BMC Medical Imaging

RESEARCH

Open Access

Automated detection of pulmonary embolism from CT-angiograms using deep learning

Heidi Huhtanen[1*], Mikko Nyman[1], Tarek Mohsen[2], Arho Virkki[3,4], Antti Karlsson[5] and Jussi Hirvonen[1]

Abstract

Background: The aim of this study was to develop and evaluate a deep neural network model in the automated detection of pulmonary embolism (PE) from computed tomography pulmonary angiograms (CTPAs) using only weakly labelled training data.

Methods: We developed a deep neural network model consisting of two parts: a convolutional neural network architecture called InceptionResNet V2 and a long-short term memory network to process whole CTPA stacks as sequences of slices. Two versions of the model were created using either chest X-rays (Model A) or natural images (Model B) as pre-training data. We retrospectively collected 600 CTPAs to use in training and validation and 200 CTPAs to use in testing. CTPAs were annotated only with binary labels on both stack- and slice-based levels. Performance of the models was evaluated with ROC and precision–recall curves, specificity, sensitivity, accuracy, as well as positive and negative predictive values.

Results: Both models performed well on both stack- and slice-based levels. On the stack-based level, Model A reached specificity and sensitivity of 93.5% and 86.6%, respectively, outperforming Model B slightly (specificity 90.7% and sensitivity 83.5%). However, the difference between their ROC AUC scores was not statistically significant (0.94 vs 0.91, $p = 0.07$).

Conclusions: We show that a deep learning model trained with a relatively small, weakly annotated dataset can achieve excellent performance results in detecting PE from CTPAs.

Keywords: Artificial intelligence, Emergency radiology, Pulmonary embolism, Deep learning, Automated detection

Background

Acute pulmonary embolism (PE) is a life-threatening condition and has an estimated incidence of 60 per 100,000 [1–3]. Symptoms of PE are often non-specific, e.g., dyspnoea and chest pain, and clinical diagnosis can be challenging. Computed tomography pulmonary angiography (CTPA) has become the golden standard in diagnosing PE [4], but accurate interpretation of CTPAs requires experience and time. However, the majority of CTPAs are negative, which might be attributed to clinicians not routinely using pre-test probability or rule-out criteria, and to a generally increasing utilization of emergency CT imaging [5, 6]. In the published literature, the yield of positive studies varies from less than 10 to 20–30% [5, 7–9]. Because increasing workload and fatigue related to after-hours work may cause more diagnostic errors in emergency radiology [10–12], an automated PE detection system could aid radiologists to avoid mistakes and prioritize the reading order of studies to ensure rapid evaluation of PE positive cases.

*Correspondence: hejohuh@utu.fi
[1] Department of Radiology, University of Turku and Turku University Hospital, Turku, Finland
Full list of author information is available at the end of the article

© The Author(s) 2022. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

The emergence of deep learning (DL), a subfield of artificial intelligence (AI), has increased interest in automatized detection and diagnostic tools in radiology [13]. Before that, multiple different computer-aided detection (CAD) models for diagnosing PE had been developed using manually encoded methods like segmentation, detection of low-attenuated areas, and/or feature analysis [14, 15]. Use of CAD as a concurrent reader has been shown to increase reader sensitivity [16], but high yield of false positives (FPs) has remained as a major drawback [17]. FPs can cause frustration to radiologists, but they can also increase the risk for false diagnoses due to automation bias—the tendency for humans to favor machine-made decision [18, 19]. Implementation of DL to PE detection has led to models generating fewer FPs without reducing sensitivity [20, 21]. The Radiological Society of North America (RSNA) chose PE detection as its AI challenge in 2020, and later published a public dataset of 12 195 annotated CTPA studies to encourage the development of PE detection models [22].

One limitation of the previous CAD systems and some of the newer DL-based models is the requirement for densely annotated training data, where each distinct embolus is marked or segmented manually. Weikert et al. tested a prototype commercial model (Aidoc Medical, Tel Aviv, Israel) based on fast region-based convolutional neural network (CNN) and trained with 28,000 segmented CTPAs. The model achieved sensitivity and specificity of 92.7% and 95.5%, respectively [23]. Buls et al. also tested the model (version 1.3) and achieved a similar specificity of 95% but a lower sensitivity of 73% [24]. However, creating an annotated training set of this magnitude for own model development may not be feasible for single hospitals or research teams. Using small datasets can lead to overfitting, meaning the model learns the training data too well and generalizes poorly to new data. Overfitting can be tackled to some degree with different regularization techniques, like dropout and early stopping [25].

Weakly supervised learning with sparser annotations is another way to reduce the manual annotation work without reducing the training set size too much. Rajan et al. proposed a sparse annotation method where emboli contours were drawn only for every 10 mm of CTPA slices, reducing the required manual work considerably, yet achieving an area under the receiving operating characteristics (ROC AUC) of 0.78 [26]. Feng et al. proposed a weakly supervised 3D CNN model for lung nodule segmentation and detection requiring only single-coordinate nodule annotations, which could also be applied to PE detection in a similar fashion [27]. Huang et al. assigned only binary labels of PE present or PE absent to CT slices, and their model achieved sensitivity and specificity of 73% and 82%, respectively [28]. Recent advancements with generative adversarial networks (GAN) have enabled creating realistic synthetic data to increase the training set as well as completely unsupervised training of anomaly detection models, although these techniques have not been studied in PE detection yet [29–31].

We sought to develop a neural network model to aid in detecting PE from CTPAs. The model could be used to reduce mistakes or to pre-screen studies to prioritize reading order. Our aim was to study whether training a well-performing model is possible even with limited resources for collecting and annotating data. We used weakly labelled data with slice-based annotations instead of annotations for each distinct embolus. We also used a relatively small training set consisting of only 600 CTPAs, which is less than in many previous studies [23, 26, 28]. To handle the volumetric nature of CT images, we used both 2D convolutional neural network (CNN) to analyze individual slices, and long-short term memory network (LSTM) to process scans as sequences of slices, a combination previously shown to be useful with weakly annotated CT data [32].

Materials and methods

For this retrospective cohort study, we obtained permission from The Hospital District of Southwest Finland. Waiver for written patient consent was not sought from the institutional review board (IRB, called the Ethics Committee of The Hospital District of Southwest Finland), because it is not required by the national legislature for retrospective studies of existing data. The study was conducted in accordance with the Declaration of Helsinki.

Data selection and labelling

The PACS records of Turku University Hospital were searched for all CTPA examinations between January 2016 and October 2018. Preliminary classification of the CTPAs was done automatically by assessing whether the patient had an ICD-10 code consistent with PE (I26) or not in the electronic health record. This first classification was used only to collect enough positive cases among the large number of negative cases; subsequently, all final labels in both classes were manually assigned. First, a randomly sampled cohort was selected from the image archive which included data from multiple sources/scanners. With preliminary classification this produced 419 positive cases, which were used for collecting the training set. Due to the small number of preliminary positive cases, another randomly sampled cohort consisting of preliminary positive cases was selected for the test set, excluding patients that were in the first set. This produced 598 cases. The first search produced 2329

preliminary negative cases, which was enough for both the training and test sets.

We chose to weakly annotate imaging data by not labelling distinct emboli with bounding boxes or segmentations, but rather assigning binary labels (PE positive or PE negative) only to axial slices and whole CT scans. The rationale was to both minimize manual work and test whether weakly annotated data suffice for a DL network to achieve reasonable results.

For this study, only CTPAs with readily available 3.0 mm axial slices were included. Scans with non-diagnostic image quality were excluded. All CTPAs were visually interpreted, and a scan was manually labelled positive if it included even one unambiguous embolus and negative if there were none. A slice was marked positive if one or more emboli were present, and negative if none were present. In the training set, the labelling work was done by a radiology resident (H.H, < 1 year of experience) specifically trained for this task by an experienced board-certified radiologist (M.N., 14 years of experience). In the test set, all scans and slices were double read by H.H. and M.N. in consensus to achieve a better reference standard. Interpretation and labelling were done using Horos software (Horos Project, Annapolis, MD, USA) and ePad annotation platform (Rubin Lab, Stanford Medicine, CA, USA). Collecting and labelling the data was performed securely on PACS inside the hospital network, and the data was de-identified before moving it to the computing platform.

For the training and test sets, 303 and 97 positive CTPAs were collected, respectively. To achieve balanced datasets, equal numbers of negative CTPAs were selected randomly for the training set (305) and test set (107). Because some of the CT studies extended craniocaudally to the level of abdomen and pelvis, the slice coverage was limited to only 96 slices starting from the top (96×3 mm = 288 mm). This was deemed sufficient, because all positive slices in the training set fell within this range. The training set had a total of 52,752 included slices, of which 7170 were positive (14%), and the test had a total of 17,778 included slices, of which 2801 were positive (16%) (Table 1). The datasets did not include the same patients. No systematic differences were observed between the training and test sets in terms of patient demographics, type of scanner used, or the time period the scans were acquired.

Table 1 Dataset information

	Training set	Test set
CTPAs (stacks)	608	204
Positive	303 (50%)	97 (48%)
Negative	305 (50%)	107 (52%)
CT slices	52,752	17,778
Positive	7170 (14%)	2801 (16%)
Negative	45,582 (86%)	14,977 (84%)
Distinct patients	569	201
Male	250 (44%)	88 (44%)
Female	319 (56%)	113 (56%)
Mean age	64	64
CT manufacturer		
Toshiba	569 (94%)	195 (96%)
GE Medical Systems	21 (3%)	5 (2%)
Siemens	18 (3%)	4 (2%)

Data augmentation and preprocessing

We augmented data to increase the number of positive slices in the training set, because the ratio between positive and negative slices was highly imbalanced. For augmentation, we used methods including translation, rotation, blur, Gaussian noise, zoom and elastic transformation. The final training set for the CNN models consisted of ~ 100,000 slices, of which half were negative and half were positive (the original 7170 real positive slices and ~ 43,000 augmented positive slices). Augmented data was not used in training the LSTM part, because it handled the data on the stack-based level which was already balanced between the positive and negative classes.

Preprocessing of the images included rescaling, resizing, segmentation and windowing. Images had originally a height and a width of 512 pixels, except for 10 scans where the original width was larger, and rescaling had to be applied. Images were resized to 386×386 pixels, which was the input size required by the CNN model. Segmentation was performed to reduce the amount of unnecessary information in the images and was done by segmenting the lungs and the heart and excluding other areas (e.g., soft tissues and objects outside the body). Images were windowed with window width 700 HU and window level 100 HU. Finally, DICOM images were converted to 8-bit PNG images.

Model architecture and training

All models were created using Keras deep learning framework (version 2.2.4) with Tensorflow backend (version 1.10.1) and were written in Python programming language (version 3.6, Python Software Foundation).

We processed whole CT scans as series of axial slices using a combination of a CNN, which analyses 2D images, and an LSTM, which processes the slice predictions created by the CNN part as sequences (Fig. 1). A major advantage in this approach is that 2D CNN models are considerably easier to train and require less data than

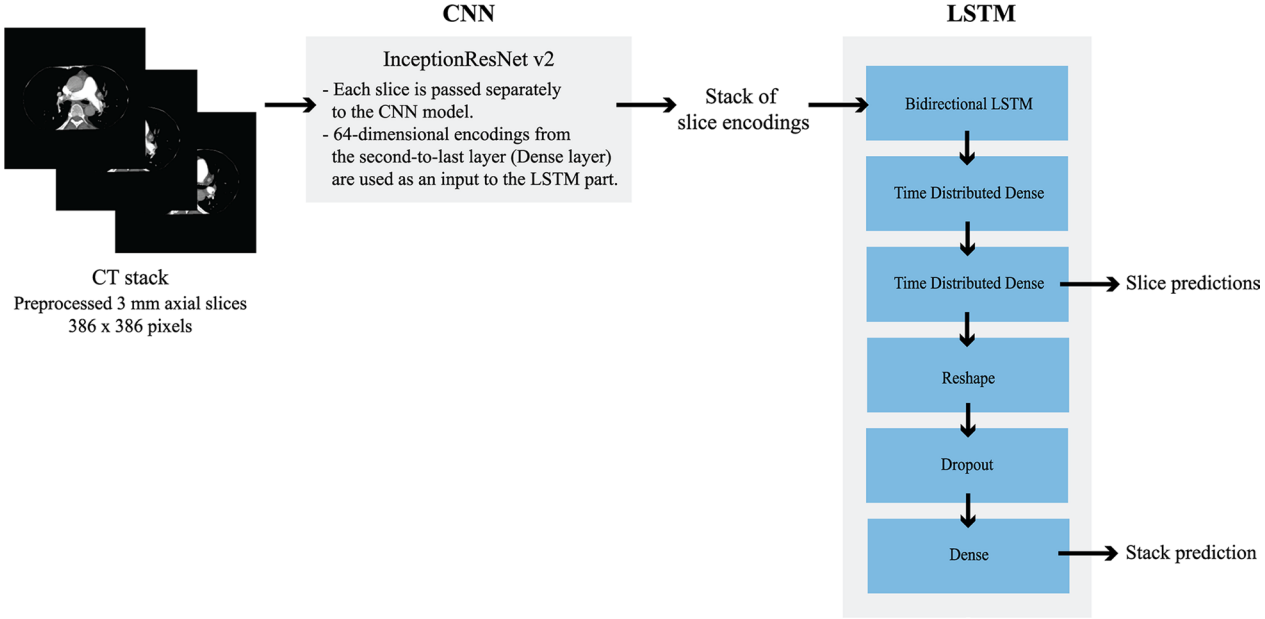

Fig. 1 Scheme of the model architecture

3D CNN models. The LSTM part also allows the model to take findings in neighbouring slices into consideration.

Based on our preliminary experiments with different pre-trained CNN model architectures, we selected InceptionResNetV2 to be used in this study [33]. A simple custom classifier was built on top of the model bottleneck. As a secondary step, a bidirectional LSTM processed data from the CNN model creating a final combination model. Detailed model architectures can be found in the Additional file 1. Transfer learning was implemented, and two separate combination models were trained with different pre-trained weights for the CNN part. Model A was pre-trained with the chest X-ray dataset provided by National Institutes of Health (over 100,000 chest X-rays, weights obtained from https://github.com/i-pan/kaggle-rsna18) and Model B was pre-trained with the ImageNet dataset (over 14 million natural images). The LSTM used 64-dimensional feature vectors from the second to last layer of the CNN as an input. We chose to pre-calculate these feature vectors and train the LSTM part independently from the CNN, because training them together would have required more time and more laborious adjustments of the hyperparameters. A fixed timestep of 96 slices starting from the top was used. CTPAs containing more than 96 slices were clipped and the CTPAs having less than 96 slices were padded with zeros.

All models were trained using a five-fold cross-validation. In this method, the training data is split into five folds of equal size, and five "copies" of the model are trained, each time using a different fold as a validation set to evaluate training results. This method gives a more truthful estimation of the model accuracy when datasets are relatively small. The folds were stratified so that each fold maintained the same distribution of both the class and the number of augmented images as the whole training set, and images from a given patient were always kept in the same fold. During training, augmented images were omitted from the validation set, which then had the original ratio of negative to positive slices (7:1) and CTPAs (1:1).

For the CNN and LSTM models, batch sizes were 48 and 16, and the numbers of epochs were 8 and 12, respectively. Binary cross-entropy was used as a loss function. The optimizers used were Adam with a decay value of 0.01 for CNN models and Nadam for LSTM models. Learning rate was 0.001 for all models. The models were trained using 4 NVidia V100 SXM2 32 GB graphics cards.

Performance evaluation and statistical analysis

The final combination models A and B produced both stack-based and slice-based predictions, and their performance was evaluated on both levels. We plotted receiver operating characteristic (ROC) and precision–recall (PR) curves for test results and calculated their area under the curve (AUC) values. We compared ROC curves of Model A and B using the DeLong method [34]. Statistical significance was set at $p < 0.05$. We determined the best operating thresholds according to the Youden Index, which gives equal weights to sensitivity and specificity as well as to positive and negative classes. Predictions above

this threshold were classified as positive and otherwise as negative. Using selected operating thresholds, we calculated confusion matrices for both stack-based and slice-based predictions. We evaluated the model performance based on five metrics, consisting of accuracy, sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV), and calculated their 95% confidence intervals (CI) using the Wilson test with continuity correction [35]. We also visually inspected example images of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). Model testing was performed using Python programming language (version 3.6, Python Software Foundation). Calculations and statistical analyses were performed using Python, R (version 4.1.0), and RStudio (version 1.4.1717).

Results

Both combination models achieved excellent results on the test set. For Model A, ROC AUC and PR AUC scores for predicting PE on the stack-based level were 0.94 and 0.94, and on the lice-based level 0.97 and 0.90, respectively. For Model B, ROC AUC and PR AUC scores on the stack-based level were 0.91 and 0.91, and on the slice-based level 0.97 and 0.88, respectively. The ROC and PR curves for both models are represented in Fig. 2. ROC curve comparisons between Models A and B showed that there was no statistically significant difference on the stack-based level ($p = 0.07$). On the slice-based level, there was a significant difference ($p < 0.001$), but this seems very minimal as the ROC curves and AUC values are almost identical.

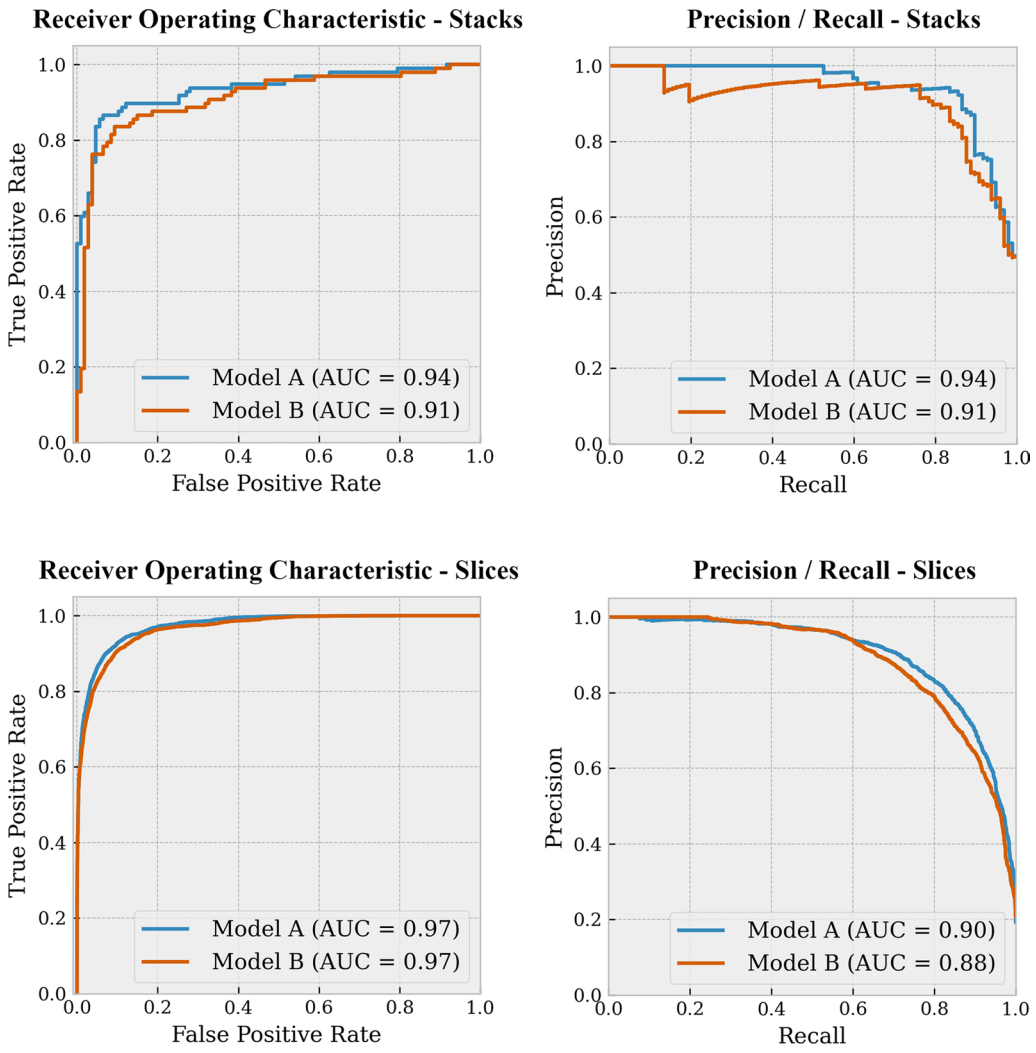

Fig. 2 ROC and PR curves for stack-based and slice-based predictions

The optimal thresholds for classifying the predictions determined by the Youden index were similar between the models, but different between the stack-based and slice-based levels. On the stack-based level, the optimal Youden indices for models A and B were 0.80 and 0.74, corresponding to thresholds 0.797 and 0.858, respectively. On the slice-based level, the Youden indices were 0.83 and 0.81, corresponding to thresholds 0.172 and 0.094, respectively.

Confusion matrices calculated using these thresholds are shown in Fig. 3.

The performance metrics for models on both stack- and slice-based predictions are represented in Table 2. Model A outperformed Model B on all metrics on the

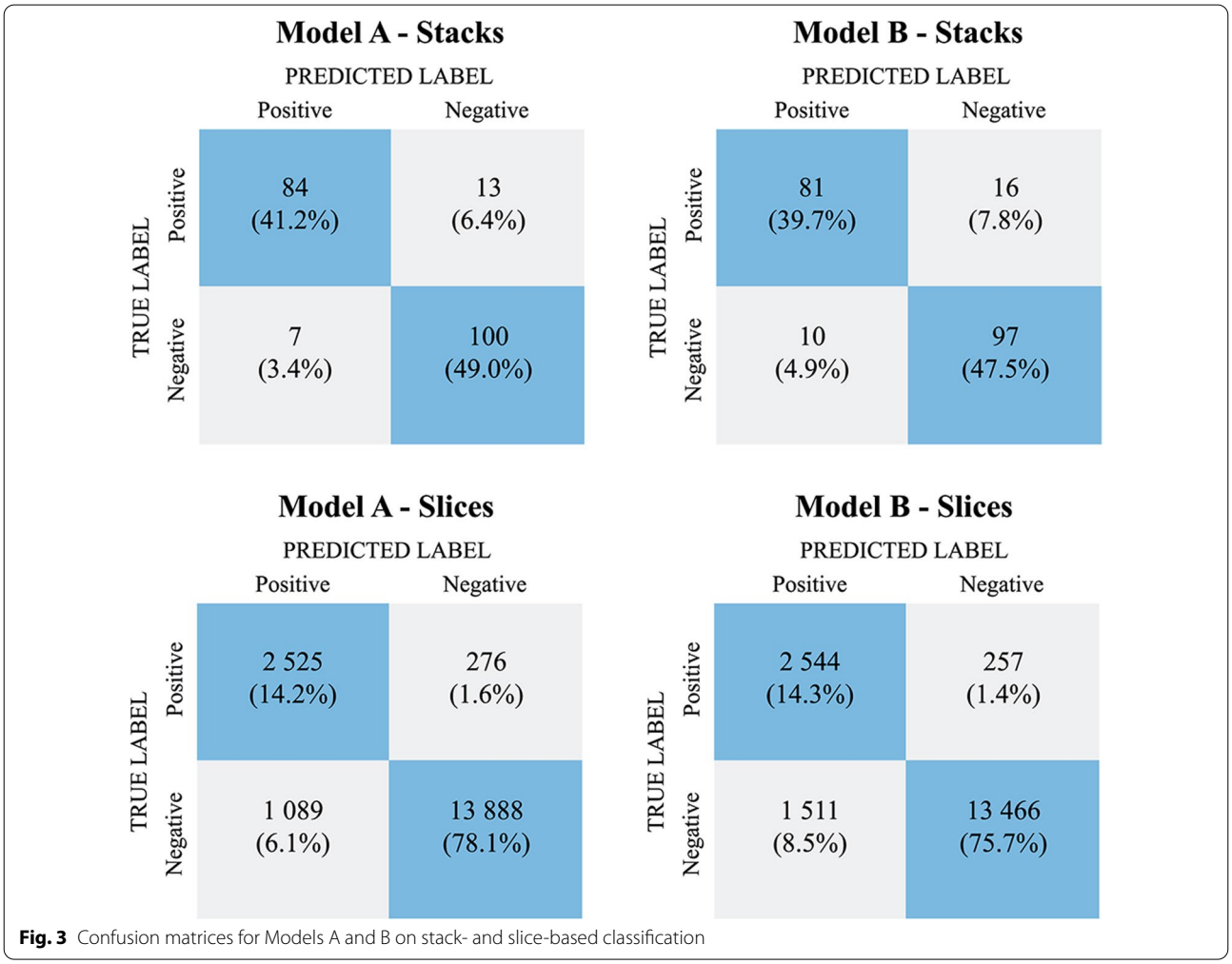

Fig. 3 Confusion matrices for Models A and B on stack- and slice-based classification

Table 2 Performance metrics for Models A and B

Metric	Model A		Model B	
	Stacks	Slices	Stacks	Slices
Accuracy	90.2 (85.1–93.8)	92.3 (91.9–92.7)	87.3 (81.7–91.4)	90.1 (89.6–90.5)
Sensitivity	86.6 (77.8–92.4)	90.1 (89.0–91.2)	83.5 (74.3–90.0)	90.8 (89.7–91.9)
Specificity	93.5 (86.5–97.1)	92.7 (92.2–93.1)	90.7 (83.1–95.2)	89.9 (89.4–90.4)
PPV	92.3 (84.3–96.6)	69.9 (68.3–71.4)	89.0 (80.3–94.3)	62.3 (61.2–64.2)
NPV	88.5 (80.8–93.5)	98.1 (97.8–98.3)	85.8 (77.8–91.4)	98.1 (97.9–98.3)

Results are in percentage (%) with 95% CI

stack-based level. The accuracy, sensitivity, and specificity of Model A were 90.2%, 86.6% and 93.5%, respectively. On the slice-based level, both models also had good performance but their PPV was considerably lower than on the stack-based level (Model A 69.9% vs. 92.3%; Model B 62.3% vs. 89.0%, respectively).

Example images from true positive, false positive and false negative classifications for both models were visually inspected (Fig. 4). We did not notice any major differences between the models. True positives seemed to be more often large proximal emboli, and false findings were more often located in the peripheral parts of the pulmonary arteries. This was expected, because small and peripheral emboli are usually difficult also for radiologists.

Discussion

The main contribution of this study is to demonstrate the development of a deep learning model for automated detection of PE from CTPAs, and to show that this is feasible even with limited data annotation resources. We found that our best model (Model A) achieved an ROC AUC of 0.94, sensitivity of 86.6% and specificity of 93.5% in predicting PE from whole CTPA stacks. We showed that these promising results could be achieved using a weakly labelled training dataset consisting of only 600 CTPAs, which is a relatively small dataset for neural networks.

Our models used less training data, yet performed better than several models presented in previous studies [26, 28, 36]. Our best model also had better specificity with almost similar sensitivity than a multimodal fusion model combining information from CT images and electronic health records, which had a specificity of 90.2% and a

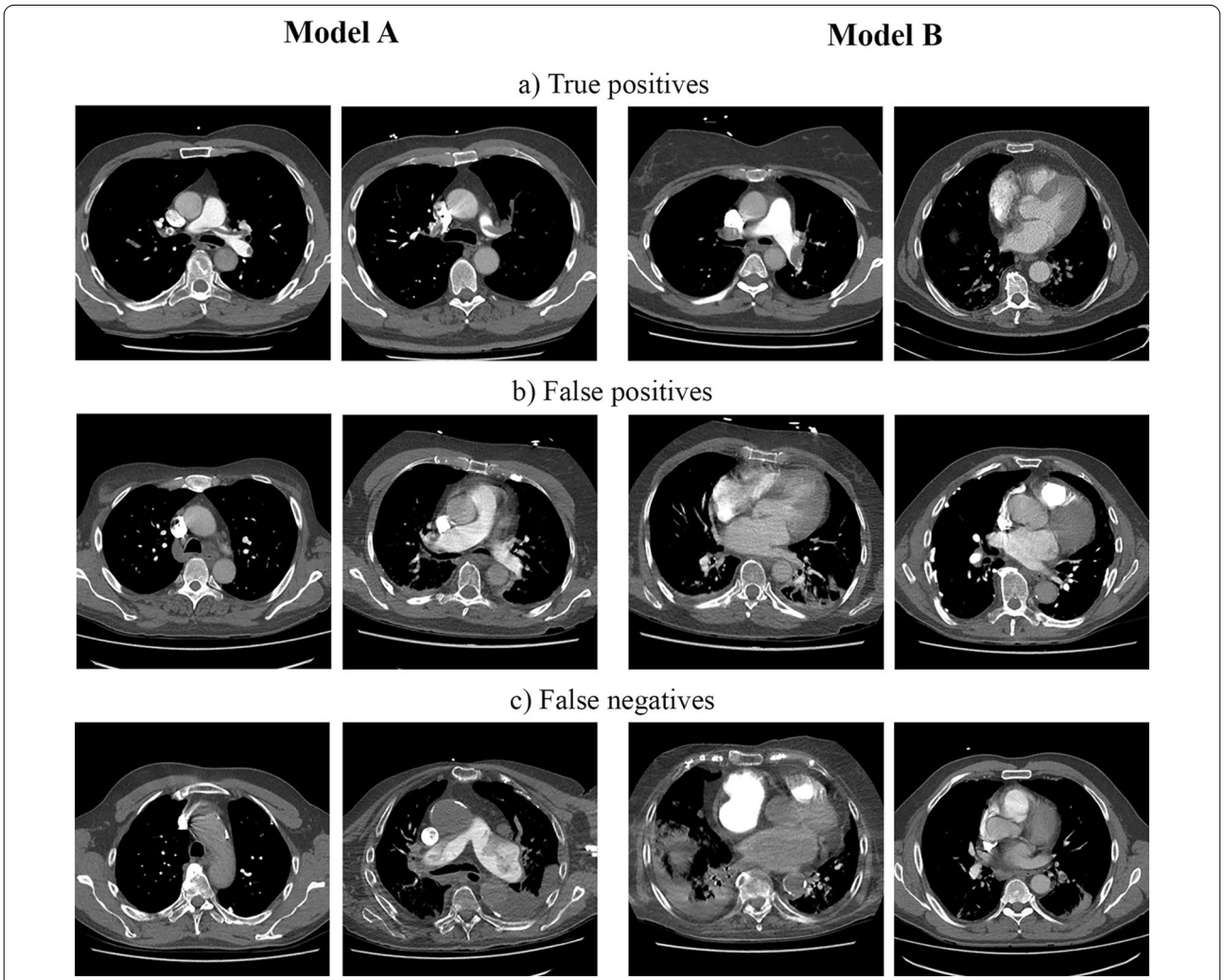

Fig. 4 The figure shows example slices where both the slice-based and stack-based predictions are **a** true positives, **b** false positives or **c** false negatives. In the left panel (the first two images on each row from left to right) are images classified by Model A and in the right panel (the last two images on each row) are images classified by Model B

sensitivity of 87.3% [37]. We achieved only slightly worse specificity with our best model than Aidoc's commercial model (93.5% vs. 95–95.5%, respectively), despite their model having been trained with a considerably larger dataset (600 vs. ~ 28,000 CTPAs) [24, 38]. Models presented by Yang et al. and Tajbaksh et al. aimed to localize each distinct embolus accurately, and their performance was evaluated differently than our model, which is why the models cannot be directly compared [21, 39].

Transfer learning and data augmentation were implemented to compensate for the small training set size. The choice of transfer learning data had only minimal impact on the model performance. Model A performed as well as or slightly better than Model B, even though it was pre-trained on a substantially smaller dataset than Model B (NIH dataset, 100,000 images vs. ImageNet dataset, 14 million images). Since the ImageNet dataset consists of color images, the model pre-trained with it learns many features irrelevant to classifying grayscale CTPA images. NIH chest X-rays are grayscale like target data, and therefore a smaller dataset seems to suffice.

Compared to many previous studies, our approach to annotate the data with only stack- and slice-based binary labels was more time-saving than marking each distinct embolus [15, 21, 39]. Rajan et al. as well as Shi et al. reduced the manual workload in annotation by doing pixel-wise segmentations only on slices at 10 mm intervals and otherwise using CT study-level binary labels [26, 36]. Despite using sparser annotations, our model performed better than their models. Huang et al. used data which was annotated in a similar fashion than in our study [28]. However, they excluded studies with only subsegmental emboli due to their unclear clinical importance, but we felt the need to include them because they represent a substantial part of the realistic data seen in clinical practice.

There was a notable difference between stack and slice level cut-off thresholds selected with the Youden index that needs to be discussed. The Youden index is a measurement for the ROC curve and it is used to select the optimal operating point on the curve. The optimal cut-off threshold corresponding to this operating point depends on the range and distribution of the model output scores, which can vary wildly between models. Without additional calibration, the model output score does not always equal the probability of the predicted class [40]. Also, another reason why the thresholds between the stack and slice levels are not comparable is that the positive class ratios were very different between the levels (48% vs. 16%, respectively).

Our study had certain limitations. First, our models do not produce precise information about the location of emboli, which might hinder the interpretation of model findings. Including localization of emboli in our model would have increased the manual annotation workload considerably. Our model can still pinpoint individual slices that are most likely to contain PE to help radiologists quickly focus on the most probable area. Exact localization of emboli is also not needed in pre-screening CT studies to prioritize the reading list. Our model could help in faster detection of PE positive cases in emergency settings as well as in non-emergency CT imaging where an incidental PE is frequently found, especially in oncologic patients, where there might be a several days delay in reading the CT scans [41, 42]. Second, we only used data from our institution, and the majority of the data was acquired with CT scanners from one vendor. Our model might need further training to perform accurately with data acquired in different institutions or with different CT scanners, as model performance often degrades on data acquired differently (e.g., different CT scanner model, manufacturer or imaging settings) [43]. Third, our model was tested only on a dataset balanced between positive and negative cases. In the clinical setting, the prevalence of positive CTPAS is much lower, varying from less than 10% to 30% [9]. Using balanced datasets is, however, a common practice in medical AI research. In the future, we aim to test the model on a dataset representing a more realistic prevalence.

Conclusions

In conclusion, we developed a deep learning model for automated PE detection which achieved substantial performance results. We showed that these results could be achieved with a relatively small, weakly labelled training set. This demonstrates that it is possible for small research groups and individual hospitals to build well-performing DL models even with limited resources. Our model could be used either as an aid in reading emergency studies to reduce mistakes, or as a pre-screening triage tool to prioritize reading order. In the future, we plan to test the model performance and usability in a clinical setting. We plan to study how the model performs when data acquisition is slightly modified (e.g., updated scanning protocol, or a new scanner), and how much additional training the model requires after each change. We also plan to develop the visualization of the output for better interpretability.

Abbreviations

PE: Pulmonary embolism; CTPA: Computed tomography pulmonary angiogram; CT: Computed tomography; DL: Deep learning; AI: Artificial intelligence; CAD: Computer-aided detection; CNN: Convolutional neural network; LSTM: Long-short term memory network; FP: False positive; FN: False negative; TP: True positive; TN: True negative; ROC: Receiver operating characteristics; PR: Precision–recall; AUC: Area under the curve; PPV: Positive predictive value; NPV: Negative predictive value; CI: Confidence interval.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12880-022-00763-z.

Additional file 1. Model architectures. Detailed model architectures for the CNN model and the CNN + LSTM combination model.

Acknowledgements
Not applicable.

Authors' contributions
All authors contributed to the study conception and design. AV, AK, MN and TM participated in setting up the technical environment and data collection. HH and MN labelled the data. TM created and trained the models. H.H. tested the models, analyzed results and prepared figures. H.H., M.N. and J.H. wrote the main manuscript text. All authors reviewed and approved the manuscript.

Funding
This work was financially supported by Turku University Hospital and the Paulo Foundation. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Availability of data and materials
Image data cannot be publicly shared because of the national legislature on patient data. The source code for image preprocessing and training the models will be published on GitHub (https://github.com/turku-rad-ai/pe-detection) by 31st of March 2022. Otherwise all relevant data are within the manuscript and its supplementary information. Further inquiries should be addressed to Jussi Hirvonen (jussi.hirvonen@utu.fi).

Declarations

Ethics approval and consent to participate
We obtained permission from The Hospital District of Southwest Finland. The study was conducted in accordance with the Declaration of Helsinki. Waiver for written patient consent was not sought from the institutional review board (IRB, called the Ethics Committee of The Hospital District of Southwest Finland), because it is not required by the national legislature for retrospective studies of existing data (e.g. registry data). Institutional review board review (approval or waiver) was not sought, because it is not required by the national legislature for retrospective studies of existing data. Registry studies in Finland are exempted from ethical approval by law, and are only subject to hospital district permission. This is based on the following legislature: Law on the secondary use of social and health data (552/2019), Data protection act (1050/2018), Act on the publicity of the activities of authorities (621/1999), and the EU GDPR (2016/679). Legislature is publicly available at www.finlex.fi (most laws and decrees are only available in Finnish and Swedish). For any additional inquiries, please contact Turku Research Services at www.turkucrc.fi/en/contact.

Consent for publication
Not applicable.

Competing interests
The authors have read the journal's policy and have the following potentially competing interests: T.M. was an employee of Reaktor Innovations Oy at the time of the study. The other authors have no competing interests.

Author details
[1]Department of Radiology, University of Turku and Turku University Hospital, Turku, Finland. [2]Reaktor Innovations Oy, Helsinki, Finland. [3]Auria Clinical Informatics, Turku University Hospital, Turku, Finland. [4]Department of Mathematics and Statistics, University of Turku, Turku, Finland. [5]Auria Biobank, Turku University Hospital, University of Turku, Turku, Finland.

Received: 3 September 2021 Accepted: 21 February 2022
Published online: 14 March 2022

References

1. Silverstein MD, Heit JA, Mohr DN, Petterson TM, O'Fallon WM, Melton LJ. Trends in the incidence of deep vein thrombosis and pulmonary embolism. Arch Intern Med. 1998;158(6):585.
2. Andersson T, Söderberg S. Incidence of acute pulmonary embolism, related comorbidities and survival; analysis of a Swedish national cohort. BMC Cardiovasc Disord. 2017;17(1):155.
3. Oger E. Incidence of venous thromboembolism: a community-based study in Western France. EPI-GETBP Study Group. Groupe d'Etude de la Thrombose de Bretagne Occidentale. Thromb Haemost. 2000;83(5):657–60.
4. Di Nisio M, van Es N, Büller HR. Deep vein thrombosis and pulmonary embolism. Lancet. 2016;388(10063):3060–73.
5. Sherk WM, Stojanovska J. Role of clinical decision tools in the diagnosis of pulmonary embolism. Am J Roentgenol. 2017;208(3):W60-70.
6. Roy P-M, Meyer G, Vielle B, Le Gall C, Verschuren F, Carpentier F, et al. Appropriateness of diagnostic management and outcomes of suspected pulmonary embolism. Ann Intern Med. 2006;144(3):157.
7. Donohoo JH, Mayo-Smith WW, Pezzullo JA, Egglin TK. Utilization patterns and diagnostic yield of 3421 consecutive multidetector row computed tomography pulmonary angiograms in a busy emergency department. J Comput Assist Tomogr. 2008;32(3):421–5.
8. Mountain D, Keijzers G, Chu K, Joseph A, Read C, Blecher G, et al. RESPECT-ED: rates of pulmonary emboli (PE) and sub-segmental PE with modern computed tomographic pulmonary angiograms in emergency departments: a multi-center observational study finds significant yield variation, uncorrelated with use or small PE rates. PLoS ONE. 2016;11(12):e0166483.
9. Dalen JE, Waterbrook AL. Why are nearly all CT pulmonary angiograms for suspected pulmonary embolism negative? Am J Med. 2017;130(3):247–8.
10. McDonald RJ, Schwartz KM, Eckel LJ, Diehn FE, Hunt CH, Bartholmai BJ, et al. The effects of changes in utilization and technological advancements of cross-sectional imaging on radiologist workload. Acad Radiol. 2015;22(9):1191–8.
11. Rohatgi S, Hanna TN, Sliker CW, Abbott RM, Nicola R. After-hours radiology: challenges and strategies for the radiologist. Am J Roentgenol. 2015;205(5):956–61.
12. Hanna TN, Zygmont ME, Peterson R, Theriot D, Shekhani H, Johnson J-O, et al. The effects of fatigue from overnight shifts on radiology search patterns and diagnostic performance. J Am Coll Radiol. 2018;15(12):1709–16.
13. Chartrand G, Cheng PM, Vorontsov E, Drozdzal M, Turcotte S, Pal CJ, et al. Deep learning: a primer for radiologists. Radiographics. 2017;37(7):2113–31.
14. Al-Hinnawi ARM. Computer-aided detection, pulmonary embolism, computerized tomography pulmonary angiography: current status. In: Pamukçu B, editor. Angiography. IntechOpen; 2018.
15. Chan H-P, Hadjiiski L, Zhou C, Sahiner B. Computer-aided diagnosis of lung cancer and pulmonary embolism in computed tomography—a review. Acad Radiol. 2008;15(5):535–55.
16. Wittenberg R, Berger FH, Peters JF, Weber M, van Hoorn F, Beenen LFM, et al. Acute pulmonary embolism: effect of a computer-assisted detection prototype on diagnosis—an observer study. Radiology. 2012;262(1):305–13.
17. van Ginneken B, Schaefer-Prokop CM, Prokop M. Computer-aided diagnosis: how to move from the laboratory to the clinic. Radiology. 2011;261(3):719–32.
18. Lyell D, Coiera E. Automation bias and verification complexity: a systematic review. J Am Med Inform Assoc. 2017;24(2):423–43.
19. Taylor SA, Brittenden J, Lenton J, Lambie H, Goldstone A, Wylie PN, et al. Influence of computer-aided detection false-positives on reader performance and diagnostic confidence for CT colonography. Am J Roentgenol. 2009;192(6):1682–9.
20. Tajbakhsh N, Gotway MB, Liang J. Computer-aided pulmonary embolism detection using a novel vessel-aligned multi-planar image representation and convolutional neural networks. In: Navab N, Hornegger J, Wells W, Frangi A, editors. Medical image computing and computer-assisted intervention—MICCAI 2015. Lecture notes in computer science, vol. 9350. Springer; 2015. p. 62–9.
21. Yang X, Lin Y, Su J, Wang X, Li X, Lin J, et al. A two-stage convolutional neural network for pulmonary embolism detection from CTPA images. IEEE Access. 2019;7:84849–57.

22. Colak E, Kitamura FC, Hobbs SB, Wu CC, Lungren MP, Prevedello LM, et al. The RSNA pulmonary embolism CT dataset. Radiol Artif Intell. 2021;3(2):e200254.
23. Weikert T, Winkel DJ, Bremerich J, Stieltjes B, Parmar V, Sauter AW, et al. Automated detection of pulmonary embolism in CT pulmonary angiograms using an AI-powered algorithm. Eur Radiol. 2020;30(12):6545–53.
24. Buls N, Watté N, Nieboer K, Ilsen B, de Mey J. Performance of an artificial intelligence tool with real-time clinical workflow integration—detection of intracranial hemorrhage and pulmonary embolism. Phys Medica. 2021;83:154–60.
25. Castiglioni I, Rundo L, Codari M, Di Leo G, Salvatore C, Interlenghi M, et al. AI applications to medical images: from machine learning to deep learning. Phys Med. 2021;83:9–24.
26. Rajan D, Beymer D, Abedin S, Dehghan E, Dalca A V, Mcdermott M, et al. Pi-PE: a pipeline for pulmonary embolism detection using sparsely annotated 3D CT images. In: Proceedings of the machine learning for health NeurIPS workshop, PMLR. 2020. p. 220–32.
27. Feng Y, Hao P, Zhang P, Liu X, Wu F, Wang H. Supervoxel based weakly-supervised multi-level 3D CNNs for lung nodule detection and segmentation. J Ambient Intell Humaniz Comput. 2019;2019:1–11.
28. Huang SC, Kothari T, Banerjee I, Chute C, Ball RL, Borus N, et al. PENet—a scalable deep-learning model for automated diagnosis of pulmonary embolism using volumetric CT imaging. NPJ Digit Med. 2020;3(1):61.
29. Han C, Kitamura Y, Kudo A, Ichinose A, Rundo L, Furukawa Y, et al. Synthesizing diverse lung nodules wherever massively: 3D multi-conditional GAN-based CT image augmentation for object detection. In: Proceedings—2019 international conference on 3D vision, 3DV 2019. 2019. p. 729–37.
30. Nakao T, Hanaoka S, Nomura Y, Murata M, Takenaga T, Miki S, et al. Unsupervised deep anomaly detection in chest radiographs. J Digit Imaging. 2021;34(2):418–27.
31. Han C, Rundo L, Murao K, Noguchi T, Shimahara Y, Milacski ZÁ, et al. MADGAN: unsupervised medical anomaly detection GAN using multiple adjacent brain MRI slice reconstruction. BMC Bioinform. 2021;22(2):1–20.
32. Braman N, Beymer D, Dehghan E. Disease detection in weakly annotated volumetric medical images using a convolutional LSTM network. arXiv:1812.01087v1 [Preprint]. 2018. http://arxiv.org/abs/1812.01087.
33. Szegedy C, Ioffe S, Vanhoucke V, Alemi AA. Inception-v4, inception-ResNet and the impact of residual connections on learning. In: Proceedings AAAI conference on artificial intelligence, vol. 31, no. 1. 2017.
34. 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(3):837.
35. Wilson EB. Probable inference, the law of succession, and statistical inference. J Am Stat Assoc. 1927;22(158):209–12.
36. Shi L, Rajan D, Abedin S, Yellapragada MS, Beymer D, Dehghan E. Automatic diagnosis of pulmonary embolism using an attention-guided framework: a large-scale study. In: Proceedings of machine learning research. PMLR; 2020. p. 743–54.
37. Huang SC, Pareek A, Zamanian R, Banerjee I, Lungren MP. Multimodal fusion with deep neural networks for leveraging CT imaging and electronic health record: a case-study in pulmonary embolism detection. Sci Rep. 2020;10(1):22147.
38. 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(1):55–9.
39. Tajbakhsh N, Shin JY, Gotway MB, Liang J. Computer-aided detection and visualization of pulmonary embolism using a novel, compact, and discriminative image representation. Med Image Anal. 2019;58:101541.
40. Guo C, Pleiss G, Sun Y, Weinberger KQ. On Calibration of modern neural networks. In: Proceedings of machine learning research. PMLR; 2017. p. 1321–30.
41. Weisberg EM, Chu LC, Fishman EK. The first use of artificial intelligence (AI) in the ER: triage not diagnosis. Emerg Radiol. 2020;27(4):361–6.
42. Klok FA, Huisman MV. Management of incidental pulmonary embolism. Eur Respir J. 2017;49(6):1700275.
43. Subbaswamy A, Saria S. From development to deployment: dataset shift, causality, and shift-stable models in health AI. Biostatistics. 2020;21(2):345–52.

Publisher's Note

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

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions