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], and Lawrence O. Hall[1]

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

Corresponding Author:
Dmitry Cherezov, PhD
University of South Florida, 2215 N Spring Glade Cir,
Tampa, Florida, USA 33613;
E-mail: cherezov@mail.usf.edu

Key Words: Convolutional neural network, explanation, lung cancer, computed tomography, camera images

Abbreviations: convolutional neural network (CNN), computed tomography (CT), region of interest (ROI), Common Objects in Context (COCO)

ABSTRACT

Noninvasive 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 (CNNs) have become widely used in medical image analysis. We study the ability of a CNN to capture nodule size in computed tomography images after images are resized for CNN input. For our experiments, we used the National Lung Screening Trial data set. Nodules were labeled into 2 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. To show the generality of our discovery, we repeated size classification experiments using Common Objects in Context (COCO) data set. From the data set, we selected 3 categories of images, namely, bears, cats, and dogs. For all 3 categories a 5- × 2-fold cross-validation was performed to put them into small and large classes. The average area under receiver operating curve is 0.954, 0.952, and 0.979 for the bear, cat, and dog categories, respectively. Thus, camera image rescaling also enables a CNN to discover the size of an object. The source code for experiments with the COCO data set is publicly available in Github (https://github.com/VisionAI-USF/COCO_Size_Decoding/).

INTRODUCTION

In radiomics studies, convolutional neural networks (CNNs) are applied to address different medical questions including diagnosing (1–4), treatment response (5–7), and patient survival time prediction (8–10). An unexpected consequence has been observed when CNNs are used in image analyses, in 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 and showed that the resulting model could identify hospitals, departments, and 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. Therefore, hospital, department and scanner information are predictive by themselves and was learned by the CNN.

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 2 years using low-dose computed tomography (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). For a given an ROI, a bounding box was defined as a rectangle whose width and height were equal to the width and the height of ROI. The rectangle was located on an image such that it enclosed the ROI. Voxels/pixels within the rectangle were extracted as a patch. 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 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 and 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 therefore learn the size of an ROI, that is, when the warping method is used, CNN learns an object's (nodule's) size. 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 test our hypothesis that nodule size was implicitly learned by our model (12), we designed a series of experiments. Lung

ISSN 2379-1381 https://doi.org/10.18383/j.tom.2019.00024

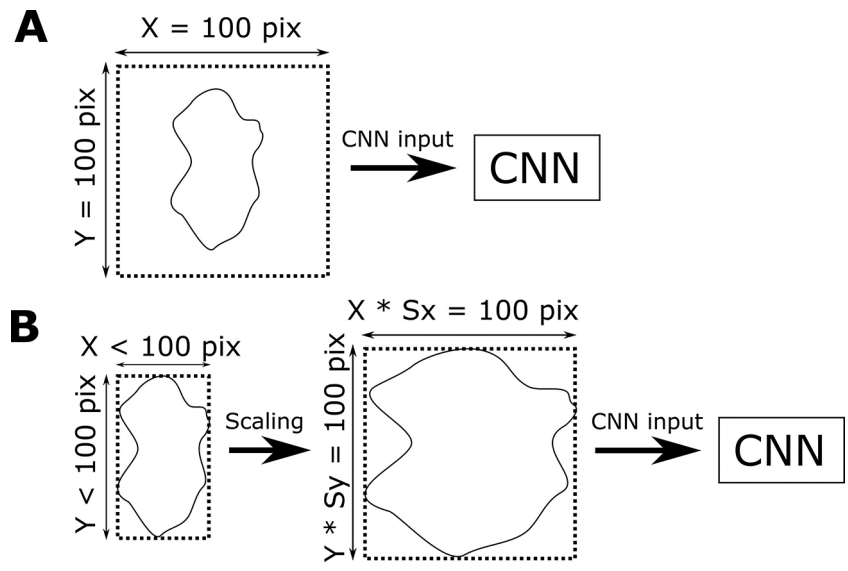

Figure 1. Cropping (A) and warping (B) patch extraction methods. The solid line represents the region of interest border. The dashed line represents an extracted patch border. This assumes that convoluted neural network (CNN) input is a 100-× 100-pixel image. X and Y represent the corresponding patch's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

nodules from low-dose CTs belonging to the National Lung Screening Trial (NLST) were divided into 2 groups, namely, small and large, using different labeling methods. For experiments with the NLST data set, 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 and tuned pretrained models. Moreover, we tested whether this phenomenon is more than a unique effect that occurred in the NLST data set (ie, if a CNN can decode size information from nonmedical images). For that, we used the Common Objects in Context (COCO) data set (13, 14). We selected 3 out of 80 object categories, namely, bears, cars, and dogs. The COCO data set 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- × 2-fold cross-validation. The COCO data set is publicly available.

As such, the goal of this work is to demonstrate that upsampling encodes nodule size information in lung CT images in which size has implications for nodule classification. Camera images were used to show that this is not a fluke phenomenon. The preprocessing, training, and testing source code is publicly available in Github (15).

MATERIALS AND METHODS

National Lung Screening Data Set

The NLST is a randomized trial of 53,439 patients that compared low-dose CT with standard chest radiography. After the baseline screening (T0), follow-up screenings (T1 and/or T2) were conducted at intervals of ∼1 year. If a screening participant was diagnosed with cancer at T0 or at T1, they did not have subsequent screening at T1 or T2, respectively. According to the NLST protocol, a screen was considered positive if a noncalcified nodule had its longest diameter >4 mm. For positive screenings, radiologists provided clinical description such as location and margins.

Based on prior work, we identified 2 cohorts from NLST (16). Patients with lung cancer in the training cohort (cohort1) had a positive screening result (noncancer) at T0 and had a positive screening result at T1 that was diagnosed as lung cancer (N = 104). Patients with lung cancer in the test cohort (cohort2) had a positive nodule result (noncancer) at T0 and T1 and had a positive screening result at T2 that was diagnosed as lung cancer. For each cancer patient, 2 positive screen noncancer subjects were selected and matched by age, sex, and smoking history. Participants were excluded if technical problems with the images or other challenges that prevented the analysis of nodules. When removing a cancer patient from the data set, the corresponding noncancer patients remained. A detailed description of the data set can be found in the study by Cherezov et al. (17).

Radiomic Features

In this work we have 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 5 methods for categorization. Longest diameters for a nodule of 6, 8, and 10 mm 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).

We used a single-click semiautomatic intensity–based segmentation algorithm with a subsequent segmentation quality check by a radiologist. The longest diameter of a nodule was computed according to the Response Evaluation Criteria in Solid Tumors (RECIST) protocol (19) using the Definiens software (20). First, the largest segmentation area slice is selected. In the resulting slice, all possible lines are plotted such that each line starts and ends in voxels that are considered as boundary voxels (a voxel for which at least one the neighboring voxels is considered as outside of the ROI). Among the plotted lines the line that has the largest length is selected and considered as the nodule longest diameter.

As shown in Figure 1, scaling parameters, S_x and S_y, for patch length and height, respectively, are independent. The

Table 1. Number of Patients in Groups after Labeling Nodules by Size

Threshold	Cohort1 T0		Cohort2 T0		Cohort2 T1		Cohort2 T2	
	Small	Large	Small	Large	Small	Large	Small	Large
Longest diameter 6 mm	57	204	44	193	39	171	44	166
Longest diameter 8 mm	129	132	126	111	106	104	89	121
Longest diameter 10 mm	183	65	172	65	140	70	126	84
Median of min size	122	139	89	148	128	82	124	86
Median nodule area	128	133	99	138	123	87	117	93
Total	261		237		210		210	

The number of patients in cohort 2 at T0 and T1/T2 vary because some patients were excluded due to low image quality or patient removal for the trial.

smaller the length/height the larger the corresponding scaling factor and influence on texture. Thus, for each patch, we selected the smallest of the 2 values, namely, 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 data set. Cohort 2 T0, T1, and T2 were used as an unseen test cohort.

COCO Data Set

The COCO data set (13, 14) consists of 330,000 large-scale images, among which >200,000 images are labeled. Overall there are 1.5 million segmented objects of 80 categories. In the COCO data set object segmentations were provided by the data set developers. These segmentations were used for patch extraction without any modifications.

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

For the selected categories of bears, cats, and dogs, 2730, 9940, and 11 452 object's patches were extracted, respectively. For patch extraction, we used bounding boxes provided by the COCO data set. The largest bounding box within each category was computed. For all 3 categories, the maximum bounding box was 640 × 640 pixels. As a part of warping method, all the patches were resampled into 640 × 640 images and used as input to a CNN for training and testing.

In the COCO data set we used only 1 labeling method. We computed the median area of extracted patches before resampling and used the resulting value for thresholding, that is, 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 Data Set

In NLST, for our experiments, we chose a CNN architecture and pretrained model presented by Paul et al. (12) because the authors used the same data set 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 2 years. Following our hypothesis, this trained model could (and did) learn nodule sizes from texture and malignancy characteristics. We studied this question in experiments described in the following sections.

The CNN model was a cascade network. There are 2 branches ("left"/"right"). The "left" branch consists of a max-pooling layer before merging. The "right" branch consists of 2 convolution layers in which each branch was 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 is represented as a vector (flattened) and is used as 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 data set. Detailed information about the architecture and performance of the model can be found in the original paper.

In comparison, Hawkins et al. (21) used 219 radiomics features (size, location intensity, and texture features) extracted from each patient in NLST cohorts to build a conventional radiomics model (naive Bayesian, Random Forests, SVM classifiers) to predict if an indeterminate nodule will evolve into a malignant tumor in 2 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

The design of experiments using the NLST data set was focused on the following 3 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?

To check the generality of a CNN implicitly learning an object's size, we designed a size detection experiment on a color (RGB) camera data set.

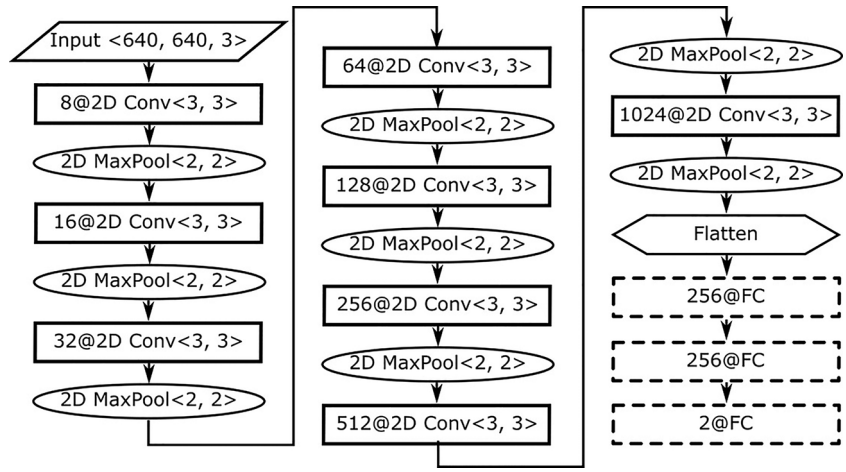

Figure 2. The CNN architecture used for size classification in the Common Objects in Context (COCO) data set. There are 8 convolution layers with 3 × 3 kernels. Each convolution layer is followed by a max-pooling layer with a 2 × 2 window and stride equal to 2. For all but the last layers, the rectified linear unit (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.

Experiment Design for the NLST Data Set

Table 1 shows the number of patients within each class after relabeling them into size categories. First (experiment 1) we trained a CNN model from scratch using Paul's architecture (12). All weights were randomly initialized and the model was trained on cohort 1 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 that 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 2 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 whether when 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 pretrained CNN model designed to predict if a benign nodule will evolve into a malignant tumor in 2 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 experiments 1 and 2, cohort 1 T0 was used for a training and cohort 2 T0, T1, T2 were used for testing. For comparability with our previous results in experiment 3, we used cohort 1 T0 for training and cohort 2 T0 for testing.

Experiment Design for the COCO Data Set

We performed 5- × 2-fold cross-validation technique for the COCO data set. At each iteration, a training fold was used to develop a CNN model capable of classifying an extracted patch into 1 of 2 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 data set split into folds and training/validation sets.

RESULTS

For the NLST data set, we assessed whether the CNN architecture from Paul et al. from our previous work (12) was capable of decoding size information when trained from scratch. The pretrained model can be tuned for size group classification. Finally,

Table 2. Accuracy and AUC (in Brackets) of a CNN Trained from Scratch for Classification a Nodule Original Size Group (Experiment 1)

Threshold	Cohort2 T0 (%)	Cohort2 T1 (%)	Cohort2 T2 (%)
Longest diameter 6 mm	95 (0.97)	79.52 (0.85)	81.4 (0.85)
Longest diameter 8 mm	89 (0.947)	79 (0.839)	76 (0.82)
Longest diameter 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 3. Accuracy and AUC (in Brackets) of a CNN Trained for Nodule Original Size Classification after Tuning for Cancer Classification (Experiment 2)

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)

Accuracy of a CNN trained from scratch to classify cancer is 76%. Accuracy of cancer classification using a tumor volume only is 71.6%.

a model trained for size classification can be tuned for tumor malignancy classification.

For the COCO data set, we checked if a CNN is capable of classifying 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 with 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 in sections above and showed in Figure 1, define S_x and S_y scaling factors and, as a result, encode size into image texture.

Hawkins et al. (21) 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 2 years. In that experiment, the accuracy was 71.6%. Paul et al. (12) using the same data set, 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 2 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, although 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 likely Paul's CNN model learned to extract additional texture features associated with cancer compared with 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, owing 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- × 2-fold cross-validation on the COCO data set is shown in Table 5. As we can see for all the selected categories accuracy and area under the curve metrics show "high" performance. Performance in the "dog" category is higher than that in the other 2 categories. We assume that this is related to the number of images among different categories. There are 11 452, 9940, and 2730 images for "dog," "cat," and "bear" categories, respectively.

DISCUSSION

In this paper, we used 2 data sets to test the hypothesis that the size of an object (ie, pulmonary nodule) is encoded into the image texture by resampling during the preprocessing step and decoded by a CNN. Using images from NLST data set, we trained a model from scratch and also tuned pretrained models from our previous work (12). Using images from the COCO data set, we performed 5- × 2-fold cross-validation in which all CNN models were trained from scratch. The results of the experiments support our hypothesis on both data sets. Thus, image warping (resampling) implicitly encodes an object's size information into texture.

It is unknown if a CNN model that was trained and tested on the NLST data set considered heterogeneity of a nodule.

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)

Threshold	Cohort2 T0 (%)	Cohort2 T1 (%)	Cohort2 T2 (%)
Longest diameter 6 mm	93.67 (0.969)	79.52 (0.82)	81.4 (0.858)
Longest diameter 8 mm	90.3 (0.923)	81 (0.8438)	80.5 (0.828)
Longest diameter 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 5. Accuracy and AUC (in Brackets) Results for 5- × 2-Fold Cross-Validation in the COCO Data Set

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)

Generally, smaller lesions are more homogeneous and become increasingly heterogeneous as the size (and volume) increases. At the same time, the COCO data set consists of objects that can be barely evaluated from a homogeneity/heterogeneity point of view. Nevertheless, the CNN efficiently differentiated "small" and "large" objects. Overall, it is possible that heterogeneity is a characteristic of a nodule that was used by a CNN for decision-making.

In the case where we used the median area of a nodule as a threshold for splitting the NLST data set into "small" and "large" nodules, the resulting classes are well balanced (Table 1). Both size categories do have benign and malignant nodules, and at the same time, results for size classification remain high. Thus, we may conclude that heterogeneity of a nodule cannot be the only texture characteristic that the CNN potentially used for decision-making.

As we can see from Tables 2 and 4, if a nodule's longest diameter feature was used as a criterion for the NLST split into size categories, then performance for size classification decreases in cohort 2 T1 and T2. Nevertheless, for the other labeling methods, classification accuracy and AUC remain high. Thus, we can conclude that classification performance depends on labeling methods that were applied.

We performed an additional experiment. Instead of classification of size groups, we tested using regression to directly predict size. Because the regression task is more complicated and requires more data in comparison to the classification task, we used the COCO data set for the experiment. The experiment is similar to the one we described in experiment design for the COCO data set section, but instead of size category, labels represent the size of extracted patches. The CNN model shown in Figure 2 was adapted to the regression task (1 output neuron, mean square error loss function, linear activation function for the last layer); 5- × 2-fold cross-validation was performed. As for performance metrics, the Pearson correlation coefficient between predicted size and the actual size was used. For the bear, cat, and dog categories, we got mean Pearson correlation coefficient values 0.861, 0.867, and 0.9 respectively. The goal of this paper is to show that upsampling encodes nodule size information in lung CT in which size has implications for nodule classification. Thus, we consider the results of the regression task, and questions such as "Why does upsampling encode size?" "How accurately can size be determined?" as material for future work. Hence, we do not include details of the experiment in this paper. Nevertheless, the provided Github source code is able to perform regression tasks.

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 data set. The code is capable of repeating the provided experiments with categories that were used in this work as well as of performing the same experiments based on the remaining categories. In addition, the code provides tools for different types of filtering (15).

ACKNOWLEDGMENTS

This research was partially supported by the National Institutes of Health under grants 4KB17 (U01 CA143062 and U24 CA180927). We thank the National Cancer Institute for access to NCI's data collected by the National Lung Screening Trial, and 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.

Conflict of Interest: None reported.

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 Med Imaging. 2019;19:51.
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;6:025503.
3. Deepak S, Ameer P. Brain tumor classification using deep CNN features via transfer learning. Comput Biol Med. 2019;111:103345.
4. Kiryu S, Yasaka K, Akai H, Nakata Y, Sugomori Y, Hara S, Seo M, Abe O, Ohtomo K. Deep learning to differentiate parkinsonian disorders separately using single midsagittal MR imaging: a proof of concept study. Eur Radiol. 2019;29:6891–6899.
5. Ha R, Chang P, Karcich J, Mutasa S, Van Sant EP, Connolly E, Chin C, Taback B, Liu MZ, Jambawalikar S. Predicting post neoadjuvant axillary response using a novel convolutional neural network algorithm. Ann Surg Oncol. 2018;25:3037–3043.
6. Wu E, Hadjiiski LM, Samala RK, Chan H-P, Cha KH, Richter C, Cohan RH, Caoili EM, Paramagul C, Alva A, Weizer AZ. Deep learning approach for assessment of bladder cancer treatment response. Tomography. 2019;5:201.
7. Xu Y, Hosny A, Zeleznik R, Parmar C, Coroller T, Franco I, Mak RH, Aerts HJ. Deep learning predicts lung cancer treatment response from serial medical imaging. Clin Cancer Res. 2019;25:3266–3275.
8. Yang CK, Yeh JC, Yu WH, Chien LI, Lin KH, Huang WS, Hsu PK. Deep convolutional neural network-based positron emission tomography analysis predicts esophageal cancer outcome. J Clin Med. 2019;8:844.
9. Balkenhol MCA, Bult P, Tellez D, Vreuls W, Clahsen PC, Ciompi F, van der Laak JA. Deep learning and manual assessment show that the absolute mitotic count does not contain prognostic information in triple negative breast cancer. Cell Oncol. 2019;42:1–15.
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 J Biomed Health Inform. 2019;23
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 Med. 2018;15: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. J Med Imag. 2018;5:011021.
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, editors. Microsoft COCO: common objects in context. European conference on computer vision; 2014: Springer.
15. Source code for size classification 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, Aberle DR, Gillies RJ. 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:e0159880.
17. Cherezov D, Goldgof D, Hall L, Gillies R, Schabath M, Müller H, Depeursinge A. Revealing tumor habitats from texture heterogeneity analysis for classification of lung cancer malignancy and aggressiveness. Sci Rep. 2019;9:4500.
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. Therasse P, Arbuck SG, Eisenhauer EA, Wanders J, Kaplan RS, Rubinstein L, Verweij J, Van Glabbeke M, van Oosterom AT, Christian MC, Gwyther SG. New guidelines to evaluate the response to treatment in solid tumors. J Natl Cancer Inst. 2000;92:205–216.
20. Definiens A. Definiens Developer 7 – Reference Book. Definiens AG, München. 2007:21–24.
21. Hawkins S, Wang H, Liu Y, Garcia A, Stringfield O, Krewer H, Li Q, Cherezov D, Gatenby RA, Balagurunathan Y, Goldgof D, Schabath MB, Hall L, Gillies RJ. Predicting malignant nodules from screening CT scans. J Thorac Oncol. 2016;11:2120–2128.