Tech Report ISL-1-2020

Lung Nodule Sizes are Encoded when Scaling CT Image for CNN's

Dmitry Cherezov[1,*], Rahul Paul[1], Nikolai Fetisov[1], Robert J. Gillies[2],
Matthew B. Schabath[3], Dmitry B. Goldgof[1], Lawrence O. Hall[1].

[1] Department of Computer Sciences and Engineering, University of South Florida. Tampa, Florida, USA
[2] Department of Cancer Physiology, H. Lee Moffitt Cancer Center and Research Institute, Tampa, Florida, USA
[3] Department of Cancer Epidemiology, H. Lee Moffitt Cancer Center and Research Institute. Tampa, Florida, USA

Abstract

Non-invasive diagnosis of lung cancer in early stages is one task where Radiomics helps. Clinical practice shows that the size of a nodule has high predictive power for malignancy. In the literature, Convolutional Neural Networks (CNN) have become widely used in medical image analysis. In this work, we study the ability of a CNN to capture nodule size in CT images after images are resized for CNN input. For our experiments, we used the National Lung Screening Trial dataset. Nodules were labeled into two categories (small/large) based on the original size of a nodule. After all extracted patches were re-sampled into 100-by-100-pixel images, a CNN was able to successfully classify test nodules into small and large size groups with high accuracy. In order to show the generality of our discovery, we repeated size classification experiments using Common Objects in Context (COCO) dataset. From the dataset, we selected three categories of images: bears, cats, and dogs. For all three categories a 5x2-fold Cross-Validation was performed to put them into small and large classes. The average Area Under Receiver Operating Curve (AUROC) is 0.954, 0.952 and 0.979 for the bear, cat and dog categories, respectively. So, camera image

[*] Corresponding author
Email address: cherezov@mail.usf.edu (Dmitry Cherezov)

rescaling also enables a CNN to discover the size of an object. The source code for experiments with the COCO dataset is publicly available in Github[†].

Keywords: Convolutional Neural Network, Explanation, Lung Cancer, Computed Tomography, Camera Images.

Introduction

In Radiomics convolutional neural networks (CNNs) are being used to try to answer all kinds of medical questions. Diagnosing (1-4), treatment response (5-7), patient survival time prediction (8-10), etc. A CNN can use 2D or 3D data.

Due to the nature of a CNN, automated extraction of features and a determination of their association with labels; it happens that CNNs may learn unexpected image properties. For example, Zech et al. (11) presented a CNN model for pneumonia detection in chest X-Ray images. Authors showed that the resulting model was able to recognize hospitals and departments, as well as the imaging device because patients with different risk scores of pneumonia were scanned using different imaging protocols. In addition, sicker patients ended up in particular locations. As a result, hospital, department and scanner information is predictive by itself and was learned.

In our previous work (12), we presented a CNN model that was trained to predict whether a benign lung nodule will become a malignant tumor in two years using low dose CT images. As one of the preprocessing steps, we used a warping technique to resize images to the CNN's input resolution. The warping method extracts a patch with a minimum bounding box, which is enough to include the region of interest (ROI). After extraction, the patch is resampled to the size required for the CNN input. The alternative for warping is cropping. Cropping extracts an ROI patch with

[†] https://github.com/VisionAI-USF/COCO_Size_Decoding/

size equal to the CNN input image, thus resampling is not used. Figure 1 shows a visual representation of the warping and cropping methods.

The warping method scales the X, and Y axes of an image using S_x, S_y coefficients respectively. These scaling coefficients depend on the size of an ROI. We hypothesize that a CNN may learn texture specific modifications associated with resampling and as a result learn the size of an ROI, i.e. CNN learns an object's (nodule's) size when the warping method is used. In lung cancer diagnosis, nodule size represented by the ROI is a highly predictive feature, thus a CNN may learn one of the most predictive diagnostic features.

To check our hypothesis that nodule size was implicitly learned by our model, from (12), we designed a series of experiments. Nodules from the National Lung Screening Trial (NLST) were divided into two groups (small/large) using different labeling methods. For experiments with the NLST dataset we used a CNN architecture from our previous work (12) which focused on Lung cancer prediction in the future. We trained a model from scratch as well as tuning pre-trained models.

Moreover, we tested whether this phenomenon is more than a unique effect that occurred in the NLST dataset i.e. if a CNN can decode size information from non-medical images. For that, we used the Common Objects in Context (COCO) dataset (13, 14). Out of 80 object categories we selected three: bears, cars, and dogs. The COCO dataset provides RGB images and segmentations of objects where the size of the objects varies. For the selected categories we repeated the size classification experiment using 5 times 2-fold cross-validation. The COCO dataset is publicly available. The pre-processing, training and testing source code is publicly available in Github (15).

Material and Methods

National Lung Screening Dataset

The NLST was a randomized trial of 53,439 patients that compared Low Dose CT (LDCT) vs. standard chest x-rays. After an initial screening (T0), follow-up screenings (T1 and/or T2) were conducted in intervals of approximately one-year. If at T1 a patient was diagnosed with cancer, he/she started treatment and did not have a T2 follow-up screening. According to the screening protocol, a screen was considered positive if a non-calcified nodule (NCN) had its longest diameter (LD) larger than 4 mm. For positive screenings, radiologists provided a clinical description such as location, margin, etc.

We extracted two cohorts from NLST (16) Cancer patients in the training cohort (Cohort1) had a positive (non-cancer) screening at time 0 and were diagnosed with cancer on the first follow-up (N = 104). Cancer patients in the test cohort (Cohort2) had a positive nodule, so non-cancer screening, at time 0 and time 1. They were diagnosed with cancer at time 2. For each cancer patient, two non-cancer subjects were selected by demographic criteria: the same age, sex, and other available criteria. Finally, we excluded cases with technical problems or other challenges that prevented the analysis of nodules. When removing a cancer patient from the dataset the corresponding non-cancer patients remained. A detailed description of the dataset can be found in Cherezov et al (17).

In this work we are not focused on Lung Cancer diagnosis, thus we relabeled patients. Labels in this study represent the size of a nodule: small or large. Different categorization methods can be used for relabeling. To analyze model performance and stability we used five methods for categorization.

Longest diameters for a nodule of 6, 8 and 10 millimeters were used as a threshold for splits. They were chosen because they are considered representative milestones in the evolution of a nodule according to Lung-RADS (18).

As we can see from Figure 1, scaling parameters, S_x and S_y, for patch length and height are independent. The smaller the length/height the larger the corresponding scaling factor and influence on texture. Thus, for each patch, we selected the smallest of the two values, length or height. For labeling, as a threshold, we used a median of the smallest values in the training cohort.

Finally, as a threshold value, we used the median value of a nodule ROI area in pixels.

The numbers of patients within each class for all labeling approaches are shown in Table 1. Cohort1 T0 was used as a training dataset. Cohort2 T0, T1, and T2 were used as an unseen test cohort.

COCO Dataset

The Common Objects in Context dataset (13, 14) consists of 330,000 large scale images among which more than 200,000 images are labeled. Overall there are 1.5 million segmented objects of 80 categories.

In our work, we used images provided by the COCO team as the training and the validation sets in 2014 and 2017 years. Images were combined into a single set. 5 times 2-fold Cross-Validation technique was performed on the combined dataset. The preprocessing, training and testing source code is publicly available in Github (15).

We selected three categories: bears, cats, and dogs. For the selected categories 2730, 9940, and 11452 object's patches were extracted, respectively. For patch extraction, we used bounding boxes provided by the COCO dataset. The largest bounding box within each category was computed. For all three categories, the maximum bounding box was 640x640 pixels. As a part of warping method, all the patches were resampled into 640 by 640 images and used as input to a CNN for training and testing.

In the COCO dataset we used only one labeling method. We computed the median area of extracted patches before resampling and used the resulting value for thresholding i.e. if a patch area is smaller than the median area of a category then the resampled image is considered small, otherwise, it is considered as a large image. Labeling was performed individually for each category before cross-validation.

Previous Results on NLST dataset

In NLST, for our experiments, we chose a CNN architecture and pre-trained model presented by Paul et al. (12) because the authors used the same dataset for training the model and showed up-to-date performance. The original model was trained to predict if a benign nodule will evolve into a malignant tumor in two years. Following our hypothesis, this trained model could (and did) learn nodule sizes from texture as well as malignancy characteristics. We studied this question in experiments described below.

The CNN model was a cascade network. There are two branches ("left"/"right"). The "left" branch consists of a max-pooling layer before merging. The "right" branch consists of two convolution layers where each of them were followed by a max-pooling layer. After the second max-pooling layers the "right" and the "left" branches are merged. After merging there are convolution and a max-pooling layers. Their result represented as a vector (flattened) and used as an input to a single fully connected layer, which is considered as an output layer in the architecture. The CNN model showed 76% accuracy on the NLST dataset. Detailed information about the architecture and performance of the model can be found in the original paper.

In comparison, Hawkins et al. (19) used 219 Radiomics features (size, location intensity and texture features) extracted from each patient in NLST cohorts, explained above, to build a conventional radiomics model (Naive Bayesian, Random Forests, SVM classifiers) to predict if a benign nodule will evolve into a malignant tumor in two years. As a baseline result, Hawkins used the accuracy of the ROI volume feature only. The accuracy of the volume feature was 71.6%. A complete list of experiments and detailed information about results can be found in the original paper.

Experiments

Design of experiments on the NLST dataset was focused on three questions: 1) Is a CNN model capable of learning an original nodule's size after image resampling? 2) Is a CNN model capable of using encoded size information in its decision-making process? 3) Does the model from our previous work implicitly use encoded size information?

In order to check the generality of a CNN implicitly learning an object's size, we designed a size detection experiments on RGB camera dataset.

Experiment design for the NLST dataset

First (Experiment 1), we trained a CNN model from scratch using Paul's architecture (12). All weights are randomly initialized and the model was trained on Cohort1 to classify nodules with respect to one of the size labeling methods described above. The goal of this experiment was to determine how much information about the size of a nodule is encoded into the texture by resampling can be extracted by a CNN.

Second, (Experiment 2), we tuned the CNN model created as a result in Experiment 1, originally trained to classify nodule size. The model was tuned (100 epochs with 0.0001 learning rate, 0.1 dropout) to predict if a benign nodule evolves into a malignant nodule in two years. Learning rates for all convolution layers were set to zero, fixing the features extracted from the image, and the last fully connected layer was randomly reinitialized. The goal of this experiment was to determine if encoded by scaling and decoded by CNN, size information can be used in a decision-making process for Lung Cancer diagnosis.

Third (Experiment 3), we tuned Paul's pre-trained CNN model designed to predict if a benign nodule will evolve into a malignant tumor in two years. The model was tuned (100 epochs with 0.0001 learning rate, 0.1 dropout) to predict nodule size. A detailed description of the model can be found in our previous work (12). Learning rates for all convolution layers, which would have extracted features from the images, were set to zero and the last layer, fully connected, was randomly reinitialized. The goal of this experiment is to determine how much information about nodule size was used by Paul's CNN (12).

In the experiments 1 and 2 Cohort1 T0 was used for a training and Cohort2 T0, T1, T2 were used for testing. For comparability with our previous results in the experiment 3 we used Cohort1 T0 for training and Cohort2 T0 for testing.

Experiment design for the COCO dataset

We performed 5 times 2-fold Cross-Validation technique for the COCO dataset. At each iteration, a training fold was used to develop a CNN model capable of classifying an extracted patch into one of two categories (small/large). The CNN architecture is shown in Figure 2. Learning rate = 0.0001. Decay = 0.001. Epochs = 100. We used early stopping techniques with patience = 10. As a Validation set, we used 20% of the training fold. The CNN was trained from scratch for each training fold. For repeatability we used predefined individual seeds for each dataset split into folds and training/validation sets.

Results

For the NLST dataset, we checked whether Paul's CNN architecture from our previous work was capable of decoding size information when trained from scratch. The pre-trained model can be tuned for size group classification. Finally, a model trained for size classification can be tuned for tumor malignancy classification.

For the COCO dataset, we checked if a CNN is capable to classify common camera images into size groups.

NLST results

Results of Experiment 1 (Table 2) show that a CNN model can distinguish the difference between small and large nodules with high accuracy. Labeling using 6, 8 and 10 mm of a nodule's longest diameter as a threshold showed smaller accuracy values compared to other labeling methods. Potentially this is caused by the fact that the longest diameter length does not take into account lengths of nodule projections onto axes, which, as we discussed above and showed in Figure 1, define S_x, S_y scaling factors and as a result, encode size into image texture.

Hawkins et al. (19) used the accuracy of an ROI volume feature in a baseline performance model for the prediction that a benign nodule evolves into a malignant tumor in two years. In that experiment, accuracy was 71.6%. Paul et al. (12) using the same dataset, but a CNN for a nodule classification improved the accuracy to 76%. These values can be considered as lower and upper bound values for Experiment 2. In the experiment we tuned a CNN model, trained to classify the size of an ROI, to classify if a benign nodule will evolve into a malignant tumor in two years. Following our assumption that if a CNN learns to extract the size of ROI then the CNN's accuracy should not be significantly smaller than the baseline result provided by Hawkins, though performance using 2D versus 3D features may vary. Paul's CNN model was trained from scratch to predict the malignancy of a nodule. Thus, results of a tuned model in Experiment 2 would not be expected to be higher because most probably Paul's CNN model learned to extract additional texture features associated with cancer compared to a model trained to extract size information.

Results from Experiment 2 (Table 3) show that a CNN trained to classify nodule size can be used for diagnosis. Nevertheless, due to the fact that accuracy values in the experiment are consistently smaller than the accuracy of the CNN trained for diagnosis, we can surmise that the model from our previous work (12) learns additional image characteristics.

Results from Experiment 3 (Table 4) show that the CNN model trained for nodule malignancy prediction (12) can be used for nodule size detection and as a result, we assume that nodule size is a feature of the image that the model learned.

COCO Results

The result for 5 times 2-fold Cross-Validation on the COCO dataset is shown in Table 5. As we can see for all the selected categories accuracy and AUC metrics show "high" performance. Performance in the "Dog" category is higher in comparison to performance in the other two categories. We assume that this is related to the number of images among different categories. There are 11452, 9940, and 2730 images for "Dog", "Cat", and "Bear" categories respectively.

Discussion

In this paper, we examined the hypothesis that an object's size is encoded into the image texture by resampling during the pre-processing step and decoded by a CNN. We used two datasets. For the National Lung Screening Trial dataset, we trained a model from scratch as well as tuned pre-trained models from our previous work. In the Common Objects in Context dataset we performed 5 times 2-fold Cross-Validation where CNN models were trained from scratch. The results of the experiments support our hypothesis on both datasets. Thus, image warping (resampling) implicitly encodes an object's size information into texture.

Radiomics, as a cross-disciplinary field, uses clinical data, imaging data, and machine learning tools. It was considered that when CNN models are used it will be hard to include clinical features into a model. Nevertheless, we showed that at least in our previous models the CNN learned to decode a nodule's size and used it in its decision-making process. As a result, this raises a question. Is it possible to encode some other clinical features into medical images such that a CNN model could use it and which will benefit the performance of the model? As we can see, there are some examples when it occurs. A model recognized hospitals, departments, and scanners from chest X-Ray images because this information was related to pneumonia risk score (11). In our work, the CNN model was able to learn tumor size because the size is an important feature in Lung Cancer diagnosis and malignancy prediction. In these examples, clinical information was encoded accidentally and researchers did not choose what information to encode. Thus, the question is if it is possible to control that process?

We shared the code which we used for experiments in the COCO dataset. The code is capable to repeat the provided experiments with categories that were used in this work as well as capable to perform the same experiments based on the remaining categories. In addition, the code provides tools for different types of filtering (15).

Conflict of interest

The authors have declared no competing interests.

Acknowledgements

This research was partially supported by the National Institutes of Health under grants 4KB17, (U01 CA143062), and (U24 CA180927). The authors thank the National Cancer Institute for access to NCI's data collected by the National Lung Screening Trial. The authors thank the National Science Foundation for equipment used in the experiments (1513126). The statements contained herein are solely those of the authors and do not represent or imply concurrence or endorsement by NCI.

References

1. Cao Z, Duan L, Yang G, Yue T, Chen Q. An experimental study on breast lesion detection and classification from ultrasound images using deep learning architectures. BMC medical imaging. 2019;19(1):51.

2. De Man R, Gang GJ, Li X, Wang G. Comparison of deep learning and human observer performance for detection and characterization of simulated lesions. J Med Imaging (Bellingham). 2019;6(2):025503.

3. Deepak S, Ameer P. Brain tumor classification using deep CNN features via transfer learning. Computers in biology and medicine. 2019;111:103345.

4. Kiryu S, Yasaka K, Akai H, Nakata Y, Sugomori Y, Hara S, et al. Deep learning to differentiate parkinsonian disorders separately using single midsagittal MR imaging: a proof of concept study. European radiology. 2019;29(12):6891-9.

5. Ha R, Chang P, Karcich J, Mutasa S, Van Sant EP, Connolly E, et al. Predicting post neoadjuvant axillary response using a novel convolutional neural network algorithm. Annals of surgical oncology. 2018;25(10):3037-43.

6. Wu E, Hadjiiski LM, Samala RK, Chan H-P, Cha KH, Richter C, et al. Deep Learning Approach for Assessment of Bladder Cancer Treatment Response. Tomography. 2019;5(1):201.

7. Xu Y, Hosny A, Zeleznik R, Parmar C, Coroller T, Franco I, et al. Deep learning predicts lung cancer treatment response from serial medical imaging. Clinical Cancer Research. 2019;25(11):3266-75.

8. Yang C-K, Yeh JC-Y, Yu W-H, Chien L-I, Lin K-H, Huang W-S, et al. Deep Convolutional Neural Network-Based Positron Emission Tomography Analysis Predicts Esophageal Cancer Outcome. Journal of clinical medicine. 2019;8(6):844.

9. Balkenhol MC, Bult P, Tellez D, Vreuls W, Clahsen PC, Ciompi F, et al. Deep learning and manual assessment show that the absolute mitotic count does not contain prognostic information in triple negative breast cancer. Cellular Oncology. 2019:1-15.

10. Ibragimov B, Toesca D, Yuan Y, Koong A, Daniel C, Xing L. Neural networks for deep radiotherapy dose analysis and prediction of liver SBRT outcomes. IEEE journal of biomedical and health informatics. 2019.

11. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. PLoS medicine. 2018;15(11):e1002683.

12. Paul R, Hawkins S, Schabath MB, Gillies RJ, Hall LO, Goldgof DB. Predicting malignant nodules by fusing deep features with classical radiomics features. Journal of Medical Imaging. 2018;5(1):011021.

13. Common Objects in Context, Image Dataset. Available from: http://cocodataset.org.

14. Lin T-Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, et al., editors. Microsoft coco: Common objects in context. European conference on computer vision; 2014: Springer.

15. Source code for size classication experiments using Common Objects in Context Image Dataset. Available from: https://github.com/VisionAI-USF/COCO_Size_Decoding.

16. Schabath MB, Massion PP, Thompson ZJ, Eschrich SA, Balagurunathan Y, Goldof D, et al. Differences in patient outcomes of prevalence, interval, and screen-detected lung cancers in the CT arm of the national lung screening trial. PloS one. 2016;11(8):e0159880.

17. Cherezov D, Goldgof D, Hall L, Gillies R, Schabath M, Müller H, et al. Revealing tumor habitats from texture heterogeneity analysis for classification of lung cancer malignancy and aggressiveness. Scientific reports. 2019;9(1):4500.

18. Radiology. ACo. Lung-RADS version 1.1 assessment categories.09-July-2019. Available from: https://www.acr.org/-/media/ACR/Files/RADS/Lung-RADS/LungRADSAssessmentCategoriesv1-1.pdf.

19. Hawkins S, Wang H, Liu Y, Garcia A, Stringfield O, Krewer H, et al. Predicting malignant nodules from screening CT scans. Journal of Thoracic Oncology. 2016;11(12):2120-8.

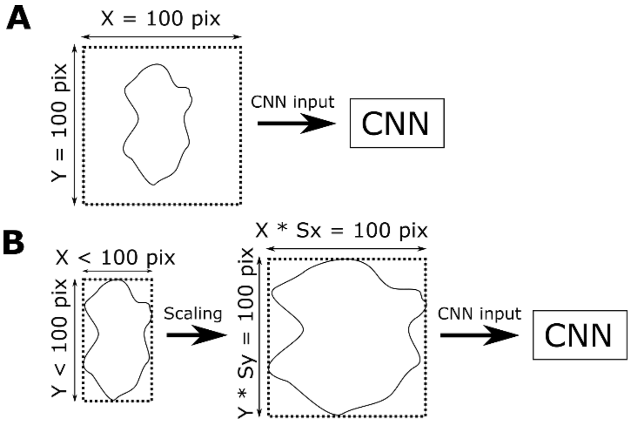

Figure 1. Cropping (A) and Warping (B) patch extraction methods. The solid line represents the region of interest border. The dashed line represents an extracted patch border. This assumes that CNN input is a 100x100-pixel image. X and Y represent the corresponding patch's width and height, respectively.

Threshold	Cohort1 T0		Cohort2 T0		Cohort2 T1		Cohort2 T2	
	Small	Large	Small	Large	Small	Large	Small	Large
LD 6 mm	57	204	44	193	39	171	44	166
LD 8 mm	129	132	126	111	106	104	89	121
LD 10 mm	183	65	172	65	140	70	126	84
Median of min axe	122	139	89	148	128	82	124	86

Median nodule area	128	133	99	138	123	87	117	93
Total	261		237		210		210	

Table 1. Number of patients in groups after labeling nodules by size. The number of patients in Cohort2 at T0 and T1/T2 vary because some patients were excluded due to low image quality or patient removal for the trial.

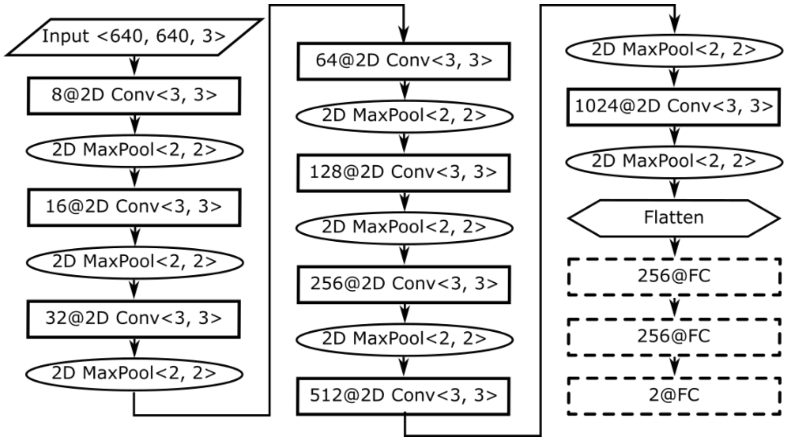

Figure 2. The CNN architecture used for size classification in the COCO dataset. There are eight convolution layers with 3x3 kernels. Each convolution layer is followed by a max-pooling layer with a 2x2 window and stride equal to 2. For all but the last layers the ReLU activation function was used. The softmax activation function was used for the last fully connected (FC) layer. Dropout for all FC layers was set to 0.75.

Threshold	Cohort2 T0 (%)	Cohort2 T1 (%)	Cohort2 T2 (%)
LD 6 mm	95 (0.97)	79.52 (0.85)	81.4 (0.85)
LD 8 mm	89 (0.947)	79 (0.839)	76 (0.82)
LD 10 mm	94.5 (0.9784)	87 (0.867)	84 (0.877)
Median of min size	99.2 (0.9998)	92.38 (0.94)	94.28 (0.95)

Median nodule size	94.93 (0.9894)	97.14 (0.9978)	95.7 (0.9974)

Table 2. Accuracy and AUC (in brackets) of a CNN trained from scratch for classification a nodule original size group (Experiment 1).

Threshold	LD 6 mm	LD 8 mm	LD 10 mm	Median of min size	Median nodule size
Accuracy (%)	72.15 (0.76)	74.26 (0.788)	75.1 (0.8182)	74.26 (0.786)	74.26 (0.794)

Table 3. Accuracy and AUC (in brackets) of a CNN trained for nodule original size classification after tuning for cancer classification (Experiment 2). Accuracy of a CNN trained from scratch to classify cancer is 76%. Accuracy of cancer classification using a tumor volume only is 71.6%.

Threshold	Cohort2 T0 (%)	Cohort2 T1 (%)	Cohort2 T2 (%)
LD 6 mm	93.67 (0.969)	79.52 (0.82)	81.4 (0.858)
LD 8 mm	90.3 (0.923)	81 (0.8438)	80.5 (0.828)
LD 10 mm	93.67 (0.9763)	87.14 (0.9235)	84.76 (0.907)
Median of min size	100 (1)	92.4 (0.937)	94.3 (0.962)
Median nodule size	97.89 (0.989)	98.57 (0.989)	98.09 (0.99)

Table 4. Accuracy and AUC (in brackets) of a CNN trained for cancer classification after tuning to classify a nodules original size group (Experiment 3).

Run	Fold	Bear	Cat	Dog
1	A	89.8 (0.942)	88.5 (0.929)	93.4 (0.983)
	B	89.9 (0.964)	88.2 (0.968)	93.9 (0.974)
2	A	85.2 (0.946)	88.6 (0.966)	93 (0.98)
	B	88.3 (0.965)	89.8 (0.956)	94.1 (0.98)
3	A	88.5 (0.964)	88.6 (0.951)	86.3 (0.971)

	B	90.7 (0.969)	86.6 (0.951)	91.6 (0.98)
4	A	88.5 (0.97)	87.2 (0.954)	92.4 (0.986)
	B	90.5 (0.944)	89.2 (0.959)	93.2 (0.981)
5	A	89.8 (0.95)	88.8 (0.953)	93.1 (0.982)
	B	88.8 (0.933)	88.8 (0.94)	93.5 (0.982)

Table 5. Accuracy and AUC (in brackets) results for 5 time 2-fold Cross-Validation in the COCO dataset.