Pulmonary Nodule Classification in Lung Cancer from 3D Thoracic CT Scans Using `fastai` and `MONAI`

Satheshkumar Kaliyugarasan[1,3**], Arvid Lundervold[2,3], Alexander Selvikvåg Lundervold[1,3 ** *]

[1] Dept. of Computer Science, Electrical Engineering and Mathematical Sciences, Western Norway University of Applied Sciences, Bergen (Norway)
[2] Dept. of Biomedicine, University of Bergen (Norway)
[3] Mohn Medical Imaging and Visualization Centre, Department of Radiology, Haukeland University Hospital, Bergen (Norway)

[**] These authors contributed equally to the current work

Received 9 November 2020 | Accepted 31 March 2021 | Published 4 May 2021

Abstract

We construct a convolutional neural network to classify pulmonary nodules as malignant or benign in the context of lung cancer. To construct and train our model, we use our novel extension of the `fastai` deep learning framework to 3D medical imaging tasks, combined with the `MONAI` deep learning library. We train and evaluate the model using a large, openly available data set of annotated thoracic CT scans. Our model achieves a nodule classification accuracy of 92.4% and a ROC AUC of 97% when compared to a "ground truth" based on multiple human raters subjective assessment of malignancy. We further evaluate our approach by predicting patient-level diagnoses of cancer, achieving a test set accuracy of 75%. This is higher than the 70% obtained by aggregating the human raters assessments. Class activation maps are applied to investigate the features used by our classifier, enabling a rudimentary level of explainability for what is otherwise close to "black box" predictions. As the classification of structures in chest CT scans is useful across a variety of diagnostic and prognostic tasks in radiology, our approach has broad applicability. As we aimed to construct a fully reproducible system that can be compared to new proposed methods and easily be adapted and extended, the full source code of our work is available at https://github.com/MMIV-ML/Lung-CT-fastai-2020.

Keywords

Convolutional Neural Networks, Fastai, Lung Cancer, Thoracic CT.

DOI: 10.9781/ijimai.2021.05.002

I. Introduction

Using convolutional neural networks is well-known to result in powerful tools to analyse medical images, across a variety of important applications [1], [2]. This approach to medical image analysis can lead to valuable insights and assistance in imaging diagnostics. The path from research to clinical practice is however slow and arduous, perhaps more so than is generally thought [2], [3]. But the number of software solutions on the market, with regulatory approval and aimed at diagnostic support, is growing, along with their adoption in hospital workflows.

In radiology, the computed tomography (CT) imaging modality is currently experiencing the highest impact of deep learning-based solutions. CT uses computer-processed combinations of many X-ray measurements taken from different angles to produce cross-sectional digital images (virtual slices) of specific regions or organs within the human body. This allows for non-invasive inspection of disease processes or lesions. Another prominent and widespread imaging modality is magnetic resonance imaging (MRI). It is based on quite different physical principles (nuclear spins in magnetic fields, spin excitation by application of radio-frequency pulses, magnetic resonance, and tissue specific and disease-related magnetization and relaxation phenomena) and enables exploitation of a large collection of measurement techniques and contrast mechanisms. Compared to CT, MRI examinations are generally more expensive, more time-consuming and less available. The signal properties are also more complex and typically multi-parametric, and proper interpretation puts high demands on radiologists' specialized training and experience. This partly explain why CT is more heavily used in daily routine radiology, and also why it is a popular target for the med-ical machine learning community [4].

Identifying and assessing structures in the lung from thoracic CT scans (chest CT) is a crucial task across multiple diseases involving the lungs and upper abdomen, e.g. lung cancer, chronic lung disease and pneumonia. Computer-aided diagnostic tools addressing chest CT is therefore an important area in medical imaging[1].

The diagnosis and follow-up of lung cancer patients using chest CT requires the identification of malignant tumors appearing as

* Corresponding author.
E-mail addresses: sathiesh.kumar.kaliyugarasan@hvl.no (S. Kaliyugarasan), arvid.lundervold@uib.no (A. Lundervold), allu@hvl.no (A.S. Lundervold).

[1] An area of particular relevance at the time of writing is the viral pneumonia caused by SARS-CoV-2 ([5], [6]).

pulmonary nodules (i.e. spots on the lungs). Distinguishing benign and malignant nodules is difficult, as the differences can be subtle and the malignancy potential is highly variable [7], but such assessment forms an important source of information for diagnosis and evaluation of progression and treatment responses. Indications of lung cancer can also appear as incidental findings on CT scans. As chest CT is widely used across a range of diseases and injuries, this represents an additional challenge for radiologists.

II. Related Work

Multiple studies have investigated how CNNs can be used in the context of lung cancer. Two recent and quite comprehensive reviews are [8], [9]. Below we highlight two illustrative examples of recent, related work.

In [10], the authors constructed an end-to-end system based on three 3D CNNs for the localization and categorization of lung cancer risk, using low-dose CT images as inputs. They achieved a test set ROC AUC of 94.4% using data from the National Lung Cancer Screening Trial (NLST), and a ROC AUC of 95.5% on an independent data set collected at Northwestern Medicine. A retrospective reader study was conducted, in which their model outperformed six experienced US board-certified radiologists. Their system had four main components: (i) a Mask R-CNN for instance segmentation used to produce lung segmentation masks; (ii) a 3D RetinaNet CNN trained to output ROIs around possible cancer lesions; (iii) a 3D version of Inception V1 trained to predict cancer diagnosis within one year directly from CT volumes; (iii) a CNN classifier trained on features extracted from the detected ROIs as well as features extracted from the volume model, outputting malignancy scores for each ROI. Their study was based on a combination of publicly available data from LUNA, LIDC and NLST, in combination with a large data set sourced from Northwestern Medicine that is not publicly available. The source code used in their work is not publicly available.

In [11], the authors construct *DeepLung*, a "cancer diagnosis system" based on two 3D CNNs that perform lung nodule detection and binary classification (benign vs. malign), respectively. For nodule detection they constructed a 3D Faster R-CNN with dual-path blocks and a similar encoder-decoder structure to the U-Net of [12], obtaining a FROC (Free Response Operating Characteristic) score of 84.2% on the LUNA16 data set [13] using a 10-fold patient-level cross-validation split. Their nodule classification model consisted of a 3D dual-path network extracting classification features, and a gradient boosting machine trained on the extracted features combined with raw nodule CT pixels and nodule size. They achieved a classification accuracy of 90.44%on the LIDC-IDRI data set using the same cross-validation approach as in LUNA16. The source code is available at https: //github.com/wentaozhu/DeepLung.

III. Main Contributions

Motivated by a lack of a common set of training data for machine learning models for lesion malignancy classification in the literature and what we see as important missing elements in how most CNNs for 3D medical imaging tasks are trained, our objectives are the following: (i) bring a set of techniques for training CNNs that have been shown to be highly impactful for 2D image classification to 3D by extending and incorporating ideas from the popular `fastai` library, and (ii) to provide a reproducible setup of data and model evaluation that can be used by other researchers aiming to train models to perform lung nodule classification. Our main contributions are:

1. We preprocessed and prepared the comparably large and well-annotated LIDC-IDRI data set (Section IV) for use in a binary malignancy prediction task, taking care to set aside a separate test set consisting of particularly well-characterized patients.
2. We constructed and trained a three-dimensional CNN using our novel extension to 3D of the `fastai` [14] deep learning library, combining it with features from `MONAI` (https://github.com/Project-MONAI/MONAI[2]), obtaining results comparable to the state-of-the-art in nodule classification and patient-level cancer diagnoses for the LIDC-IDRI data set.
3. We investigated the malignancy predictions by integrating a 3D version of gradient-weighted class activation mapping (Grad-CAM) [16] in our framework, enabling some element of *explainable AI* [17].
4. To ensure reproducibility and to ease further extensions or adaptions of our approach, we have made the source code openly available under a permissive open source license at https://github.com/MMIV-ML/Lung-CT-fastai-2020, in a tutorial-like Jupyter Notebook [18] that step through the process from data loading to result interpretations.

IV. Methods and Materials

A. Data Set

Using supervised learning with CNN models requires large amounts of labelled training data. For pulmonary nodule analysis, the data is typically obtained by manually labelling nodule locations and outlining lesions on CT images, a costly and hard to scale process hampered by intra- and inter-rater variability. Nevertheless, reasonably large annotated data sets with benign and malignant pulmonary nodules have been made openly available for researchers, reducing the entry price and increasing the pace of new research.

We used the Lung Image Database Consortium image collection (LIDC-IDRI), consisting of diagnostic and clinical lung cancer screening thoracic computed tomography (CT) scans with marked-up annotated lesions [19][3]. The images were extracted from the picture archiving and communication systems (PACS) of seven different institutions and anonymized in accordance with HIPAA guidelines. The data collection was approved by the local IRBs of the seven participating LIDC-IDRI institutions. To each image there is associated the results of a two-stage anno-tation process involving four experienced thoracic radiologists. First, in a blinded-read phase, each radiologist independently reviewed the CT scans, marking lesions belonging to one of three categories (*nodule ≥ 3 mm*, *nodule < 3 mm*, and *non-nodule ≥ 3 mm*), where the concept of "nodule" refers to a focal abnormality[4]. Then each radiologist (among a total of 12 radiologists coming from altogether five LIDC-IDRI institutions) assessed independently and subjectively each *nodule ≥ 3 mm* for characteristics such as subtlety, internal structure, spiculation, lobulation, shape (sphericity), solidity, margin, and likelihood of malignancy. Each such nodule, having (by its size) a greater probability of malignancy than lesions in the other two categories, was marked regardless of presumed histology, e.g. a primary lung cancer, metastatic disease, a noncancerous process, or indeterminate in nature.

By design, reader consistency studies are not possible with the LIDC-IDRI data set as the order of the readers varies from instance to instance. However, the marks from up to four readers for a given

[2] Originally, we developed our extension of `fastai` and `MONAI` for 3D MRI of the head, as a tool for the estimation of brain age from MRI recordings (unpublished work and [15]) indicating our framework's general utility.

[3] See also https://wiki.cancerimagingarchive.net/display/Public/LIDC-IDRI

[4] Some radiologists will argue that these three lesion categories could be somewhat artificial relative to clinical practice.

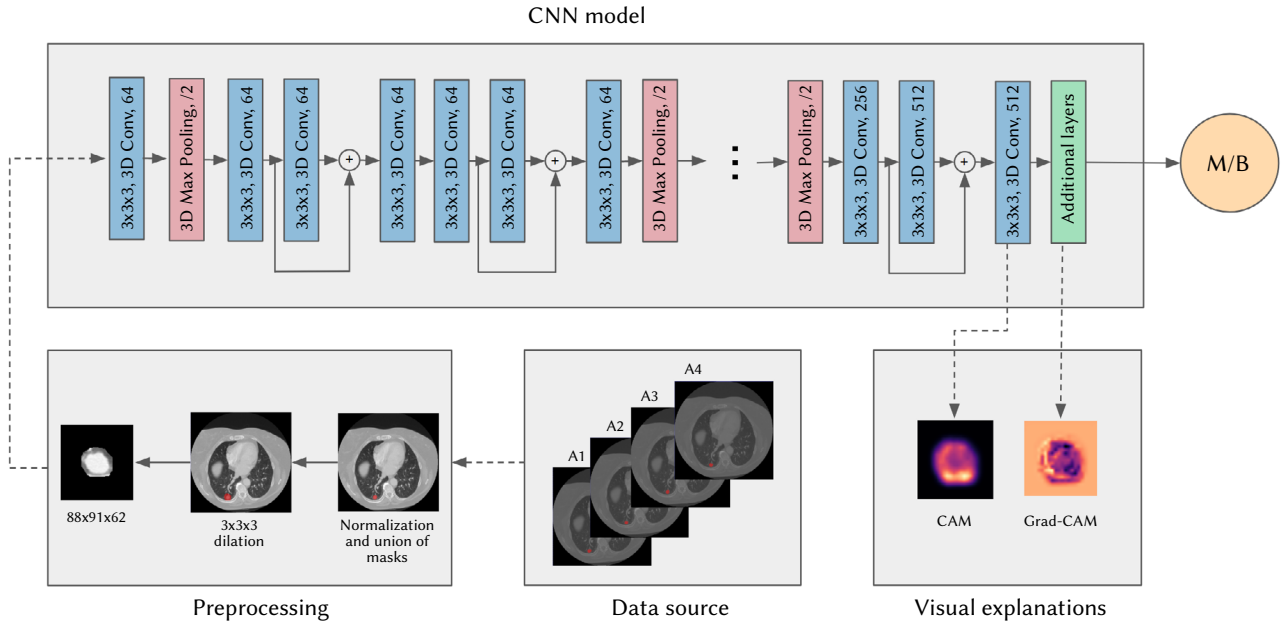

Fig. 1. The annotated images from our data source, LIDC-IDRI, are preprocessed by extracting 3D regions of interests around each of the nodules by taking the union of all the masks provided by expert annotations (e.g. A1−A4), before dilating the image slightly to capture some of the nodule surroundings. Using the expert assessment of malignancy, the resulting nodule images are used to train a 3D CNN model. This results in a nodule classification model with binary output: malignant (M) or benign (B). Our 3D implementation of class activation maps provides a visual explanation, here shown as a pair of 2D slices, indicating areas impacting our model's nodule classification decision. For further details, see the text and the accompanying code repository: https://github.com/MMIV-ML/Lung-CT-fastai-2020

lesion, using a five-point scale (a low score denoted likely benign nodule, a high score likely malignant), makes it possible to assess different degrees of reader agreement. Assessing inter-rater variability is very important to gauge the performance of systems aiming to automate the process. We therefore made an analysis of inter-rater variability regarding the "likelihood of malignancy" characteristic using the *Krippendorff's alpha* coefficient [20].

In our study, we have used a total of 2662 annotated nodules that were annotated as *nodule* $\geq$ *3 mm* by at least one radiologists, collected from clinical thoracic CT scans of 1018 patients in the LIDC-IDRI data set.

B. Preprocessing

The voxels in a 3D CT recording are displayed in terms of relative radiodensity. More specifically, the signal intensities or attenuations in CT are expressed in Hounsfield units (HU). This is based on a linear transformation of the original attenuation coefficients in which the radiodensity of distilled water has HU = 0 and the radiodensity of air is set to HU = −1000. According to this HU scale, lung parenchyma is in the range [−700, −600], fat is [−120, −90], lymph nodes [+10, +20], and blood [+13, +50], to mention a few relevant tissue types. In our CT data we considered voxels within a HU-range of [−1200, +600], and voxel values were normalized to the interval [0, 1] according to the transformation $x'' \mapsto x'' : x' = (x + 1200)/(1200 + 600)$; $x'' = 0$ if $x' < 0$, $x'' = 1$ if $x' > \ldots 1$, else $x'' = x'$.

For each CT scan of a subject, we collected all the radiologists segmentation masks. To ensure that we captured entire nodules we took the union of the masks. To make some of the surrounding context of each nodule available for the classification model, we dilated the resulting mask by adding 3 voxels to its boundary. The data set used to construct and evaluate our models was the constructed by applying the masks to the corresponding normalized CT and cropping to a cube containing the nodules. This gave us a total of 2662 3D images containing nodules. See Fig. 1 for an illustration of the preprocessing process.

We extracted each of the radiologists' subjective assessments of malignancy likelihood and computed the median scores across the readers for each nodule. If the median score for a nodule was < 3 we marked it as *benign*, if $> \ldots 3$ as *malignant*. The nodules with median score 3 (indeterminant) were dropped from our data set. This gave us a total of 1106 benign nodules and 525 malignant.

C. Our `fastai` Extension and the 3D CNN Architecture

Our work is based on a combination of the `MONAI` deep learning framework and our own extension of the powerful `fastai` library built on top of PyTorch [14]. We have added functionality to support the construction, training and evaluation of three-dimensional convolutional neural networks, tailored for medical imaging-specific problems and file formats. In short, we have extended `fastai` to support 2D and 3D MRI and CT images by constructing new data loaders and data augmentation capabilities, and enabled the use of custom 3D CNNs while still supporting the highly impactful training techniques of `fastai`. This includes the learning rate finder [21] to find the optimum learning rate and the one-cycle learning rate policy (i.e. specific learning rate changes during the training, related to the concept of super-convergence [22], [23]).

The architecture of our 3D CNN is shown in Fig. 1. Each convolutional layer in our network consists of $3 \times 3 \times 3$ convolutions, followed by a batch normalization layer [24] and a rectified linear unit (ReLU) layer [25]. We add residual connections after each second convolutional layer. Each down-sampling block has a two-stride $2 \times 2 \times 2$ max-pooling layer.

To enable *discriminative learning rates*, i.e. different learning rates for different parts of the network, we divide the network into two layer groups: convolutional layers and additional layers. This also allow us to do gradual unfreezing, and eases the potential re-use of trained weights from the early layers for other tasks (i.e. *transfer learning*).

D. Training and Evaulation

To evaluate and get a robust estimate of our model's performance, we selected all the subjects in the LIDC-IDRI data set that have corresponding patient-level diagnoses as our test set (99 subjects, 238 nodules). The remaining data were divided into a training set (526 subjects, 1140 nodules) and a validation set (90 subjects, 255 nodules), using stratified sampling and no patient overlap between the sets. In order to deal with imbalanced classes in the training set (802 benign, 338 malignant), we over-sampled the malignant class by duplicating each sample.

Before feeding the images into the network, each image was padded to have the same volume dimension as the largest volume data × a scaling factor. We used data parallelism to train our model on four NVIDIA Tesla V100 32GB GPUs. Our training process was composed of two phases:

- Training a model on $44 \times 46 \times 31$ volumes, with weights randomly initialized (He initialization [26]).
- Training a final model on $88 \times 91 \times 62$ volumes, with weights initialized by copying the weights of the previous model.

This approach is known as progressive image resizing [27], a technique used to both reduce training time and to increase model performance. In our case, we found that it improved the accuracy on the validation set by almost two percentage points.

Our model was trained end-to-end in mixed precision [28] using the Adam optimizer [29]. The base value for the cyclic learning rate in the final model was set to 6×10^{-4} for frozen layers and 5×10^{-5} after unfreezing the layers, with learning rates for earlier layers scaled down by a factor of 20. We trained the model using a batch size of 128. For data augmentation we used random scaling with a factor from 1.0 to 1.1 and random rotation by an angle in the range [-35, 35]. As the geometry of the nodules can contain information about their malignancy, we only used shape-preserving morphisms. For regularization, we used a weight decay rate of 0.01 and a dropout ratio of 0.4, selected based on the performance on the validation data. Our final model was trained on the combined training and validation data for a few epochs, with a small cyclic learning rate, to also make use of the information contained in the validation data and its labels during model training.

E. Explainable AI and Class Activation Maps

As deep learning models are highly complex hierarchical objects with enormous amounts of parameters, there is an inherent "black-boxiness" to them. As they are increasingly being implemented across the medical imaging and decision making domains, this raises both technical challenges (how to open the black box?) and ethical conundrums (when is it OK to use predictions you cannot fully understand?). Using our extension of `fastai` we can produce what are called class activation maps (CAM) [30] and gradient-weighted class activation mapping (Grad-CAM) [16]. These are heat maps that can be used to indicate the importance of regions of an image for the model's classification, providing a relatively simple way to gain some explainability for image classification models, and potentially also to gain useful insights into the data used to construct the model.

CAM generates heat maps from the adaptive pooling layer, where the average of each cell across every channel is calculated. On the other hand, Grad-CAM uses the gradient information flowing into the last convolutional layer to produce heat maps, making it applicable to any CNN architecture.

A problem with these methods is that the resolution of the heat maps are the same size as the final convolutional layer. This means that we have to upsample them to the same size as the input images to highlight class-specific image regions. To mitigate this problem one can remove the pooling layers, but this will require more computational power due to larger spatial dimensions. In addition, overfitting is more likely to occur, which might reduce the performance of the network.

V. Experimental Results

Our test set consisted of 238 nodules from 99 subjects, 146 benign and 92 malignant. There were no overlap among train and test subjects. In addition to predicting nodule malignancy, we further investigated the models predictive capabilities by using the ground truth labels of patient diagnosis available in the LIDC-IDRI data set. The 99 patients in our test set were all diagnosed as either *malignant* or having *benign or non-malignant disease*. If one or more nodules from a patient was predicted to be *malignant*, we predicted malignant, else *benign or non-malignant disease*.

The results are displayed in Table I, Fig. 2 and Fig. 3.

TABLE I. Performance Metrics of Our Binary Classifier Predicting Single Nodules (N=238) and Patient Cases (N=99) in the Test Data Set: Accuracy (ACC), Precision (PREC) and Recall (REC). For the Patient Predictions We Give Performance Values Separately for Those Obtained by Our Model (CNN) and for Those Obtained by the Median Radiologist Assessments (Rad)

Classification task						
Nodule classification (%)			Patient classification (%)			
ACC	PREC	REC	Source	ACC	PREC	REC
92.4	85.6	96.7	CNN	75	86.8	78.7
			Rad	70	88.1	69.3

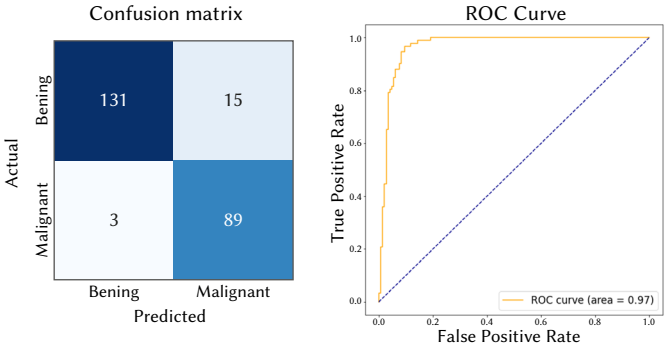

Fig. 2. Predicting the "likelihood of malignancy" in the test set of 238 nodules. (a) Confusion matrix. (b) ROC curve.

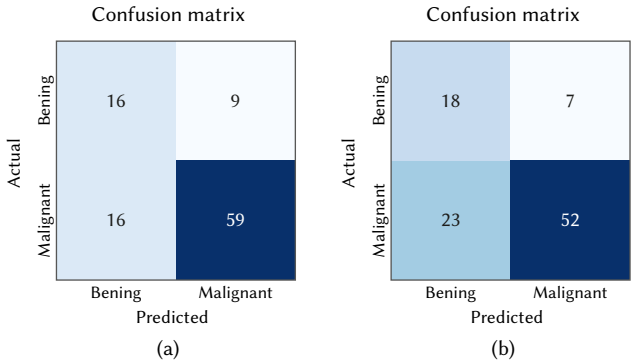

Fig. 3. Confusion matrices: (a) for the CNN predictions, (b) for the median malignancy scores by the radiologists. Note the additional cancer diagnoses captured by our CNN.

The mean score assigned to each nodule classified correctly as benign was 1.91 (SD 0.56) and as malignant 4.18 (SD 0.56). The nodules

misclassified as benign had a mean score of 3.5 (SD 0.0) and those misclassified as malignant had a mean score of 2.23 (SD 0.4).

To assess the inter-rater variability and how the model compares to the human raters, we calculated the *Krippendorff's alpha* coefficient [20] for the 238 nodules. Krippendorff's alpha applies to any measurement level, can handle various number of raters and is invariant to the permutation and selective participation of raters. It also ignores missing data entirely. The independent and interchangeable rater panel per unit consisted of one to five radiologists using scores $s \in \{1$ (*most likely benign*), $2, \ldots, 5$ (*most likely malignant*)$\}$.[5] We note that the agreement on these subjective assessments were not very high. For the Krippendor's $\alpha \in [0, 1]$, $\alpha = 0$ is absence of agreement, and $\alpha = 1$ is perfect agreement. For the "likelihood of malignancy" we found Krippendorff's $\alpha = 0.49$, $\mathrm{CI}_{.025,.975} = [0.43, 0.54]$ (obtained by bootstrapping), indicating poor agreement among the raters.

The Krippendorff's alpha coefficient (in this case equivalent to Cohen's Kappa score) comparing the model's rating to the ground truth (determined by the median radiologist rating) was 0.84, $\mathrm{CI}_{.025,.975} = [0.78, 0.91]$.

The Krippendorff's alpha of the binary assessments of malignancy among the radiologists was $\alpha = 0.58$. By including the independent, CNN-based rater we obtained an increased alpha score to 0.68, indicating the usefulness of including this rater in the assessment of each nodule.

We applied our class-activation map approach described in Section IV.E to a selection of test nodules and CNN predictions. In general, getting better insight into CNN behavior and model predictions, both in cases where it classifies correctly and in cases where it fails, is of interest for several reasons. The class activation maps can provide discriminative information in image regions or part of the lesion being used by the model to predict the class label for the particular instance. This ability can at best introduce interpretability and trust in the model, or facilitate exploration and discovery of new features (image biomarkers) that might have a mechanistic relation to the disease process or disease state. In the present study, we did not fully explore the CAM approach or its potential by involving radiologists or pathologists, and the CAM results are anecdotal and not rigorously validated.

Some of the generated heat maps from our CNN model are presented in Fig. 4. By examining the malignant nodules (nodule 1 and nodule 2) and their corresponding heat maps, we can see that the lesion brims are highlighted, indicating that these regions are most important for the predictions. This might reflect typical malignant tumor growth characterized by central necrosis and viable tumor cells in a well-vascularized periphery. Another interesting finding was nodule 4, a nodule rated benign but classified as malignant by our model. This nodule was assessed by two radiologists deciding malignancy likelihood 2 and 3, respectively (i.e. towards benign), whereas the biopsy done on this nodule concluded that it was a malignant primary lung tumor.

VI. Discussion and Perspectives

We have addressed an important field of oncological radiology: the use of 3D CT scans to characterize focal lung lesions as benign or malignant. Using a large multi-center collection of well-organized CT examinations we constructed and trained a 3D CNN model to perform nodule malignancy classification.

[5] The "likelihood of malignancy" characteristic is particularly subjective since the radiologists were not provided with any clinical information about the patients. As a general scaling guide, the likelihood of malignancy was rated under the assumption that the lesion was associated with a 60-year-old male smoker.

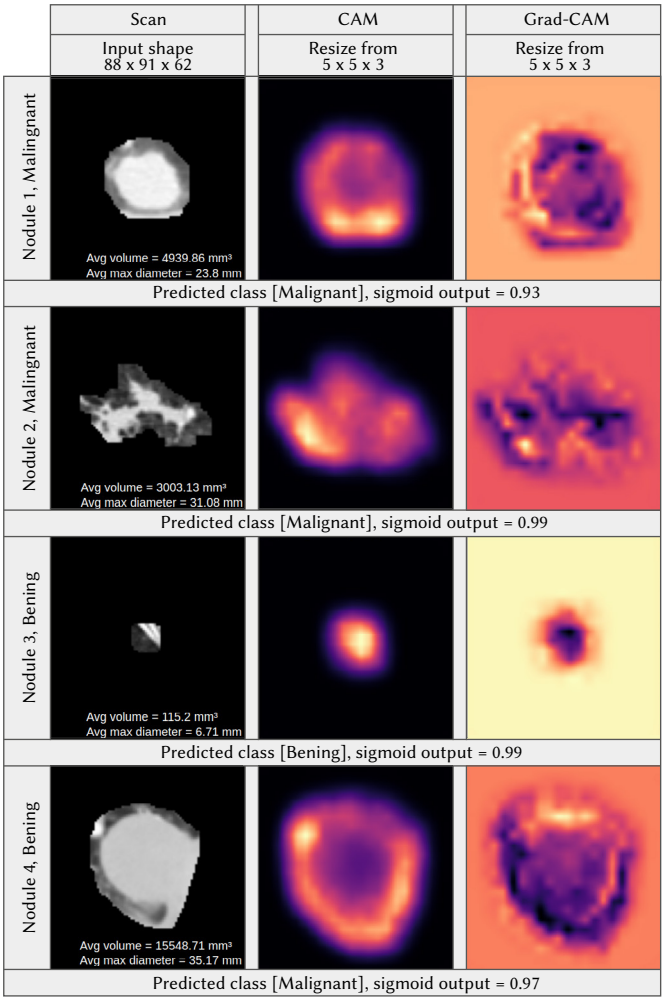

Fig. 4. Examples of CAMs and Grad-CAMs for our model and the corresponding predictions and sigmoid outputs for the respective classes on a selection of four test set nodules.

Because CNNs automatically extract features from data, both interpretation and troubleshooting are more difficult compared to traditional machine learning models. For domains like medical diagnosis, where decision confidence is crucial, it is important to make sure that the results make sense. Otherwise, these models can easily end up performing worse than expected when used for real-world decision making. CAMs and Grad-CAMs generated from CNN models can be valuable for developers to gain some visual insights into models decision processes, helpful to identify data leakage, structural bias and for more comprehensive performance evaluation. In addition, the heat maps have the potential to detect local features that can be used as a biomarker for identifying malignant nodules. We implemented and explored these simple "explainable AI" techniques, assessing successful and unsuccessful nodule predictions.

Our model had a test set accuracy of 92.4% on the per-nodule malignancy classification task. On the patient-level malignancy classification task, our model had an accuracy of 75%. This gave an indication of the network's ability to pick up patterns corresponding to real nodule malignancy. As shown in Fig. 4, class activation maps can highlight regions of particular relevance for the nodule classifications, further indicating that the reasonableness of the features picked up by our CNN model.

In further work we will use the present system as a component in a detection + classification framework, obviating the need for manual annotation steps. We will test the system in the established

radiology research workflow at our hospital, through our "research PACS and RIS" system, enabling us to run arbitrary algorithms on locally recorded images. Such real-world testing is crucial to uncover and surmount the many technical obstacles faced when attempting to bring deep learning-based systems into practice [3]. Especially as it facilitates prospective investigations of the effect of combining the algorithm's predictions with radiologists' expertise, arguably the most interesting next step for research into applications of deep learning in medicine.

Acknowledgment

This work was supported by the Trond Mohn Research Foundation, grant number BFS2018TMT07.

References

[1] A. S. Lundervold, A. Lundervold, "An overview of deep learning in medical imaging focusing on MRI," *Zeitschrift für Medizinische Physik*, vol. 29, no. 2, pp. 102–127, 2019.

[2] X. Liu, L. Faes, A. U. Kale, S. K. Wagner, D. J. Fu, A. Bruynseels, T. Mahendiran, G. Moraes, M. Shamdas, C. Kern, *et al.*, "A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis," *The lancet digital health*, vol. 1, no. 6, pp. e271–e297, 2019.

[3] M. Nagendran, Y. Chen, C. A. Lovejoy, A. C. Gordon, M. Komorowski, H. Harvey, E. J. Topol, J. P. Ioannidis, G. S. Collins, M. Maruthappu, "Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies," *BMJ*, vol. 368, 2020.

[4] M. Brown, P. Browning, M. W. Wahi-Anwar, M. Murphy, J. Delgado, H. Greenspan, F. Abtin, S. Ghahremani, N. Yagh-mai, I. da Costa, *et al.*, "Integration of chest ct cad into the clinical workflow and impact on radiologist efficiency," *Academic radiology*, vol. 26, no. 5, pp. 626–631, 2019.

[5] C. Bao, X. Liu, Z. H., Y. Li, J. Liu, "Coronavirus Disease 2019 (COVID-19) CT Findings: A Systematic Review and Meta-analysis," *J Am Coll Radiol*, vol. Mar 25, pp. 1–9, 2020.

[6] G. D. Rubin, C. J. Ryerson, L. B. Haramati, N. Sverzellati, P. Kanne, S. Raoof, N. W. Schluger, A. Volpi, J.-J. Yim, B. Martin, *et al.*, "The role of chest imaging in patient management during the covid-19 pandemic: a multinational consensus statement from the fleischner society," *Chest*, vol. 158, no. 1, pp. 106–116, 2020.

[7] P. de Groot, B. Carter, G. F. Abbott, C. C. Wu, "Pitfalls in chest radiographic interpretation: blind spots," in *Seminars in roentgenology*, vol. 50, 2015, pp. 197–209, WB Saunders Ltd.

[8] D. Li, B. Mikela Vilmun, J. Frederik Carlsen, E. Albrecht-Beste, C. Ammitzbøl Lauridsen, M. Bachmann Nielsen, Lindskov Hansen, "The performance of deep learning algorithms on automatic pulmonary nodule detection and classification tested on different datasets that are not derived from LIDC-IDRI: a systematic review," *Diagnostics*, vol. 9, no. 4, p. 207, 2019.

[9] A. Halder, D. Dey, A. K. Sadhu, "Lung Nodule Detection from Feature Engineering to Deep Learning in Thoracic CT Images: a Comprehensive Review," *Journal of Digital Imaging*, pp. 1–23, 2020.

[10] D. Ardila, A. P. Kiraly, S. Bharadwaj, B. Choi, J. J. Reicher, Peng, D. Tse, M. Etemadi, W. Ye, G. Corrado, *et al.*, "End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography," *Nature medicine*, vol. 25, no. 6, pp. 954–961, 2019.

[11] W. Zhu, C. Liu, W. Fan, X. Xie, "Deeplung: Deep 3d dual path nets for automated pulmonary nodule detection and classification," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2018, pp. 673–681, IEEE.

[12] O. Ronneberger, P. Fischer, T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*, 2015, pp. 234–241, Springer.

[13] A. A. A. Setio, A. Traverso, T. De Bel, M. S. Berens, C. van den Bogaard, P. Cerello, H. Chen, Q. Dou, M. E. Fantacci, B. Geurts, *et al.*, "Validation, comparison, and combination of algorithms for automatic detection of pulmonary nodules in computed tomography images: the LUNA16 challenge," *Medical image analysis*, vol. 42, pp. 1–13, 2017.

[14] J. Howard, S. Gugger, "fastai: A Layered API for Deep Learning," *Information*, vol. 11, no. 2, p. 108, 2020.

[15] S. Kaliyugarasan, A. Lundervold, A. Lundervold, *et al.*, "Brain age versus chronological age: A large scale mri and deep learning investigation," 2020, European Congress of Radiology-ECR 2020.

[16] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

[17] D. Gunning, "Explainable Artificial Intelligence (XAI)," *Defense Advanced Research Projects Agency (DARPA), nd Web*, vol. 2, 2017.

[18] T. Kluyver, B. Ragan-Kelley, F. Pérez, B. Granger, M. Bussonnier, J. Frederic, K. Kelley, J. Hamrick, J. Grout, S. Corlay, *et al.*, "Jupyter Notebooks — a publishing format for reproducible computational workflows," *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, p. 87, 2016.

[19] S. G. Armato III, G. McLennan, L. Bidaut, M. F. McNitt-Gray, C. R. Meyer, A. P. Reeves, B. Zhao, D. R. Aberle, C. I. Henschke, E. A. Hoffman, *et al.*, "The lung image database consortium (LIDC) and image database resource initiative (IDRI): a completed reference database of lung nodules on CT scans," *Medical physics*, vol. 38, no. 2, pp. 915–931, 2011.

[20] K. Krippendorff, "Reliability in Content Analysis: Some Common Misconceptions and Recommendations," *Human Communication Research*, vol. 30, no. 3, pp. 411–433, 2004.

[21] L. N. Smith, "No more pesky learning rate guessing games," *CoRR, abs/1506.01186*, vol. 5, 2015.

[22] L. N. Smith, "A disciplined approach to neural network hyper-parameters: Part 1–learning rate, batch size, momentum, and weight decay," *arXiv preprint arXiv:1803.09820*, 2018.

[23] L. N. Smith, N. Topin, "Super-convergence: Very fast training of neural networks using large learning rates," in *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications*, vol. 11006, 2019, p. 1100612, International Society for Optics and Photonics.

[24] S. Ioffe, C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[25] V. Nair, G. E. Hinton, "Rectified Linear Units Improve Restricted Boltzmann Machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807–814.

[26] K. He, X. Zhang, S. Ren, J. Sun, "Delving Deep Into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.

[27] T. Karras, T. Aila, S. Laine, J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," *arXiv preprint arXiv:1710.10196*, 2017.

[28] P. Micikevicius, S. Narang, J. Alben, G. Diamos, E. Elsen, D. Garcia, B. Ginsburg, M. Houston, O. Kuchaiev, G. Venkatesh, *et al.*, "Mixed precision training," *arXiv preprint arXiv:1710.03740*, 2017.

[29] D. P. Kingma, J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[30] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921–2929.

Satheshkumar Kaliyugarasan

Satheshkumar Kaliyugarasan is a doctoral researcher at the Mohn Medical Imaging and Visualization Center focusing on machine learning in radiological imaging. In 2019 he completed his MSc degree in soft-ware engineering at the Western Norway University of Applied Sciences, Norway.

Arvid Lundervold

Arvid Lundervold is a professor of medical information technology at the University of Bergen and head of the Neuroinformatics and Image Analysis Laboratory in the Neural Networks Research Group, and co-leader of the Computational Medical Imaging and Machine Learning Group at the MMIV center. His research interests are image processing and pattern recogni tion, functional imaging, image registration, quantification and visualization, and mathematical modeling. Lundervold received an MD from the University of Oslo and a PhD in physiology from the University of Bergen.

Alexander S. Lundervold

A.S. Lundervold has a PhD in mathematics from the University of Bergen, Norway. He's currently working as an associate professor at the Western Norway University of Applied Sciences, and as a senior data scientist at the Dept. of radiology, Haukeland University Hospital, Norway. He leads the Computational Medical Imaging and Machine Learning Group at MMIV, together with A.L. His expertise lies in medical data analysis, with a particular focus on medical image processing and applications of machine learning to medicine.