tim Dettmers [12]. Alternatively, data can be processed using off-site cloud solutions such as Amazon Web Services.

Software

The ML community is fully digital and publishes mostly open source. Practically all relevant resources like software libraries and discussion forums are freely accessible. Jupyter Notebook is a commonly chosen web-based platform to compile ML algorithms (jupyter.org). The platform allows the use of multiple programming languages, including Python, which currently is the most prevalent language in the field of ML (python.org). A programming language can be thought of as the vocabulary and rule system that is used

to instruct a computer to perform tasks. ML algorithms can be developed using software libraries like TensorFlow (tensorflow.org), scikit-learn (scikit-learn.org), and PyTorch (pytorch.org). These libraries contain pre-written code and procedures that enable easier and faster software code development. Other alternatives are MATLAB (mathworks.com) and R (r-project.org). Once the code is created, it should preferably be shared publicly. GitHub is a common online Git repository for sharing and discussing software code with version control function that allows to retrace a project's source code history (github.com). Furthermore, anonymization tools are important for ML projects in radiology, because sensitive patient information is part of the DICOM header of each image and data exchange is needed to build large databases with studies from multiple centers. Fortunately, there are numerous free stand-alone tools with batch processing function for Mac OS (e.g., dicomanonymizer.com) and Windows (e.g., rubomedical.com/dicom_anonymizer). The RSNA's Clinical Trials Processor (CTP) is open-source software that covers the whole image transfer pipeline between data acquisition sites and a principal investigator site with build-in anonymization capability (mirc.rsna.org). Figure 3 provides an overview of useful software and online resources.

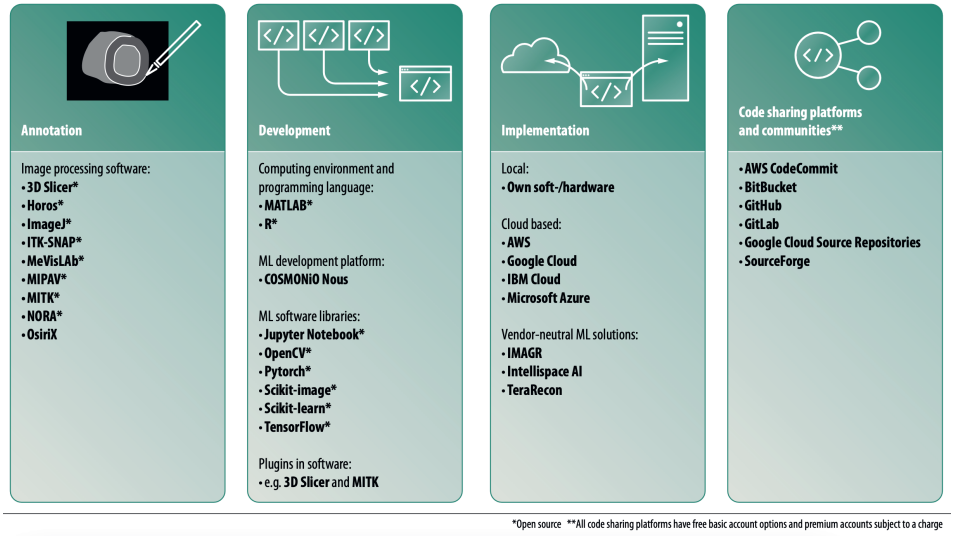


Fig. 3. Useful software at different stages of a ML project in radiology

Recommendations regarding study design and reporting

Consensus statement

•Based on existing study quality standard frameworks such as SPIRIT and STARD, we propose a list of quality criteria for ML studies in radiology.

ML studies should be held to the same quality standards as any other diagnostic or prognostic study. Several frameworks exist that define standard protocol items for clinical trials as well as for reporting the results of diagnostic and prognostic studies. Clinical trial protocols should conform to the Standard Protocol Items: Recommendations for Interventional Trials (SPIRIT) checklist [13]. Diagnostic accuracy studies to the Standards for Reporting of Diagnostic Accuracy Studies (STARD) requirements and, at a minimum, should report essential items listed in the 2015 version of the STARD checklist [14]. For prognostic studies, the Transparent Reporting of Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) guideline and checklist [15] should be followed. Although these guidelines were not designed with ML studies in mind, they do form a solid basis for providing the details of a ML study in a protocol (SPIRIT), and for reporting results of studies in which ML has been applied (STARD and TRIPOD). Because these guidelines have not been taken up widely in the ML community, efforts are underway to develop ML-specific versions of each of these frameworks. In the meanwhile, we attempt to provide guidance by offering a checklist of items for researchers designing ML studies and for readers assessing the quality of publications. Our efforts expand upon the recently published editorial by Bluemke et al, which also addresses this topic [16].

Recommended items for designing and reporting ML studies

In the following section, we provide a list of important considerations when designing and reading studies that employ ML. We have summarized these considerations in a checklist (Table 1) and apply them to a research article that aimed to design a DL algorithm for automatic cardiac chamber segmentation and quantification of left ventricular ejection fraction (LVEF; [17]).

Table 1 Checklist of items to include when reporting ML studies

1. Which clinical problem is being solved?
   - □ Which patients or disease does the study concern?
   - □ How can ML improve upon existing diagnostic or prognostic approaches?
   - □ What stage of diagnostic pathway is investigated?
2. Choice of ML model
   - □ Which ML model is used?
   - □ Which measures are taken to avoid overfitting?
3. Sample size motivation
   - □ Is the sample size clearly motivated?
   - □ Which considerations were used to prespecify a sample size?
   - □ Is there a statistical analysis plan?
4. Specification of study design and training, validation, and testing datasets
   - □ Is the study prospective or retrospective?
   - □ What were the inclusion and exclusion criteria?
   - □ How many patients were included for training, validation, and testing?
   - □ Was the test dataset kept separate from the training and validation datasets?
   - □ Was an external dataset used for validation?*
   - □ Who performed external validation?
5. Standard of reference
   - □ What was the standard of reference?
   - □ Were existing labels used, or were labels newly created for the study?
   - □ How many observers contributed to the standard of reference?
   - □ Were observers blinded to the output of the ML algorithm and to labels of other observers?
6. Reporting of results
   - □ Which measures are used to report diagnostic or prognostic accuracy?
   - □ Which other measures are used to express agreement between the ML algorithm and the standard of reference?
   - □ Are contingency tables given?
   - □ Are confidence estimates given?
7. Are the results explainable?
   - □ Is it clear how the ML algorithm came to a specific classification or recommendation?
   - □ Which strategies were used to investigate the algorithm's internal logic?
8. Can the results be applied in a clinical setting?
   - □ Is the dataset representative of the clinical setting in which the model will be applied?
   - □ What are significant sources of bias?
   - □ For which patients can it be used clinically?
   - □ Can the results be implemented at the point of care?
9. Is the performance reproducible and generalizable?
   - □ Has reproducibility been studied?
   - □ Has the ML algorithm been validated externally?
   - □ Which sources of variation have been studied?
10. Is there any evidence that the model has an effect on patient outcomes?
    - □ Has an effect on patient outcomes been demonstrated?
11. Is the code available?
    - □ Is the software code available? Where is it stored?
    - □ Is the fully trained ML model available or should the algorithm be retrained with new data?
    - □ Is there a mechanism to study the algorithms' results over time?

Application of the checklist to the research article "Automated cardiovascular magnetic resonance imaging analysis with fully convolutional networks" by Bai et al [17].

1. Which clinical problem is being solved?

A clear description of the clinical problem and rationale for the study should be provided, taking into account existing approaches and how they fall short. This includes the specification of the disease in question and a clear description of the subjects or patients studied. It is also important to hypothesize how ML approaches may improve upon existing approaches such as conventional statistical approaches to solve the problem. Other relevant questions include the stage of the disease in question and place in the diagnostic pathway.

2. Choice of ML model

The choice of ML model should be clearly motivated since there is a wide variety of approaches, which may result in different results. It is also important to explicitly discuss overfitting and approaches used to mitigate this problem. Overfitting occurs when ML models are trained to predict training data too well, which results in the inability to generalize to new, unseen data. An overview of commonly used ML models and their characteristics as well as approaches that can be used to deal with overfitting is provided by Liu et al in their review article [18]. Technical details of the algorithm including hyperparameters should be specified to foster transparency and replicability.

3. Sample size motivation

In contrast to the recommendations made in the STARD and CONSORT guidelines, most ML studies have not explicitly considered sample size when designing the study and are often based on convenience samples. However, sample size and a statistical analysis plan should ideally be prespecified. Although there are presently no clear guidelines on how to calculate a sample size in ML studies, the number of subjects or datasets can be prespecified according to considerations such as the minimal clinical difference of interest or the expectation that ML is able to generate equivalent results to human observers on a certain task. Furthermore, sample sizes used by other researchers to solve comparable problems might be a good indicator.

4. Specification of study design and training, validation, and testing datasets

Algorithm development demands data for training, validation, and testing. Investigators should specify how the data was split into each of these categories. It is of utmost importance to strictly separate the testing dataset from the other datasets to obtain a realistic estimate of model performance. This is also a requirement for regulatory approval of ML-based computer-assisted detection devices from the United States Food and Drug Administration (FDA) [19]. Ideally, validation is performed not only on internal data (from the same department or institute) but also on an external dataset by independent researchers.

5. Standard of reference

A key consideration in ML studies is selection and quality of the reference standard or ground truth. Researchers should precisely specify how and by whom ground truth data were labeled, including the level of experience of each observer. It is important to take into account interobserver variability between experts and to describe how disagreements are resolved (e.g., by demanding that observers reach a consensus, or by adjudicating any differences by a separate observer). It should be noted whether existing labels were used (e.g., from radiology reports or electronic health records), or new labels were created. Finally, experts labeling the data should ideally work independently from each other because this will facilitate measurement of interobserver agreement between human experts.

6. Reporting of results

Analogous to conventional diagnostic studies, contingency tables with the number of true positive, true negative, false positive, and false-negative classifications should be given at the prespecified chosen classifier threshold. Other useful measures include the area under the receiver operating curve (AUC) and Bland-Altman plots [20]. It is important to note that terminology in ML studies may be different from the terminology used in the medical literature. Sensitivity is equivalent to "recall" and "precision" denotes positive predictive value. The F1 score is a compound measure of precision and recall and its use is therefore highly recommended. Table 3 summarizes measures frequently used in ML. Confidence intervals should be reported for all of these measures. In image segmentation and analysis tasks, measures of how well the ML algorithm performs compared to the standard of reference should be given. These typically include the Dice coefficient (a measure of how well the ML generated contours overlap with the standard of reference contours), the mean contour distance (the mean distance between two segmentation contours), and the Hausdorff distance (the maximum distance between the 2 segmentation contours) [11].

7. Are the results explainable?

Because of the large number of parameters involved, interpreting the results of ML studies can be challenging, especially when working with DL algorithms. This consideration is particularly pertinent when important treatment decisions are contingent upon the results generated by the algorithm. Saliency mapping enables the identification of morphological features in the input image underlying the model's prediction and can help to investigate the algorithm's internal logic. Visual feedback about the model's predictions is very important to understand whether networks learn patterns agreeing with accepted pathophysiological features or biologically unknown, potentially irrelevant features.

8. Can the results be applied in a clinical setting?

Machine learning studies designed to solve a specific clinical problem should explicitly consider whether the results apply to a real-world clinical setting. This includes discussion of how representative the dataset used for derivation and testing of the model is of the clinical setting in which it will be applied. Any sources of bias, in particular class imbalance and spectrum bias, should be identified and discussed. Considering these factors can enable more precise identification of patients in which the algorithm can be used clinically, or in which groups of patients and clinical scenarios additional validation is needed. Investigators should also consider if and how the algorithm can be used at the

point of care, including issues like availability of the algorithm (e.g., on-premise or via cloud solutions), how fast results are available (e.g., in real-time or with a delay), and how results are visualized in order to check the model's predictions.

9. Is performance reproducible and generalizable?

To date, in most reports on ML, model development, tuning, and testing have been performed on a convenience sample of locally available data. Although many of these reports have demonstrated encouraging results, it is important to investigate the reproducibility of the results and to perform an external validation, preferably on multiple datasets from other independent institutes and investigators. External validation is important to investigate the robustness of the model to e.g., differences in image acquisition and reconstruction methods between vendors and institutes and differences in referral patterns and variability in the prevalence of the condition of interest. Conversely, we also believe it is advisable to validate external algorithms prior to local use, especially if the algorithms` results are used for automated analysis with results directly transferred into clinical reports instead of use as a second reading tool.

10. Is there any evidence that the model has an effect on patient outcomes?

Although one of the first proofs of concept in the development of an ML algorithm is the investigation of its diagnostic accuracy, investigators and readers should ask themselves the question whether there is any evidence of an effect on patient outcomes. This is especially important for algorithms used for treatment recommendations and detection of unrequested findings. Ideally, this should be investigated in prospective, randomized clinical trials, as is the case for conventional interventions. These considerations also help to detect and mitigate reasons for missing impact of diagnostically well performing algorithms on patient outcomes, such as suboptimal communication of results.

11. Is the code available?

Transparency regarding an ML model's design and function is key to clinical acceptance. Making the computer code available to other investigators is a major step towards this goal and is increasingly becoming a condition for obtaining funding as well as acceptance of studies in high-quality, peer-reviewed journals. The GitHub platform facilitates free and rapid dissemination of software code with basic quality checks. Investigators should state whether the source code of their algorithm will be made available and under which conditions. If not, specific reasons should be given. Making the software code available

enables other researchers to independently investigate whether reported results can be reproduced and to improve model performance. Furthermore, it enables the evaluation of a model's performance over a prolonged period of time.

Table 3 Performance metrics frequently used in ML

Metric	Definition and details
Recall	Fraction of true positive (TP) instances among the instances predicted to be positive by an algorithm, including false positive (FP) instances (synonym for "positive predictive value") Recall = $\frac{TP}{TP+FP}$
Precision	Fraction of the instances predicted to be positive by an algorithm among all TP instances, including false negative (FN) instances (synonym for "sensitivity") Precision = $\frac{TP}{TP+FN}$
Accuracy	Fraction of TP and true negatively (TN) predicted instances among all instances. Accuracy = $\frac{TP+TN}{TP+TN+FP+FN}$
F_1-score	Harmonic mean of precision and recall. Ranges from 0 to 1 (meaning perfect precision and recall). Important measure, because both high precision and recall are needed for high F_1 scores. $F_1 = 2 * (\frac{precision*recall}{precision+recall})$
False-positive findings	Negative instances falsely predicted to be positive by an algorithm. Numbers of false-positive findings are very important in ML, because too many of them render algorithms useless. Investigating the reasons for false-positive findings may help to develop strategies to avoid them, but requires domain knowledge in the field of application.
ROC curve	Receiver operating characteristic curve. Graph illustrating the discriminative ability of a classifier. Sensitivity (*Y*-axis) plotted against the false-positive rate (*X*-axis) for different classification thresholds. The area under the curve (AUC) measures the 2D area underneath the ROC curve and provides an aggregate measure of performance.
Intersection-Over-Union (IoU)	Important measure to assess the performance of algorithms for segmentation tasks. Overlap between two regions of interest, mostly of a ground truth segmentation and a predicted segmentation, e.g., of the left ventricle. Ranges from 0 to 1, with 1 indicating perfect overlap. IoU = $\frac{Area\ of\ overlap}{Area\ of\ union}$.
Dice similarity coefficient (DSC)	Another important measure in assessing segmentation algorithms. Ranges from 0 to 1, with 1 indicating perfect overlap. DSC = $\frac{2*Area\ of\ overlap}{Total\ \ area\ of\ objects}$

Insights of a systematic literature review on applications of ML in cardiac radiology

To identify articles on the application of ML in cardiac radiology, a comprehensive search for articles in PubMed and EMBASE databases was conducted. The search identified all articles in the English language registered no later than 31.01.2020 ($n = 599$ in PubMed; $n = 2559$ in EMBASE). Supplement 1 documents the search strings. Figure 4 displays the exact search and review workflow that included the removal of duplicates ($n = 506$) with the auto-function of the literature management software (Mendeley) and the exclusion of articles that were not on ML in cardiac radiology by manual screening ($n = 2466$). In the next step, the remaining relevant articles ($n = 222$) were classified into five categories according to the function of the ML applications: (a) image acquisition and preprocessing, (b) detection, (c) segmentation, (d) diagnosis, (e) prediction, and (f) other. The relation of those categories is sequential; e.g., detection is a prerequisite for segmentation. The studies were attributed to the most advanced category according to the purpose of the given algorithm.

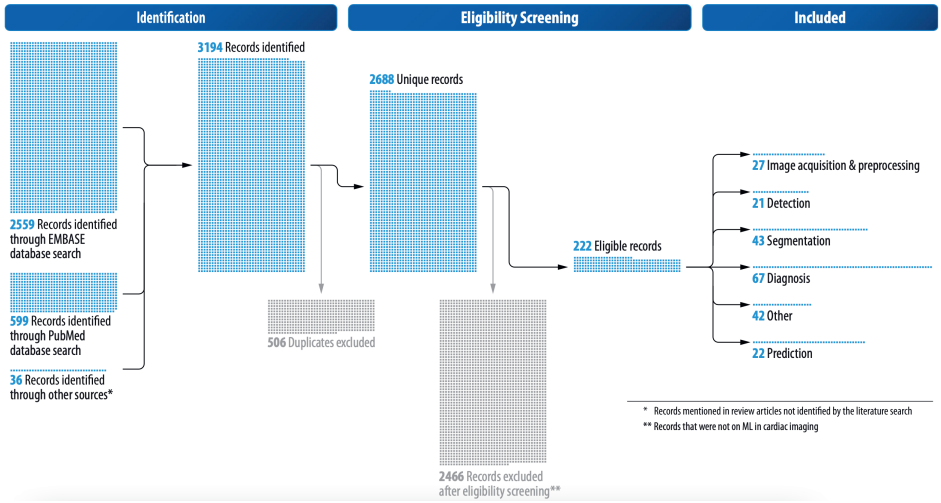


Fig. 4. Search and review flow diagram

At this point, we briefly mention an example per category; Fig. 5 presents corresponding images: (a) Tatsugami et al used a DCNN with 10 layers to reduce the image noise of CT angiography images. The mean image noise was significantly lower than that of images reconstructed with standard hybrid iterative reconstruction alone (18.5 ± 2.8 HU vs. 23.0 ± 4.6 HU) [21]. (b) Howard et al developed five neural networks on 1676 images to detect and identify cardiac pacemakers and defibrillators on chest radiographs. They report an accuracy of 99.6% and even classified specific model groups of the devices [22]. (c) Romaguera et al used a DCNN to segment the left ventricle in short-axis cardiac MRI images and found a Dice score of 0.92, a sensitivity of 0.92, and a specificity of 1.00 [23]. (d) Lessmann et al developed and tested DCNNs for an automated calcium scoring on 1744 non-ECG-gated CT scans without contrast. They report an F1 score of 0.89 for calcium scoring of coronaries on soft kernel reconstructions [24]. (e) Coenen et al used a neural network with four layers to predict the hemodynamic relevance of coronary artery stenoses from CTA data alone by using the ML-based FFR (fractional flow reserve) with invasively measured FFR as a standard of reference. They report an improved diagnostic accuracy of CTA-based assessment of stenosis from 71 to 85% (sensitivity: 89%; specificity 76%) [25].

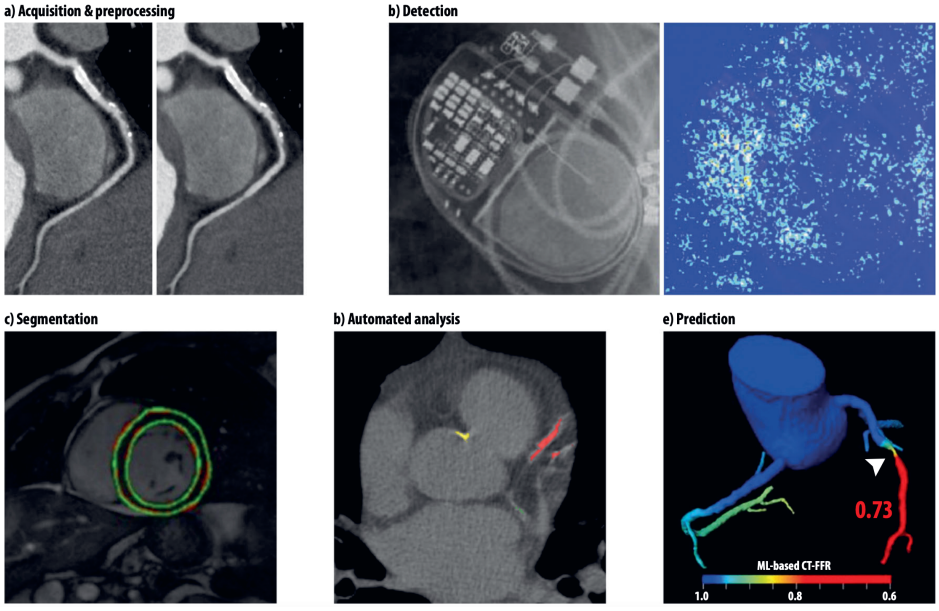


Fig. 5. Examples of application of ML in cardiac radiology. a Curved multiplanar reformation of CTAs with multiple plaques and a stent in the right coronary artery; standard hybrid iterative image reconstruction on the left, image processed with an ML algorithm with reduced noise on the right [21]. b Correctly identified Advisa device on a plain radiograph (left) with the according saliency map (right) that visualizes the neural networks attention [22]. c Segmentation of the LV on MRI by a DCNN with automatically detected contours in green colorC [23]. d Automated detection and quantification of calcifications on non-contrast CT scans (red: left anterior descending coronary artery; green: left circumflex coronary artery; yellow: thoracic aorta) [24]. e ML-based CT fractional flow reserve predicting obstructive stenosis in the mid left anterior descending coronary artery [25]

Figure 6 demonstrates the exponentially increasing number of publications on ML in cardiac radiology since 2013. Figure 7 shows the distribution of modalities and the ML techniques that were covered in the research articles, with MRI being the predominant modality (41.4%) and DL being the most frequently used ML technique (63.1%).

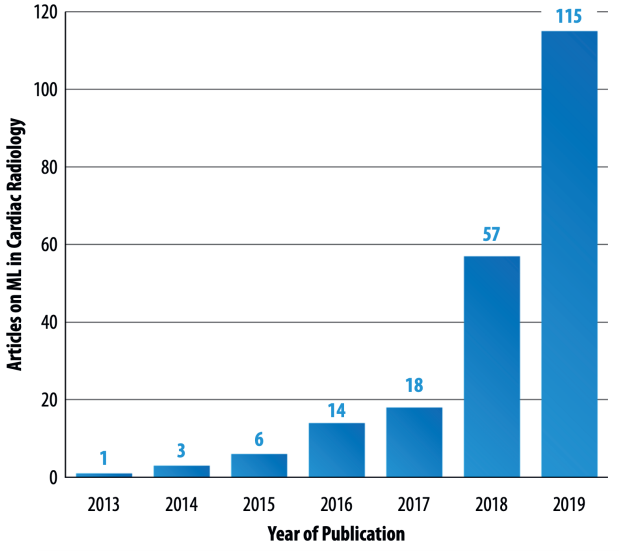


Fig. 6. Number of articles on ML in cardiac radiology (Y-axis) published per year between 2013 and 2019 (X-axis), as resulting from the structured literature review. Studies published earlier than 2013 and in 01/2020 are not included for reasons of clarity

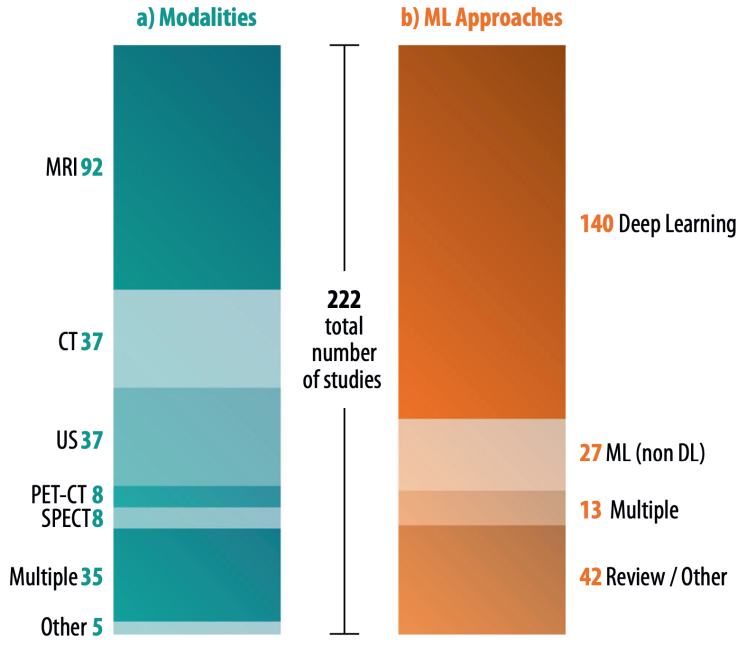


Fig. 7. (a) Modalities involved and (b) types of ML approaches used in studies on ML in cardiac radiology (n = 222)

Supplement 2 provides a detailed literature review on ML in cardiac radiology highlighting and discussing important studies in all categories. Supplement 3 contains the complete reference list of all studies resulting from the literature search and a table with detailed information on the studies.

ML in cardiovascular radiology: opportunities and challenges

1. Opportunities

Consensus statement

•ML algorithms provide opportunities along the whole task-pipeline of cardiovascular radiology.

Machine learning offers great opportunities in cardiovascular imaging from the perspective of multiple stakeholders. First and foremost, from a patient perspective, there are opportunities to avoid unnecessary imaging. Should imaging be deemed necessary, ML offers opportunities to do so with shorter imaging protocols and lower radiation doses. Once the images are acquired, automated post-processing facilitated by ML offers quicker results compared to current workflows and could reduce interobserver variation and aid in prioritizing studies with urgent findings. ML can also be used to extract additional relevant information from images. Some examples relevant to cardiovascular imaging include extraction of volumes of *all* cardiac chambers instead of just the LV, more detailed analysis of cardiac motion patterns, and quantification of the amount of pericoronary and pericardial fat as well as the amount of liver fat. When reliable algorithms capable of assigning a diagnosis become available, this could reduce diagnostic error by serving as a "second reader." Finally, ML can aid in automatic extraction of unrequested but prognostically relevant information. For example, automated detection or exclusion of pulmonary nodules or abnormalities in other organ systems depicted in the field-of-view would be useful for radiologists specialized in cardiovascular imaging. Furthermore, ML algorithms can be used to create detailed local, national, and international databases with normal values for clinical comparison. This will also enable the detection of smaller effect sizes and more precise results.

2. Challenges

Consensus statement

•ML algorithms were initially developed to solve problems in non-medical domains. Due to peculiarities of health-related data like high inter-reader variability, dispersed data storage, and data privacy issues, ML projects in radiology are facing specific challenges.

•The medical research community should strive for the compilation of multicenter datasets that are currently lacking in the field of cardiovascular imaging.

•Further challenges encompass rare disease and/or anatomical variants and compliance with legal frameworks.

The compilation of high-quality datasets for ML projects in radiology is hampered by some peculiarities of health-related data. First, there is the issue of significant inter- and intrareader variability fostered by the fact that many categories in medicine are not as distinct as those of everyday objects such as dogs or cars. Second is the complexity of medical image interpretation. For example, a small hyperintense streak in late gadolinium enhancement imaging may be a hyperintensity artifact or a true scar. For clarification, one needs to integrate additional information such as whether there is an implanted cardiac device or not. Third, there is a lack of standardization in the acquisition of medical data. In radiology, heterogeneity is introduced by differing vendors of hardware, software, and unstandardized acquisition parameters. This is true for imaging data, and also for other diagnostic tests and therefore prevents "one-fits-all" solutions. The fourth challenge concerns the non-standardized format and dispersed nature of health data. While in other areas like engineering data is registered in interchangeable systems often designed from scratch, data in hospitals are mostly stored in dispersed, historically grown data silos in multiple data formats. The fifth challenge, especially relevant to cardiovascular radiology, is dealing with higher dimensional imaging data. For example, cardiac cine images are four-dimensional, while ML algorithms are traditionally designed to cope with data in two dimensions. Solutions to this challenge are either complex, or reduce information (processing of a 3D CT dataset as a series of multiple 2D images). Another important challenge to ML projects in radiology is strict standards of data privacy.

All this makes the creation of high-quality ground truth datasets in healthcare challenging and expensive. As a result, datasets in healthcare-related ML projects tend to be much smaller compared to the non-medical domain: the famous contest for everyday object

detection, the ImageNet challenge, encompasses over 14 million images (image-net.org/about-stats), while ML studies in cardiovascular radiology often comprise less than 100 cases. Public datasets like the ChestX-ray8 dataset provided by the NIH containing more than 100,000 frontal-view radiographs with eight disease labels are important initiatives to overcome this problem. However, labels need critical quality review, which requires medical domain knowledge. For cardiovascular radiology, comparable datasets are currently lacking and professional societies can play an important role in the assembly of large publicly available datasets with high-quality ground truth labels to allow for an objective comparison of different ML algorithms. Registries like the ESCR MR/CT registry (mrct-registry.org), providing large standardized data sets for further analyses with currently > 300,000 de-identified examinations [26], are also an important contribution in this direction.

Apart from image data–related problems, there are other challenges. First, there is the problem of rare disease entities. ML algorithms need a sufficient amount of training examples to detect patterns, ideally including examples at the extreme ends of the disease spectrum. However, many radiologic disease patterns are rarely seen, like congenital heart diseases, and ensuring a fully representative training dataset remains a difficulty. Second, legal issues: Machine learning algorithms, although highly accurate for many tasks, are never perfect and discussions on legal liability for incorrect or missed diagnoses are ongoing [27]. Third, the question of acceptance: will physicians and patients be willing to trust judgments of algorithms that are a "black box" to them? However, trust in systems not fully understandable to us is part of day-to-day life.

Conclusion

The number of scientific studies published on ML in cardiovascular imaging has been exponentially growing with more than 100 research articles in 2019. The majority concerned MRI studies using DCNNs for image segmentation tasks, but ML algorithms can also help to shorten imaging protocols and extract more information from the same imaging data. The prerequisites for ML to make important contributions in the field of radiology are now in place: freely available open-source software, vast amounts of digital radiology data in most countries, and an increasing presence of well-trained experts to train and clinically supervise ML. Furthermore, online transfer of data and ML models has become convenient.

However, to accomplish this enormous potential, the field of radiology needs to develop common quality standards regarding ML applications and studies. We highlight the need for a detailed description of datasets and methodology used. Furthermore, in the course of ML algorithm development that aims at having a clinical impact, cooperation of professionals from multiple backgrounds is required.

CHAPTER 8

Discussion, Conclusion, and Outlook

Discussion

Driven by recent technological advancements, a large number of ML applications has emerged in the field of radiology. While the main focus was on image recognition tasks in recent years, the spectrum of applications is much broader including text-recognition, speech-recognition, image reconstruction, quality control, patient scheduling, predictive analytics and many more. The thesis at hand combines six articles with a focus on cardiothoracic imaging that make a small contribution to the vast body of literature in this field. This chapter first summarizes key insights gained during the five projects, then discusses chances and challenges for ML in radiology with a focus on cardiothoracic imaging, and finally provides a conclusion and outlook.

Key insights

This section provides five key insights gained during the projects.

First, ML is the undisputed standard method for image recognition and segmentation tasks at this moment, and therefore of high interest to radiologists. While a combination with traditional image processing methods may improve the performance in specific cases, the "go-to" procedures for detection and segmentation of medical images are based on ML. To be more precise, on the ML-subcategory deep learning, which uses neural networks. For object detection, ResNets are a frequent choice [19]. The output can be visualized as bounding boxes around findings of interest. To each bounding box, a category (e.g., "lung tumor") and probability (e.g., "0.96") are assigned. Examples are found in **Chapter 3** (detection of pulmonary emboli) and **Chapter 4** (detection of rib fractures). For image segmentation, U-Net, a convolutional neural network [2], is state-of-the art and was used in **Chapter 6** to segment the aorta and define pericardial volumes. For organ segmentation in general, U-Nets are an excellent choice.

Second, the answer of whether ML will be part of the future of radiology can be answered with a clear "yes": ML is already highly productive in assisting radiologists. Examples are the algorithm for detection of pulmonary embolism discussed in **Chapter 3** and the algorithm for detection and segmentation of pulmonary nodules used in **Chapter 5**, which have become part of products running in hospitals around the world. On the other hand, it has also become clear by now, that the statement by ML-expert Geoffrey Hinton, who claimed at a conference in 2016 that "(...) people should stop training radiologists

now, it's just completely obvious that within five years deep learning is going to be better than radiologists" [20] was an exaggeration probably based on lacking knowledge of the complexity of medicine. To conclude, whereas we are far from complete automatization of the radiology workflow, multiple applications have reached a productive stage.

Third, implementation of ML applications in the clinic is at least as challenging as training the underlying model. Whereas the body of literature on ML in radiology is vast [21], only a tiny fraction finally makes it into clinic. Besides expert knowledge in radiology and computer science, many more competences are needed for successful implementation. This includes legal expertise to cope with the complex legal framework governing medical products and expertise in user interface design to ensure usability. Furthermore, the explicit will of all stakeholders is needed. Whereas there is usually a lot of initial enthusiasm towards new technologies, the concrete and sustained operation in a clinical environment is a challenge demanding perseverance and attribution of resources and responsibilities.

Fourth, the visual inspection of outputs and analysis of reason of failure should be part of ML studies in radiology, because they can guide the way to further algorithmic improvements. An example is provided in **Chapter 3**, which found that an algorithm for detection of pulmonary emboli frequently marked unremarkable pulmonary veins. A potential way to improve this algorithm is to additionally segment pulmonary veins and disregard all embolus candidates located within the vein vessel mask. Similarly, **Chapter 4** reports that a relevant number of FP detections of rib fractures are found outside the region of the body ("out of bounds") – in this case, an algorithm based on ML or traditional image processing techniques such as HU thresholding could improve performance by classifying voxels into body region vs. surroundings.

Fifth, the consequences of large amounts of FP findings caused by parallel application of multiple algorithms are currently being ignored. Sensitivity and the number of FP findings are a trade-off: the more an algorithm is tuned towards sensitivity, the more FP it produces and vice-versa. Whereas current algorithms, such as the ones presented in **Chapters 3 and 4**, reach high levels of sensitivity at only small numbers of FPs per case, they only solve one specific task each: detection of pulmonary embolism and rib fractures, respectively. Presumably, the use of ML in radiology will increase and each examination will be analyzed by multiple algorithms. An average of 20 false alerts per examination is

the consequence of running 100 algorithms for detection of various critical findings, assuming an excellent number of only 0.2 FPs per case per algorithm. Those would have to be checked by a radiologist and decisively inhibit the radiology workflow. Potential solutions include accepting lower sensitivities, adding further layers of analysis as mentioned above, or pre-checking algorithmic results by educated auxiliary staff.

Chances

Machine learning offers great chances to radiology in general. This is especially true for cardiovascular imaging with its time-consuming post-processing workflows and relatively radiation-intensive diagnostic methods such as cardiac CT. For example, algorithms can help to achieve lower radiation doses by optimizing image acquisition and reconstruction [22]. Automated post-processing can save immense amounts of time, especially in cardiac MRI and vascular imaging. At the same time, those automated analyses reduce interobserver variation and help to replace qualitative statements in radiology reports with quantitative imaging parameters. Another opportunity is that ML facilitates the use of all information available in an examination. For example, by delivering metrics on the amount of fat in the liver that is partially included in the field-of-view of a chest CT. This also helps not to miss findings that are not in the focus of the given examination, such as pulmonary nodules or osseous processes in a cardiac CT. Prioritization of examinations with suspected critical findings can optimize the allocation of resources in radiology. Furthermore, ML algorithms facilitate the retrospective analysis of whole imaging archives, thereby enabling the creation of databases with normal values on a population level.

Challenges

There are relevant challenges that ML projects in radiology have to deal with. One is the availability of high-quality data. ML models need large amounts of training data. Everyday objects such as cars or cats are distinct categories, images are available in large quantities, and labelling does not require specific education. This makes high-quality training data cheap. In contrast, in radiology, labels are less distinct (is this slight thickening of the pleura suspicious or not?), less data is available (also due to regulatory barriers protecting individual rights) and expertise in medical imaging is a prerequisite for labelling. All this makes high-quality training data expensive. Additionally, in rare disease, there is not enough data to reach satisfying performance. Furthermore, medical imaging data is heterogenous due to the variety of vendors, acquisition parameters, and software

providers. To complicate things even more, disease prevalence differs between regions, which hampers "one-size-fits-all" solutions. Another challenge is legal issues: even if modern ML algorithms reach high performance, they will never be perfect and debates about liability for mis-diagnoses are ongoing.

Conclusion and Outlook

This thesis provided examples of application of ML to solve problems in the field of cardiothoracic radiology. It was demonstrated that this is not confined to the core duty of radiology departments, image analysis, but includes the complete range of tasks, beginning from data acquisition and data curation to clinical application and predictive analytics. It was found that ML is the state-of-the art for image recognition tasks and that performances of current applications are at levels that allow for clinical application. At the same time, it was noticed that we are far from a complete auto-analysis of medical images and, also due to the complexity of the field, it is fair to predict that radiologists as domain experts will be needed in the future.

The most likely scenario for the mid-term future is the ML-enhanced radiology department, where algorithms perform repetitive tasks such image pre-processing, segmentation of organs, and provide volumetric information to support radiologists in their daily routine. Prepopulation of written radiology reports on the basis of algorithm findings has tremendous potential. It is then the job of the radiologist to monitor the algorithm outputs, adjust their settings where needed, and consolidate as well as interpret this information. The time that is saved by unburdening radiologists from repetitive tasks can be used to put more effort into timely and comprehensive communication to referring physicians and patients. Cardiothoracic radiology is likely to be among the sub-specializations of radiology that profit most, due to the considerable proportion of time-consuming and repetitive tasks such as segmentation of ventricles in cardiac MRI.

To make the most of ML, an open discussion among imaging professionals and computer scientists on where its application makes sense and where not is warranted. As a general rule of thumb, using ML to solve tasks that are repetitive, frequently performed, allow for a clear definition of ground-truth, and offer a lot of training data seems promising. This involves guideline-compliant diameter measurements of the aorta, cardiac volumetry, and

chest CT for follow-up of pulmonary nodules. Furthermore, in recent years, the focus of research was the development and evaluation of specific ML applications using retrospective data. This was reasonable at this early stage of development. In the future, clinical application and impact on clinical workflows as well as patient outcomes need to be addressed. For that, prospective controlled clinical trials are needed.

To conclude, ML has already arrived at the heart of cardiothoracic radiology. The reasonable integration of resulting applications into clinical workflows is among the central fields of activity in the years to come.

References

1. Ronneberger O, Fischer P, Brox T (2015) U-Net: Convolutional Networks for Biomedical Image Segmentation. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 9351:234–241
2. He K, Zhang X, Ren S, Sun J (2015) Deep Residual Learning for Image Recognition. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2016-December:770–778
3. Geoff Hinton: On Radiology - YouTube. www.youtube.com/watch?v=2HMPRXstSvQ&t Accessed 5 Feb 2022
4. Weikert T, Francone M, Abbara S, et al (2020) Machine learning in cardiovascular radiology: ESCR position statement on design requirements, quality assessment, current applications, opportunities, and challenges. European Radiology 1–14
5. Brady SL, Trout AT, Somasundaram E, Anton CG, Li Y, Dillman JR (2021) Improving image quality and reducing radiation dose for pediatric CT by using deep learning reconstruction. Radiology 298:180–18

APPENDIX

- Summary
- Nederlandse samenvatting
- Acknowledgments
- Curriculum Vitae
- List of Publications
- Oral Presentations at Conferences

Summary

This thesis applies machine learning to a broad spectrum of tasks in the field of cardiothoracic radiology.

Data curation: Chapter 2 demonstrates how ML can be used for time-efficient data curation. Three methods, a convolutional neural network (CNN), a support vector machine and a random forest classifier were trained on the impression sections of 2,801 CT pulmonary angiograms, which had been manually labeled by a radiologist according to whether they describe the presence of pulmonary embolism or not. The performance of the three approaches for report classification (pulmonary embolism: yes or no) was assessed on a test set of 1,377 CT pulmonary angiogram impression sections. The CNN reached the highest accuracy (99.1%; 95% CI 98.5-99.6%). Of interest, all three approaches reached an accuracy of >93% after training with a subset of 470 labelled impressions (=16.8% of available training data). Because the time-consuming compilation of large curated datasets is a prerequisite for most scientific projects and quality control measures, these ML techniques are a valuable tool for both researchers and clinicians.

Detection: Chapter 3 evaluates the performance of a deep convolutional neural network (CNN) for detection of pulmonary embolism on CT pulmonary angiograms. In times of ever-increasing numbers of imaging tests performed, the swift identification and communication of exams with critical findings becomes one of the major tasks in radiology. In this study, the two-step algorithm pipeline consisted of a 3D CNN with ResNet architecture for embolus candidate generation and a feature-based algorithm for false-positive reduction. It had been trained on 28,000 CTPAs from other institutions. The test set comprised 1,465 consecutive CTPAs acquired during one year at a university hospital. On a per-examination level, the model's sensitivity was 92.7% (95% CI: 88.3-95.5) and specificity was 95.5% (95% CI: 94.2-96.6). Most false-positive findings were caused by contrast agent-related flow artifacts and (unremarkable) pulmonary veins. This algorithm is currently implemented in various hospitals and used to trigger sending warning messages to radiologists and for worklist prioritization.

Detection: Chapter 4 investigates the performance of a ResNet-based CNN for detection of rib fractures on a large dataset of whole-body trauma CTs acquired at a level-1 trauma center. The CNN had been trained on 11,455 chest CTs from eight medical institutions.

The independent test set included 511 whole-body trauma CTs. On the per examination level, sensitivity was 87.4% (95 CI: 81.2-92.1) and specificity 91.5% (95% CI: 88.0-94.2). Sensitivity on a per finding level was lower with 65.7% (95% CI: 68.8-92.4). Ninety-seven true rib fractures that had not been described in the official written radiology reports were detected by the algorithm. Therefore, the algorithm can help to avoid underreporting. Detection of rib fractures has therapeutical consequences such as close monitoring for secondary consequences like pneumothorax.

Segmentation: Chapter 5 investigates the performance of ML for detection and segmentation of pulmonary tumors of various size on FDG-PET/CT. Therefore, the PET/CTs of 320 patients with histologically confirmed lung cancer were processed with an algorithm pipeline consisting of a lung segmentation algorithm (deep image-to-image network), an algorithm for nodule candidate generation (faster RCNN) and a false-positive reduction stage (ResNet). The mean lung tumor diameter in the test set of 5.0 cm (SD:3.4) was significantly larger than the pulmonary nodules that had been used for training (<3.0 cm). Consequently, the tumor detection rate dropped from 90.4% for T1 tumors to 8.8% for T4 tumors. Besides lesion size, pleural contact was the most relevant predictor of non-detection. Regarding segmentation, predicted volumes correlated strongly with ground-truth volumes for T1-tumours (r=0.91). Correlation coefficients dropped for larger tumors. Whereas this performance pattern was expected due to the structure of the training data, the study emphasizes the general importance of performance sub-analyses for different stages of disease.

Prediction: Chapter 6 uses multiple CNNs to extract cardiovascular and pulmonary imaging parameters from 120 chest CTs of patients RT-PCR positive for SARS-CoV-2, acquired at time of initial presentation at the hospital. For cardiovascular metrics, algorithms based on U-Net and ResNet were used to extract five cardiovascular metrics, including the volume of coronary calcifications, the total pericardial volume and aortic diameters. For extraction of pulmonary metrics, a DenseUNet trained on 901 chest CTs segmented all pulmonary opacities. On this basis, by means of traditional image processing and computation, six pulmonary imaging metrics were calculated. Furthermore, patient demographics and laboratory markers of inflammation were retrieved at the day of presentation at the hospital and combined to models predicting the patients' future treatment journey (outpatient vs. hospital admission vs. ICU admission), which was ascertained 12 weeks after the end of data collection. The model based on the

image parameters alone was an excellent classifier for differentiation of patients eventually needing ICU-care vs. those who did not (AUC=0.88; 95% CI: 0.79-0.97). Adding demographic information and laboratory findings to the model further improved the performance (AUC=0.91; 95% CI: 0.85-0.98). Therefore, the algorithm was able to identify patients at high risk of needing intensified care within the following weeks at the time of initial hospital presentation.

Framework: Chapter 7 provides a framework for the design and evaluation of ML studies in cardiothoracic radiology. This includes prerequisites of ML projects with regard to hardware, software, and expertise. Furthermore, it contains a checklist of items that should be reported in research articles in the field of cardiothoracic imaging.

Conclusion and Outlook: This thesis provided examples of application of ML to solve problems in the field of cardiothoracic radiology. It was demonstrated that this is not confined to the core duty of radiology departments, image analysis, but includes the complete range of tasks, beginning with data acquisition and data curation to clinical application and predictive analytics. It was found that ML is state-of-the art for image recognition tasks and that performances of current applications are at levels allowing for clinical application. At the same time, it was noticed that we are far from a complete automatization of the radiology workflow.

ML has arrived at the heart of cardiothoracic radiology. The reasonable integration of resulting applications into clinical workflows will be a central field of activity in the years to come.

Nederlandse samenvatting

Dit proefschrift past machine learning toe op een breed spectrum van taken in de cardiothoracale radiologie.

Data curatie: Hoofdstuk 2 laat zien hoe ML kan worden gebruikt voor tijdsefficiënte data curatie. Drie methoden, een convolutioneel neuraal netwerk (CNN), een support vector machine en een random forest classifier werden getraind middels 2.801 CT pulmonale angiogrammen, die handmatig door een radioloog waren gelabeld op basis van de vraag of ze de aanwezigheid van longembolie beschrijven of niet. De prestaties van de drie methoden voor rapportclassificatie (longembolie: ja of nee) werden beoordeeld met een testset van 1.377 CT-pulmonaire angiogrammen. Het CNN bereikte de hoogste nauwkeurigheid (99,1%; 95% CI 98,5-99,6%). Van belang is dat alle drie methoden een nauwkeurigheid van >93% bereikten na training met 470 gelabelde samenvattingen (16,8% van de beschikbare trainingsdata). Omdat de tijdrovende compilatie van grote gecureerde datasets een voorwaarde is voor de meeste wetenschappelijke projecten en kwaliteitscontrolemaatregelen, zijn de gepresenteerde ML-technieken een waardevol hulpmiddel voor zowel onderzoekers als clinici, om de tijd die nodig is voor de voorbereiding van de data, duidelijk te reduceren.

Detectie: Hoofdstuk 3 evalueert de prestaties van een diep CNN voor detectie van een longembolie op CT angiogrammen (CTPA). Bij het stijgend aantal uitgevoerde onderzoeken per jaar, wordt de snelle identificatie en communicatie van onderzoeken met kritische diagnosen een van de meest uitdagende taken in de radiologie. In deze studie bestond de tweetraps algoritmepijplijn uit een 3D CNN met ResNet architectuur voor het genereren van emboluskandidaten en een feature-based algoritme voor reductie van vals-positieve resultaten. Het was getraind op 28.000 CTPA's van externe ziekenhuisen. De testset bestond uit 1.465 opeenvolgende CTPA's die gedurende een jaar in een academisch ziekenhuis waren afgenomen. Per CTPA bedroeg de sensitiviteit van het model 92,7% (95% CI: 88,3-95,5) en de specificiteit 95,5% (95% CI: 94,2-96,6). De meeste vals-positieve bevindingen werden veroorzaakt door contrastmiddel-gerelateerde flow artefacten en onopvallende pulmonale vaten. Dit algoritme wordt momenteel geïmplementeerd in verschillende ziekenhuizen en gebruikt. Ook is een workflow gerealiseerd om waarschuwingsberichten naar radiologen te sturen en voor hun werklijst te prioriteren.

Detectie: Hoofdstuk 4 onderzoekt de prestaties van een ResNet-gebaseerde CNN voor detectie van ribfracturen op een grote dataset van trauma CT scans, verkregen in een level-1 traumacentrum. Het CNN was getraind op 11.455 thorax CT scans van acht medische instellingen. De onafhankelijke test set bestond uit 511 trauma CT scans verkregen vanuit traumacentra. Per onderzoek gezien was de sensitiviteit 87,4% (95 CI: 81,2-92,1) en de specificiteit 91,5% (95% CI: 88,0-94,2). De sensitiviteit per laesie was lager met 65,7% (95% CI: 68,8-92,4). Zevenennegentig ribfracturen die niet waren beschreven in het klinische verslag werden door het algoritme gedetecteerd. Het algoritme kan dus helpen om onderrapportage te voorkomen. Detectie van ribfracturen heeft therapeutische gevolgen zoals nauwgezette controle op secundaire gevolgen zoals pneumothorax.

Segmentatie: Hoofdstuk 5 onderzoekt de prestaties van ML voor detectie en segmentatie van pulmonale tumoren van verschillende grootte op FDG-PET/CT. Daartoe werden de PET/CT's van 320 patiënten met histologisch bevestigde longkanker verwerkt met een algoritme-pijplijn bestaande uit een longsegmentatie-algoritme (deep image-to-image network), een algoritme voor het genereren van nodule-kandidaten (faster RCNN) en een vals-positieve reductiefase (ResNet). De gemiddelde longtumordiameter in de testset van 5,0 cm (SD: 3,4 cm) was aanzienlijk groter dan de longnodules die voor de training waren gebruikt (<3,0 cm). Als gevolg daarvan daalden de tumordetectiepercentages van 90,4% voor T1 tumoren tot 8,8% voor T4 tumoren. Naast de grootte van de laesie was pleuracontact de meest relevante voorspeller van niet-detectie. Wat segmentatie betreft, correleerden voorspelde volumes sterk met de refferentievolumes voor T1-tumoren (r=0,91). De correlatiecoëfficiënten daalden voor grotere tumoren. Hoewel dit prestatiepatroon werd verwacht als gevolg van de structuur van de trainingsgegevens, benadrukt de studie het algemene belang van prestatie-subanalyses voor verschillende ziektestadia.

Voorspelling: Hoofdstuk 6 worden meerdere CNNs gepresenteerd waarmee cardiovasculaire en pulmonaire beeldvormingsparameters uit 120 thorax CTs van patiënten die positief zijn voor SARS-CoV-2 kunnen worden gemeten. Alle onderzoeken werden verkregen op het moment van de eerste presentatie in het ziekenhuis. Voor cardiovasculaire parameters werden algoritmen op basis van U-Net en ResNet gebruikt om vijf cardiovasculaire parameters te extraheren, waaronder het volume van coronaire verkalkingen, het totale pericardiale volume en de aortadiameter. Voor de extractie van pulmonale metriek werden met een DenseUNet, getraind op 901 CT's van de thorax, alle

pulmonale verdichtingen gesegmenteerd. Op basis hiervan werden, door middel van traditionele beeldverwerking en -berekening, zes pulmonale beeldvormingskenmerken berekend. Bovendien werden demografische gegevens van de patiënt en laboratoriummarkers van ontsteking verzameld op de dag van de presentatie in het ziekenhuis en gecombineerd tot modellen die het toekomstige behandeltraject van de patiënten voorspelden (poliklinisch vs. ziekenhuisopname vs. IC-opname), dat 12 weken na het einde van de gegevensverzameling werd vastgesteld. Het model op basis van de beeldparameters alleen was een uitstekende classificator voor het onderscheiden van patiënten die uiteindelijk IC-zorg nodig hadden en patiënten die dat niet hadden (AUC=0,88; 95% CI: 0,79-0,97). Toevoeging van demografische informatie en laboratoriumresultaten aan het model verbeterde de prestatie (AUC=0,91; 95% CI: 0,85-0,98). Kortom, het algoritme was in staat om patiënten te identificeren met een hoog risico op geïntensiveerde zorg binnen de volgende weken op het moment van de eerste ziekenhuisopname.

ML-studiedesign: Hoofdstuk 7 biedt een kader voor het ontwerp en de evaluatie van ML-studies in de cardiothoracale radiologie. Dit omvat randvoorwaarden van ML projecten met betrekking tot hardware, software, en expertise. Verder levert het een checklist met items die gerapporteerd moeten worden in een wetenschappelijk artikel op het gebied van cardiothoracale beeldvorming.

Conclusie en vooruitzichten: Dit proefschrift beschrijft voorbeelden van toepassing van ML om uitdagingen op het gebied van cardiothoracale radiologie op te lossen. Aangetoond werd dat dit niet beperkt blijft tot de kerntaak van radiologie-afdelingen, beeldanalyse, maar het complete takenpakket omvat, beginnend bij data-acquisitie en data-curatie en rijkend tot de klinische toepassing en voorspellende analyses. ML is de state-of-the-art voor beeldherkenningstaken en dat de prestaties van de huidige toepassingen op een niveau liggen dat klinische toepassing mogelijk maakt. Tegelijkertijd werd opgemerkt dat we nog ver verwijderd zijn van een volledige automatisering van de radiologie-workflow.

ML heeft een belangrijke plek verworven in het hart van de cardiothoracale radiologie. De zinvolle integratie van resulterende toepassingen in klinische workflows zal in de komende jaren tot de centrale werkterreinen behoren.

Acknowledgments

Throughout the writing of this thesis and the research projects it is based on I have received a great deal of support.

I would first like to thank my promotor and supervisor, Professor Tim Leiner, a true role model and expert in the field of cardiothoracic imaging and machine learning, who provided guidance at any stage. Besides his many obligations at Universiteit Utrecht, Mayo Clinic, and elsewhere, he always found time for highly productive meetings. His advices were always precise and valuable- Tim, I want to thank you very much for your effort, I absolutely do not take this for granted.

Furthermore, I would like to thank Professor Alexander W. Sauter, Dr. Bram Stieltjes and Dr. Gregor Sommer, who were indispensable advisors and partners during the conduction of the scientific projects. All have become good friends. Alex, I want to especially thank you for the countless hours we spent together discussing the projects' methods, manuscript drafts and conference contributions. Bram, you for your always enlightening opinions and advices that are based on great experience. And you, Gregor, for your thorough commitment to research and education. If all three of you enjoyed the common times only half as much as I did, you had a great time.

Machine Learning in radiology is a team effort that requires people with different professional backgrounds. I was fortunate enough to find a group of highly talented and motivated computer scientists, medical doctors and other scientists, with whom I worked together closely. Shan Yang, Joshy Cyriac, Manuela Moor, Michael Bach, Maurice Henkel, Ivan Nesic, Marko Obradovic, Victor Parmar, Jakob Wasserthal, Jan A. Roth, Eleftherios Remoundos and Tugba Akinci D'Antonoli are among them. A big "thank you" to all of you and all other team members for your expertise, advices and friendship. I will never forget you.

Every great team needs someone who provides the organizational framework. Besides Dr. Bram Stieltjes, I want to express my special gratitude to Professor Elmar Merkle, head of the Department of Radiology, and Professor Jens Bremerich, head of the Department of Cardiothoracic Imaging, for doing so. They have gathered an excellent team at the Department of Radiology at the University Hospital Basel. Their excellent sense for talent and team-building is the prerequisite for all successes.

I also want to thank Dr. Christoph Aberle from the Radiological Physics Department at University Hospital Basel for providing technical information for many projects. Furthermore, I want to thank all younger colleagues who I was fortunate enough to support with their projects, especially Luca, Andrej, Lorraine, Adrian and Julien- thank you for the common time, discussions and your ideas. Furthermore, I want to thank all colleagues of the radiology residence program at the Department of Radiology, many of whom became co-authors.

But my gratitude is not limited to colleagues in Utrecht and Basel- I also want to thank all cooperation partners at the German Cancer Research Center who were involved in the PET/CT ML project, especially Dr. Paul Jäger and Professor Klaus H. Maier-Hein. Last but not least, I want to thank Dr. Bruno Dietsche, who helped me with my first steps in the world of science at the University of Marburg - I will always be grateful for the time spent together and the many hours discussing research topics while playing table tennis in the basement of the University Hospital in Marburg.

Curriculum Vitae

I am a resident and research fellow at the Department of Radiology at the University Hospital Basel, Switzerland, and PhD candidate at the Utrecht University with more than five years of professional experience. I have a strong interest and expertise at the interface of healthcare, machine learning, research, politics and business.

Work Experience

since 10/2016	Resident and Research Fellow at the Department of Radiology, University Hospital Basel, Switzerland
since 01/2020	PhD candidate at Utrecht University, The Netherlands

Education

10/2008 – 06/2016	Medical School and MD, University of Marburg, Germany
	- Internships in Germany, Poland, Austria, and Switzerland
10/2005 – 09/2008	BA in Politics and Management, University of Konstanz, Germany
	- Semester abroad at Warsaw School of Economics, Poland
09/2004 – 08/2005	Voluntary Service at the German-Polish Youth Office, Warsaw, Poland

Language Skills

German	native
English	proficient
Polish	proficient
French	intermediate
Latin	intermediate

List of Publications

Articles in peer-reviewed journals

Weikert T, Litt HI, Moore WH, Abed M, Azour L, Noor AM, Friebe L, Linna N, Yerebakan HZ, Shinagawa Y, Hermosillo G, Allen-Raffl S, Ranganath M, Sauter AW (2023) Reduction in radiologist interpretation time of serial CT and MR imaging findings with deep learning identification of relevant priors, series and finding locations. *Academic Radiology*

Weikert T, Jaeger PF, Yang S, Baumgartner M, Breit HC, Winkel DJ, Sommer G, Stieltjes B, Thaiss W, Bremerich J, Maier-Hein KH, Sauter AW (2023) Automated lung cancer assessment on 18F-PET/CT using Retina U-Net and anatomical region segmentation. *European Radiology*

Sexauer R, Hejduk P, Borkowski K, Ruppert C, **Weikert T**, Dellas S, Schmidt N (2023) Diagnostic accuracy of automated ACR BI-RADS breast density classification using deep convolutional neural networks. *European Radiology*

Breit HC, Vosshenrich J, Clauss M, **Weikert T**, Stieltjes B, Kovacs BK, Bach M, Harder D (2023) Visual and quantitative assessment of hip implant-related metal artifacts at low field MRI: a phantom study comparing a 0.55-T system with 1.5-T and 3-T systems. *European Radiology Experimental*

Weikert T, Friebe L, Wilder-Smith A, Yang S, Sperl JI, Neumann D, Balachandran A, Bremerich J, Sauter AW (2022) Automated quantification of airway wall thickness on chest CT using retina U-Nets–Performance evaluation and application to a large cohort of chest CTs of COPD patients. *European Journal of Radiology*

Saba L, Loewe C, **Weikert T**, Williams MC, Galea N, Budde RPJ, Vliegenthart R, Velthuis BK, Francone M, Bremerich J, Natale L, Nikolaou K, Dacher JN, Peebles C, Caobelli F, Redheuil A, Dewey M, Kreitner KF, Salgado R (2022) State-of-the-art CT and MR imaging and assessment of atherosclerotic carotid artery disease: standardization of scanning protocols and measurements—a consensus document by the European Society of Cardiovascular Radiology (ESCR). *European Radiology*

Binsfeld Gonçalves L, Nesic I, Obradovic M, Stieltjes B, **Weikert T**, Bremerich J (2022) Natural Language Processing and Graph Theory: Making Sense of Imaging Records in a Novel Representation Frame. *JMIR Medical Informatics*

Saba L, Loewe C, **Weikert T**, Williams MC, Galea N, Budde RPJ, Vliegenthart R, Velthuis BK, Francone M, Bremerich J, Natale L, Nikolaou K, Dacher JN, Peebles C, Caobelli F, Redheuil A, Dewey M, Salgado R (2022) State of the art CT and MR imaging and assessment of carotid artery disease: the reporting-a consensus document by the European Society of Cardiovascular Radiology (ESCR). *European Radiology*

Wilder-Smith AJ, Yang S, **Weikert T**, Bremerich B, Haaf P, Segeroth M, Ebert LC, Sauter AW, Sexauer R (2022) Automated Detection, Segmentation, and Classification of Pericardial Effusions on Chest CT Using a Deep. *Diagnostics*

Poletti J, Bach M, Yang S, Sexauer R, Stieltjes B, Rotzinger DC, Bremerich J, Sauter AW, **Weikert T** (2022). Automated lung vessel segmentation reveals blood vessel volume redistribution in viral pneumonia. *European Journal of Radiology*

Sexauer R, Yang S, **Weikert T**, Poletti J, Bremerich J, Roth JA, Sauter AW, Anastasopoulos C (2022) Automated detection, segmentation, and classification of pleural effusion from computed tomography scans using machine learning. *Investigative Radiology*

Seyam M, **Weikert T**, Sauter A, Brehm A, Psychogios M-N, Blackham KA (2022) Utilization of Artificial Intelligence–based Intracranial Hemorrhage Detection on Emergent Noncontrast CT Images in Clinical Workflow. *Radiology AI*

Weikert T, Rapaka S, Grbic S, et al (2021) Prediction of Patient Management in COVID-19 Using Deep Learning-Based Fully Automated Extraction of Cardiothoracic CT Metrics and Laboratory Findings. *Korean Journal of Radiology*

Weikert T, Francone M, Abbara S, et al (2021) Machine learning in cardiovascular radiology: ESCR position statement on design requirements, quality assessment, current applications, opportunities, and challenges. *European Radiology*

Pusterla O, Heule R, Santini F, **Weikert T**, et al (2021) MRI lung lobe segmentation in pediatric cystic fibrosis patients using a recurrent neural network trained with publicly accessible CT datasets. *arXiv*

Romanov A, Bach M, Yang S, Franzeck FC, Sommer G, Anastasopoulos C, Bremerich J, Stieltjes B, **Weikert T** (shared senior author), Sauter AW (2021) Automated CT Lung Density Analysis of Viral Pneumonia and Healthy Lungs Using Deep Learning-Based Segmentation, Histograms and HU Thresholds. *Diagnostics*

Abel L, Wasserthal J, **Weikert T**, et al (2021) Automated Detection of Pancreatic Cystic Lesions on CT Using Deep Learning. *Diagnostics*

Pradella M, **Weikert T**, Sperl JI, et al (2021) Fully automated guideline-compliant diameter measurements of the thoracic aorta on ECG-gated CT angiography using deep learning. *Quantitative Imaging in Medicine and Surgery*

Breit HC, Block KT, Winkel DJ, Gehweiler JE, Henkel MJ, **Weikert T**, Stieltjes B, Boll DT, Heye TJ (2021) Evaluation of liver fibrosis and cirrhosis on the basis of quantitative T1 mapping: Are acute inflammation, age and liver volume confounding factors? *European Journal of Radiology*

Schmuelling L, Franzeck FC, Nickel CH, **Weikert T**, et al (2021) Deep learning-based automated detection of pulmonary embolism on CT pulmonary angiograms: No significant effects on report communication times and patient turnaround in the emergency department nine months after technical implementation. *European Journal of Radiology*

Winkel DJ, Breit HC, **Weikert T**, Stieltjes B (2021) Building Large-Scale Quantitative Imaging Databases with Multi-Scale Deep Reinforcement Learning: Initial Experience with Whole-Body Organ Volumetric Analyses. *Journal of Digital Imaging*

Weikert T, Cyriac J, Yang S, Nesic I, Parmar V, Stieltjes B (2020) A practical guide to artificial intelligence-based image analysis in radiology. *Investigative Radiology*

Weikert T, Nesic I, Cyriac J, Bremerich J, Sauter AW, Sommer G, Stieltjes B (2020) Towards automated generation of curated datasets in radiology: Application of natural language processing to unstructured reports exemplified on CT for pulmonary embolism. *European Journal of Radiology*

Weikert T, Noordtzij LA, Bremerich J, Stieltjes B, Parmar V, Cyriac J, Sommer G, Sauter AW (2020) Assessment of a Deep Learning Algorithm for the Detection of Rib Fractures on Whole-Body Trauma Computed Tomography. *Korean Journal of Radiology*

Weikert T, Winkel DJ, Bremerich J, Stieltjes B, Parmar V, Sauter AW, Sommer G (2020) Automated detection of pulmonary embolism in CT pulmonary angiograms using an AI-powered algorithm. *European Radiology*

Henkel M, **Weikert T** (corresponding author), Marston K, et al (2020) Lethal COVID-19: Radiological-Pathological Correlation of the Lungs. *Radiology: Cardiothoracic Imaging*

Anastasopoulos C, **Weikert T**, Yang S, et al (2020) Development and clinical implementation of tailored image analysis tools for COVID-19 in the midst of the pandemic: The synergetic effect of an open, clinically embedded software development platform and machine learning. *European Journal of Radiology*

Winkel DJ, **Weikert T**, Breit HC, Chabin G, Gibson E, Heye TJ, Comaniciu D, Boll DT (2020) Validation of a fully automated liver segmentation algorithm using multi-scale deep reinforcement learning and comparison versus manual segmentation. *European Journal of Radiology*

Weikert T, Sommer G, Tamm M, Haegler P, Cyriac J, Sauter AW, Hostettler K, Bremerich J (2019) Centralized expert HRCT Reading in suspected idiopathic pulmonary fibrosis: Experience from an Eurasian teleradiology program. *European Journal of Radiology*

Weikert T, Akinci D'Antonoli T, Bremerich J, Stieltjes B, Sommer G, Sauter AW, Bonanno E (2019) Evaluation of an AI-Powered Lung Nodule Algorithm for Detection and 3D Segmentation of Primary Lung Tumors. *Contrast Media & Molecular Imaging*

Weikert T, Maas OC, Haas T, Klarhöfer M, Bremerich J, Forrer F, Sauter AW, Sommer G (2019) Early Prediction of Treatment Response of Neuroendocrine Hepatic Metastases after Peptide Receptor Radionuclide Therapy with ^{90}Y-DOTATOC Using Diffusion Weighted and Dynamic Contrast-Enhanced MRI. *Contrast Media & Molecular* Imaging

Winkel DJ, Heye T, **Weikert T**, Boll DT, Stieltjes B (2019) Evaluation of an AI-Based Detection Software for Acute Findings in Abdominal Computed Tomography Scans: Toward an Automated Work List Prioritization of Routine CT Examinations. *Investigative Radiology*

Sexauer R, **Weikert T**, Mader K, Wicki A, Schädelin S, Stieltjes B, Bremerich J, Sommer G, Sauter AW (2018) Towards More Structure: Comparing TNM Staging Completeness and Processing Time of Text-Based Reports versus Fully Segmented and Annotated PET/CT Data of Non-Small-Cell Lung Cancer. *Contrast Media & Molecular Imaging*

Sauter AW, Stieltjes B, **Weikert T**, Gatidis S, Wiese M, Klarhöfer M, Wild D, Lardinois D, Bremerich J, Sommer G (2017) The spatial relationship between apparent diffusion coefficient and standardized uptake value of ^{18}F-fluorodeoxyglucose has a crucial influence on the numeric correlation of both parameters in PET/MRI of lung tumors. *Contrast Media & Molecular Imaging*

Sauter AW, Stieltjes B, **Weikert T**, Gatidis S, Wiese M, Klarhöfer M, Wild D, Lardinois D, Bremerich J, Sommer G (2017) The Spatial Relationship between Apparent Diffusion Coefficient and Standardized Uptake Value of 18 F-Fluorodeoxyglucose Has a Crucial Influence on the Numeric Correlation of Both Parameters in PET/MRI of Lung Tumors. *Contrast Media & Molecular Imaging.*

Dietsche B, Backes H, Laneri D, **Weikert T**, Witt SH, Rietschel M, Sommer J, Kircher T, Krug A (2014) The impact of a CACNA1C gene polymorphism on learning and hippocampal formation in healthy individuals: a diffusion tensor imaging study. *NeuroImage 89:256–261*

Book chapters

Weikert T, Leiner T (2022). How to Write and Review an Artificial Intelligence Paper. In: De Cecco, C.N., van Assen, M., Leiner, T. (eds) Artificial Intelligence in Cardiothoracic Imaging. Contemporary Medical Imaging. *Humana, Cham*. https://doi.org/10.1007/978-3-030-92087-6_53

Oral Presentations at Conferences

RSNA 2022 Significantly Accelerated Longitudinal Assessment of Imaging Findings on CT and MRI with Deep Learning Identification of Relevant Prior Exams and Finding Locations

ECR 2022 A Laboratory-Medicine-Like Approach to the Analysis of Unremarkable Chest Radiographs Using Artificial Intelligence

ECR 2021 Advanced Education Session: Training Data for Deep Learning: What is Needed?

ECR 2021 Session Moderation: Advanced New Techniques in Cardiovascular Imaging

FIL ROUGE 2021: Artificial Intelligence Applied to Carotid Imaging: From Detection to Characterization and Stratification

RSNA 2021 Automated and Comprehensive Quantification Of Airway Wall Thickness In A Large Patient Collective As CT Imaging Biomarker Of Chronic Obstructive Pulmonary Disease

RSNA 2021 Automated Detection Of Primary Lung Cancer Of All Stages And Associated Metastases On FDG-PET/CT Using A Retina-U-Net Algorithm

ECR 2020 Liver Size Estimation Revisited: Why we Should Replace Distance Measurements in Midclavicular Line by Automated Volumetry

ECR 2020 Automated Detection of Primary Lung Cancer of All Stages and Associated Metastases on FDG-PET/CT Using Three Retina U-Net Algorithms

ECR 2020 Cross-border Knowledge Sharing in the Diagnostic Workup of Rare Disease- Experiences with a Transeuropean Teleradiology Project on Idiopathic Pulmonary Fibrosis.

ECR 2019 AI-powered Detection of Pulmonary Embolism in CT Pulmonary Angiograms: a Validation Study of the Diagnostic Performance of Prototype Algorithms

ECR 2019 Generation of a Curated Dataset from Unstructured Reports Using Natural Language Processing, Illustrated on CT Reports Regarding Pulmonary Embolism

ECR 2019 Assessing the Precision of an AI-powered Algorithm for Automatic Detection and 3D Segmentation of Primary Tumours in NSCLC

SCR 2019 Creation of Curated Datasets from Unstructured Radiology Reports using Natural Language Processing- Exemplified on CT Reports Regarding Pulmonary Embolism

SCR 2019 Automated Detection of Pulmonary Embolism in CT Pulmonary Angiograms: Testing the Diagnostic Performance of a Prototype Algorithm

SCR 2019 Evaluation of an AI-powered Algorithm for the Automated Detection and 3D Segmentation of Primary Tumours in NSCLC

SCR 2019 Evaluation of an AI-based Detection of Acute Findings in Abdominal CTs: Towards an Automated Work List Prioritization of Routine CT Exams

RSNA 2019 Patterns of Failure of an AI-based Software: A Report on False Positive Findings of an Algorithm Detecting Pulmonary Embolism on CT Pulmonary Angiograms

RSNA 2019 Reducing the Amount of Training Data for Natural Language Processing (NLP) in Radiology by an Active Learning Approach

ECR 2018 Automated Translation of Radiologic Reports with Deep Learning Powered Translation Engines: a Feasibility Study