Unboxing AI - Radiological insights into a deep neural network for Lung Nodule Characterization

Vasantha Kumar Venugopal[1] Kiran Vaidhya [2] Abhijith Chundur [2] Vidur Mahajan [1] Murali Murugavel [1] Suthirth Vaidhya [2] Digvijay Mahra[2] Akshay Rangasai [2] and Harsh Mahajan M.D.[1]

[1]Centre for Advanced Research in Imaging, Neurosciences & Genomics (CARING), Mahajan Imaging, New Delhi, INDIA

[2]Predible Health, Bangalore, INDIA

ABSTRACT

Rationale and Objectives:

To explain predictions of a deep residual convolutional network for characterization of lung nodule by analyzing heat maps

Materials and Methods

A 20-layer deep residual CNN was trained on 1245 Chest CTs from NLST trial to predict the malignancy risk of a nodule. We used occlusion to systematically block regions of a nodule and map drops in malignancy risk score to generate clinical attribution heatmaps on 103 nodules from LIDC-IDRI dataset, which were analyzed by a thoracic radiologist. The features were described as heat inside nodule (IH)-bright areas inside nodule, peripheral heat (PH)-continuous/interrupted bright areas along nodule contours, heat in adjacent plane (AH)-brightness in scan planes juxtaposed with the nodule, satellite heat (SH)- a smaller bright spot in proximity to nodule in the same scan plane, heat map larger than nodule (LH)-bright areas corresponding to the shape of the nodule seen outside the nodule margins and heat in calcification (CH)

Results

These six features were assigned binary values. This feature vector was fed into a standard J48 decision tree with 10-fold cross-validation, which gave an 85 % weighted classification accuracy

with a 77.8 %TP rate, 8% FP rate for benign cases and 91.8% TP and 22.2 %FP rates for malignant cases. IH was more frequently observed in nodules classified as malignant whereas PH, AH,
and SH were more commonly seen in nodules classified as benign.

Conclusion

We discuss the potential ability of a radiologist to visually parse the deep learning algorithm generated 'heat map' to identify features aiding classification.

Introduction

The recent surge in applications built on artificial intelligence (AI) has resulted in the need for greater training, understanding and studying the ethical implications of AI in the field of medical imaging (1,2). The excellent image classification capabilities of deep learning (DL) are often approached with caution due to perceived lack of explainability of their functioning (3). This perceived opacity in the functioning of DL algorithms has led to such networks being referenced as a 'black box'. Democratization in the understanding of the network architecture with the help of activation maps, has led to greater emphasis on explainable AI (exAI or xAI). This development has found favor as a method which can lead skeptical clinicians and radiologists towards informed adoption of the DL techniques in their practice (4). Such works can help the transition from in-silico testing of AI solutions to prospective clinical evaluation. In this paper, we attempt to explain predictions of a deep residual convolutional network developed for characterization of lung nodules by analyzing the class attribution maps - more generally known "heat maps".

Material and Methods

All required Ethics approval was obtained from the institutional ethics committee.

The class attribution map analysis featured a 20-layer deep residual convolutional neural network trained to identify and characterize the pulmonary nodules. This network was trained on 1245 Chest CT scans from the NLST (5) trial to predict the malignancy risk of a given CT.

The training consisted of two stages where the network was trained to detect pulmonary nodules in the first stage and pool the nodules through a leaky noisy-OR gate in the second stage to predict overall malignancy risk for the CT scan.

Training:

A deep learning system based on convolutional neural networks was trained to predict the malignancy status from CT scans of the chest. The deep learning system comprises of a nodule detector and a malignancy estimator. The nodule detection system was trained and validated to pick up pulmonary nodules >= 3mm on 554 CT scans from NLST and 888 CT scans from LIDC-IDRI (8,9). The malignancy estimator was trained on the 1245 CT scans and validated on 350 CT scans from NLST.

CADe (Nodule Detector):

The CADe system is an ensemble of 3 single-shot 3D Feature Pyramid Networks (FPN) (8,9) which is trained to detect lung nodules from the CT scans. The 3D FPN is built with a U-Net encoder-decoder architecture, composed of 3D convolutions, to maximize the effective receptive field and fuse multi-scale information. Multi-scale information is essential for differentiating pulmonary nodules from vasculature present in the organs. The network takes in a 3D patch of size 128x128x128 as input and gives out 32x32x32x3x5 as output, with 3 anchor boxes of varying size limits for each network in the ensemble. During inference, the ensemble is rolled over the CT scan with 128x128x128 overlapping patches and the predicted bounding boxes are fused with non-maximum suppression to provide the final candidates for lung nodules.

The CADe system is trained on 711 CTs from LIDC-IDRI and 554 CTs from NLST with Adam optimizer and a learning rate of 0.0001, a weight decay of 0.0005 and a dropout of 0.5. Validation was done on 177 CTs from LIDC-IDRI. The data was augmented with nodules of different sizes to ensure training was not biased towards detecting small nodules. The network architecture is depicted in Figure 1.

CADx (Malignancy Estimator):

The CADx system is a leaky Noisy-OR gate (10) based on deep convolutional neural networks. The noisy-OR model operates on 96x96x96 patches from each detected nodule, fuses the information from each nodule and gives the probability of the patient being affected by lung cancer. During training, top 5 nodule candidates, based on their nodule probabilities, are taken from the CADe system and fed to the noisy-OR model. A leakage probability is assigned to the CADx model during training to account for missed primary nodules/masses by the CADe system. The CADx model shares the same backbone as the CADe model with the convolutional layers sharing their weights to avoid over-fitting.

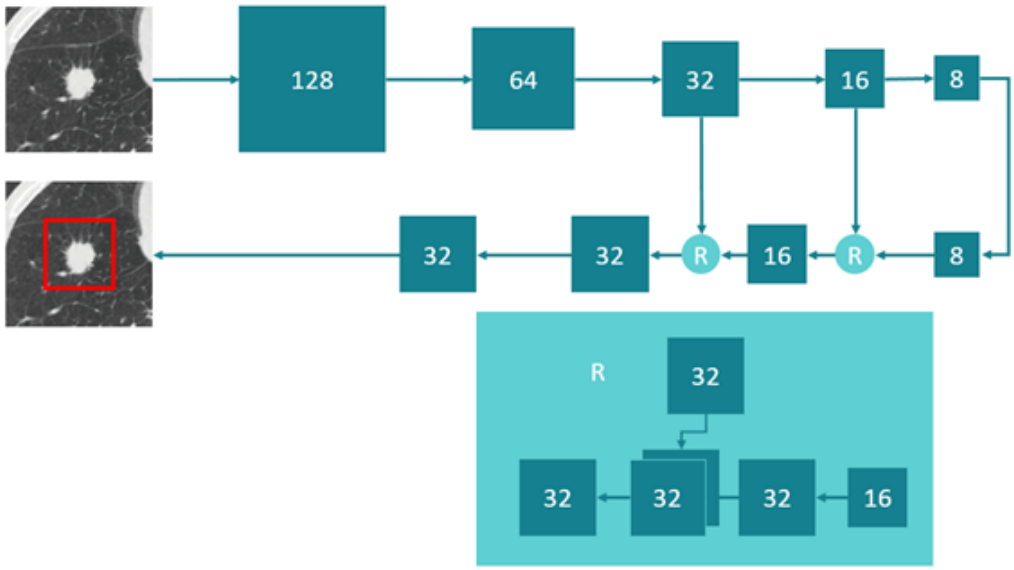

Figure 1. The deep learning network architecture

The CADx system was trained on 1245 CT scans and validated on 350 CT scans from NLST. During inference, all the detected nodules are considered to compute the overall malignancy risk at a scan-level.

Occlusion

The CADx network was taken to understand the decision-making process of predicting malignancy for a given pulmonary nodule. The analysis of the functioning was initiated by a technique called “occlusion” (10), wherein, systematic blocking of various regions of the nodule were done and drops in the malignancy risk score by the CADx network were recorded. The changes in malignancy scores are mapped back to form a clinical attribution heat map for the said nodule. A patch size of 4x4x4 was taken to block the image of the nodule with a stride of 4 steps per block. For an input patch of 96x96x96, an output heatmap of size 24x24x24 was obtained. The 24x24x24 patch was then zoomed back to 96x96x96 with trilinear interpolation to map it back to the original nodule patch.

Saliency Maps

The saliency maps are generated by systematically blinding the classifier to specific regions of the nodule. This can be visualized as breaking down the nodule into multiple 3D tiles, and individually single tiles are removed and replaced during prediction through the neural network. The difference in classification outputs of the neural network is monitored and used to generate a final heatmap. The process is based on the intuition that the probability of malignancy should drop when a relevant region on the nodule is blinded to the neural network during prediction. This technique is depicted in figure 2.

Such heat maps were generated for randomly selected 103 unseen nodules from LIDC-IDRI dataset. It comprised of 49 benign and 54 malignant nodules. The network correctly predicted malignancy in 38 nodules and absence of malignancy in 37 nodules. There were 12 benign nodules incorrectly classified as malignant and 16 malignant nodules incorrectly classified as benign. The sensitivity and specificity values were 70.3 % and 75 % respectively. The distribution of these nodules is summarized in **table 1**.

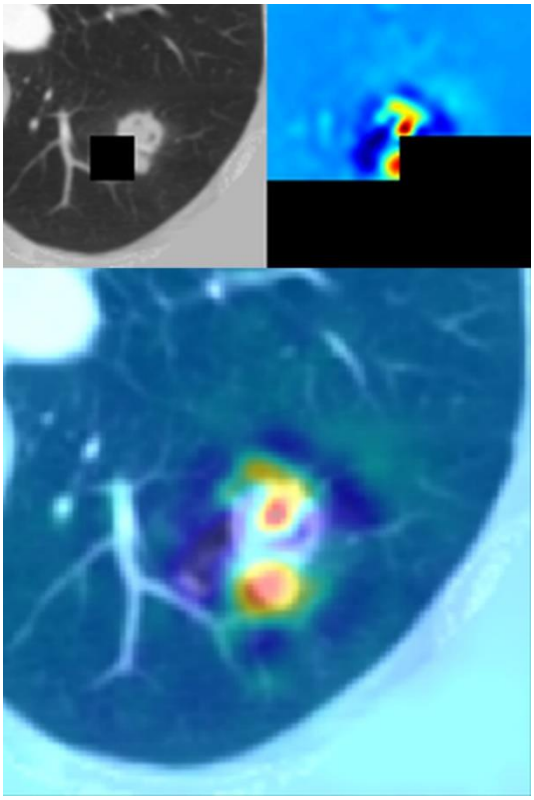

Figure 2. The occlusion technique used for creating the class activation maps

	Positive (Biopsy)	Negative (Biopsy)	Total Biopsy proven results
Malignant (AI)	38	16	54
Benign (AI)	12	37	49
Total network labels	50	53	103

Table 1. Confusion matrix comparing the network labels with the ground truth.

These heat maps were analyzed by a radiologist with more than 8 years' experience in chest imaging. The radiologist analyzed the pattern and discernible features in the heat maps generated for network-benign and network-malignant nodules separately. The common features observed in the heat maps for both the set of nodules on axial planes were described using the following terminology: 'heat inside the nodule', 'peripheral heat rim', 'heat in adjacent scan plane', 'satellite heat', 'heat map larger than nodule' and 'heat in calcification'.

1. **Heat inside the nodule:**

Any activation or bright areas observed inside the nodule contours is considered as positive for heat inside the nodule activation. It included both homogenous and heterogenous activations (Fig 3a)

2. **Peripheral heat rim**

This is defined as peripheral activation or bright area along the boundaries of the nodule. It includes both continuous as well as interrupted areas of activation along the margins (Fig 3b).

3. **Heat in calcification**

This feature is defined as presence of activation within the calcified regions of the nodule. When confounded by presence of other defined features like 'heat inside the nodule' or 'peripheral heat rim', this feature is distinguished by positive activation more prominent than the other forms of activation (Fig 3c).

4. **Satellite heat**

This includes areas of activation, near but distinct from the nodule. It is also typically smaller than the size of the nodule and is seen in the same scan plane of the parent nodule (Fig 3d).

5. **Heat map larger than the nodule**

This is described as areas of activation larger than the size of the nodule extending outside its margins (Fig 3e).

6. **Heat in adjacent scan planes**

This is described as areas of activation seen in immediately adjoining scan planes both above and below the plane in which the nodule is seen (Fig 3f).

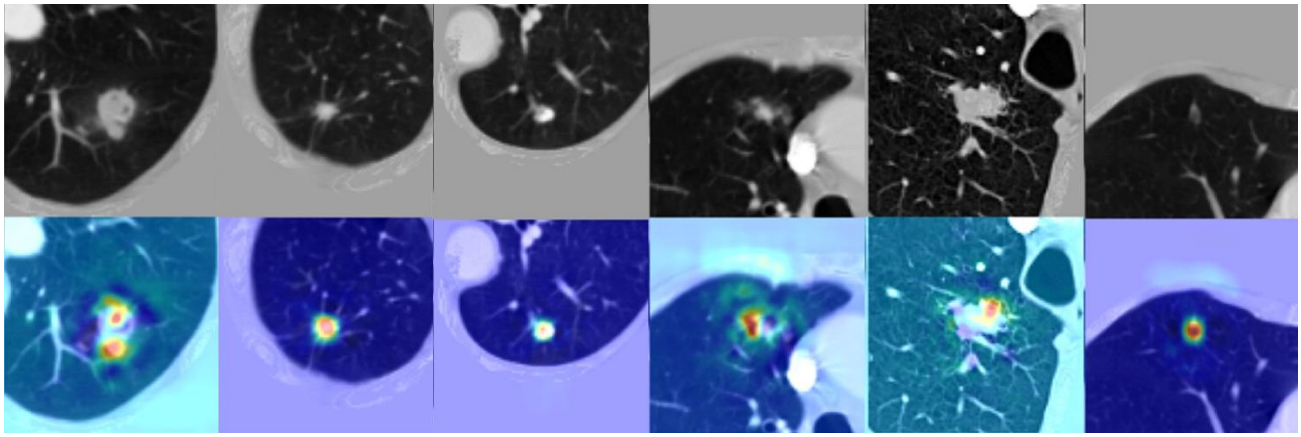

Figure 3. Heat Map features overlaid on the CT scan. 3a. Heat inside nodule, 3b. peripheral rim heat, 3c. Heat inside calcification, 3d. Satellite heat, 3e. Heat map larger than nodule, 3f. Parenchymal heat in adjacent scan plane

The radiologist analyzed all the class activation maps generated for all these 103 nodules.

Some of these heat map features are depicted without the CT overlay in figure 4.

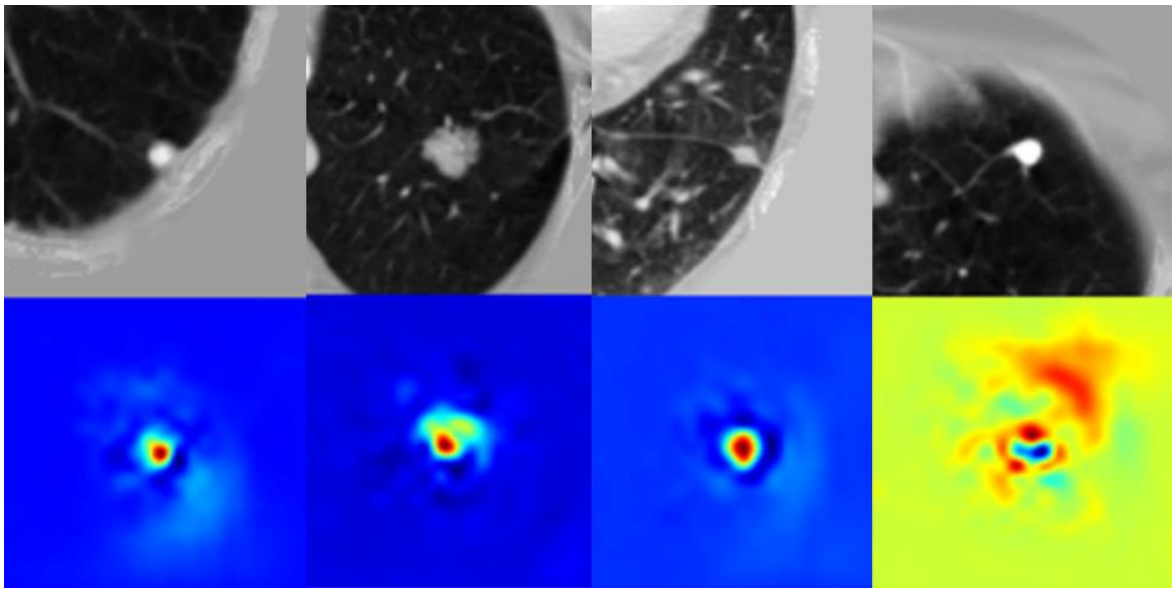

Figure 4a-b. Heat maps represented separately without overlay on corresponding CT images. Figure 4a. Peripheral rim in a subpleural calcified benign nodule with activation seen in an area larger than the nodule. Figure 4b. Heat map matching the size of a malignant nodule. No activation in the adjacent areas. Figure 4c. Peripheral rim and heat inside the nodule seen in a benign nodule. Figure 4d. Heat map larger than the nodule in a calcified benign nodule.

The presence or absence of the six described features in the heat maps generated for each of these nodules was binarized - 1 for the presence of the feature and 0 for absence of it. This feature vector was then fed into a standard J48 decision tree with 10-fold cross-validation.

Results

The presence or absence of each of these features were analyzed in the clinical attribution maps generated for these nodules. The distribution of these features and the correlation with the network labels and the ground truth is summarized in Table 2.

1. Heat inside the nodule:

This was seen in 67 of the 103 nodules. 19 of these nodules were classified as benign by the network with two of them being false negative predictions on comparison with ground truth. 48 of these nodules were classified as malignant with 11 false positive predictions.

Features	True Benign	Benign False positive	True malignant	Malignant false negative	Total
Heat inside the nodule	17 (25%)	11(16%)	37(55%)	2 (3%)	67
Peripheral heat rim	29 (38%)	6 (8%)	25 (33%)	16 (21%)	76
Heat in calcification	17 (74%)	5(21%)	1 (4 %)	0 (0%)	23
Satellite heat	12 (57%)	1 (5%)	4 (19%)	4 (19%)	21
Heat map larger than nodule	25 (33%)	12 (16%)	23 (30%)	16 (21%)	76
Lung heat in adjacent planes	18 (51%)	1 (3%)	5 (14%)	11 (31%)	35

Table 2. Summary of distribution of the radiologist described features and their correlation with the network labels and the ground truth

2. Peripheral heat rim

Peripheral rim of activation is seen in 86 nodules. 45 of these nodules were classified as benign by the network whereas 31 were classified as malignant. Among these there were 16 false negative and 6 false positive predictions. This peripheral rim of activation appeared

interrupted in 23 of the true benign predictions and 8 of the false negative predictions. In all the remaining cases, the peripheral rim of activation appeared continuous.

3. Heat in calcification

51 of the 103 nodules revealed some degree of calcification. 34 of these nodules were classified as benign by the network with false negatives predicted in four cases. 17 of those nodules with calcification were classified as malignant with a false positive in 6 nodules.

Heat in calcification was noted in 23 of these 51 nodules. 17 of them were classified as benign with all predictions corresponding to the ground truth. None of the four calcified nodules which were falsely classified as benign by the network showed any activation within the calcification whereas five out of the six nodules which were incorrectly classified as benign showed activation with their calcification. Also, out of the 11 true positive calcified malignant nodules, only one showed activation within the calcification.

4. Satellite heat

Satellite nodule like activation was seen in 21 nodules, 16 of them were classified as benign and 5 of them classified as malignant by the network. There were four false negative predictions in the 16 network-benign nodules and one false positive prediction among the five network-malignant nodules.

5. Heat map larger than the nodule

This was noted in 76 nodules. 41 of these nodules were classified as benign by the network with 16 false negative predictions. 35 nodules with heat map larger than their size were classified as malignant with 12 false positive predictions.

6. Heat in adjacent scan planes

Activation in adjacent scan planes were observed in 35 nodules out of 103. Among these there were 29 nodules that were classified as benign with 11 of them being false negative

classifications. Six nodules were classified as malignant by this network with one false positive prediction.

The ability of using these described features to predict the network classification was then assessed by feeding this feature vector into a standard J48 decision tree with 10-fold cross validation. The J48 decision tree gave 85 % weighted classification accuracy for predicting the output based on these activation maps. The use of this decision tree independently yielded a true positive prediction rate of 77.8 %, False positive prediction rate of 8% for the network-benign cases. It also resulted in a true positive prediction rate of 91.8% and False Positive prediction rate of 22.2 % for network-malignant cases. It is important to note that the decision tree was used to predict the network classification, but not the true malignancy probability of the nodules. Hence the ground truth for the comparisons are the network predicted classes and not the biopsy results. These results are summarized in Table 3.

	Malignant network label	**Benign network label**
TP rate	92%	78%
FP rate	22%	8%

Table 3. Results of the decision tree prediction of the network labels analyzing the heat maps

The decision tree highlighting the potential ability of a radiologist to visually parse the DL algorithm generated heat map to identify features aiding diagnosis is shown in Figure 5.

Among the individual features, the interior heat had the highest correlation with malignancy prediction (correlation coefficient = 0.19, p = 0.06) whereas peripheral heat rim correlates with benignity (correlation coefficient = 0.34, p = 0.0007).

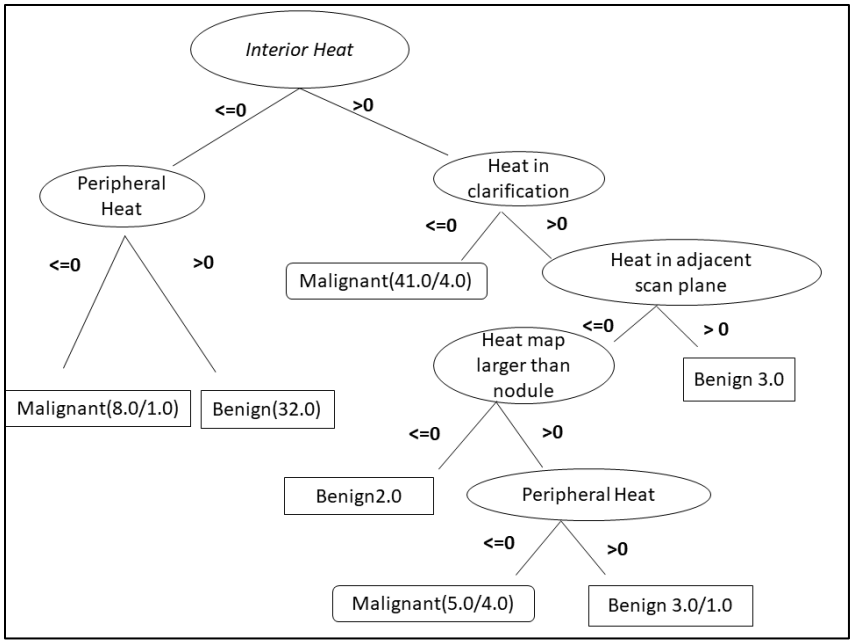

Figure 5. The decision tree highlighting the potential ability of a radiologist to visually parse the Deep Learning algorithm generated heat map to identify features aiding diagnosis

The morphological features of these nodules recorded by three radiologists provided with the LIDC-IDRI dataset (7) were compared with these heat maps to evaluate any correlation and Pearson correlation co-efficient was calculated (table 4). The morphological features compared included diameter of the nodule, lobulation, margins, sphericity, spiculation, texture (solid, part-solid or ground glass) and perceived malignancy risk (subjective assessment of the likelihood of malignancy, assuming the scan originated from a 60-year-old male smoker).

None of the heat map features had strong positive correlation with the morphological features defined by the radiologists. There was a moderate negative correlation between the perceived malignancy on morphology and the peripheral interrupted heat as well as heat in calcification (r = 0.4).

Discussion

There is a certain degree of overlap in usage of the terms like explainable AI, interpretable machine learning, intelligible DL. At a higher level, these terms imply the need for assisting

physicians and other downstream DL users make appropriate decisions by providing a greater insight into the functioning of the DL. In its true sense, explainable AI is not limited to mere transparency of a model, but the need for every constituent element of the system to be understandable (11,12). The granularity of the explanations should be at the level of human cognition with the assumption that the end user has basic domain knowledge.

In this study, we attempt to explain the interpretations of a specific radiology focused DL network by analyzing the weighted outputs generated by systematic occlusion. The outputs are categorized in to features relevant to the classification problem as well as the domain knowledge and understanding of the end-user.

Interior heat or heat inside the nodule was found to be statistically significant classifier for malignancy. This is consistent with the current radiology practice of examining the presence or absence of heterogeneity in a nodule while characterizing it on CT scan (13,14). Peripheral heat rim was the statistically significant classifier for benignity of the nodule. This again is explainable by a radiologist, that smooth contours without spiculated margins is a feature characteristic of benign nodules.

Heat inside a calcification was found to be a very specific indicator of benignity. There was calcification in 36 benign and 15 malignant nodules in the test set. The network showed positive activation in 21 of the benign nodules, classifying 17 of them as benign, all of them corresponding to the biopsy ground truth. The remaining five were incorrectly classified as malignant. There was activation in five of the malignant nodules with one correct prediction and five false negative classifications. It can be proposed from these results that the presence of calcification and probably patterns of calcification were weighted by the network in classifying a nodule is benign. Further analysis by classifying the patterns of calcification may provide further insights into the network.

Satellite heat was found to increase the likelihood of classification of a nodule as benign. Even though this observation is concordant with the clinical observations that the presence of a satellite nodule is a feature seen in benign granulomatous nodules (15), many of the cases with activations did not show a visible satellite nodule. No simple hypothesis can be derived for this observation.

Heat in adjacent scan planes was again found to favor a benign classification with nearly 83% of the nodules with such activation being classified as benign. The nearest domain correlate for this observation is the knowledge that presence of ground glass densities surrounding the nodule is known to be associated with malignancy in most cases (16,17).

Heat map larger than the nodule was the most frequent observation without significant classifying power. This observation is an in-plane correlate of the activation in the adjacent scan planes.

The absence of significant correlation between the morphological features and the features defined on the heat maps as shown in Table 4 can be attributed to new non-classical associations learned by the deep learning algorithm. Further studies are needed to evaluate this hypothesis.

	Periphery continuous rim	Periphery Heat Interrupted	satellite heat	Heat in adjacent scan plane	heat in calcification	heat map larger than nodule
DIAMETER	0.24112932	-0.16534159	-0.166364188	-0.284376328	-0.298716381	0.013059719
LOBULATION	0.030248106	-0.21193904	-0.048335144	-0.346816471	-0.279608714	0.032982279
MALIGNANCY	0.3318012	-0.411120725	-0.229251406	-0.336958386	-0.418976478	-0.084846449
MARGIN	0.018399385	0.107230571	0.023395494	0.210285525	0.322800471	0.061185738
SPHERICITY	0.084191996	0.01062422	0.021640558	-0.029768654	-0.070725147	-0.128055609
SPICULATION	0.051792968	-0.171801518	-0.12164727	-0.280054649	-0.357049494	-0.023466616
TEXTURE	0.163299316	-0.280252341	-0.163441332	-0.095849599	0.126491106	-0.046915743

Table 4. Pearson correlation index for heatmap features vs morphological features of the nodule

This algorithm makes predictions based on features learned by correlation with the ground truth established by biopsy. The focus of our research work is not on the accuracy of these

predictions but rather on the ability of a human reader to predict the predictions of the algorithm from the saliency maps.

Strengths and Limitations

The strength of the study is that it is attempting to offer clinical insights into the functioning of a DL network by analyzing the clinical attribution maps. We also compare the correlates between the imaging features of characterizing a pulmonary nodule to the features on the activation maps.

The major limitation of the study is that the features were described and observed by a single reader. The downside of using multiple blinded readers would have been too many non-overlapping features to evaluate and build hypotheses towards explainability. However future approaches can include multi-reader consensus-based methods.

The other limitation of the study can be traced to the argument that truly explainable system should integrate reasoning that explains the decision-making process of the model via user understandable features of the input data. One school of thought **(12)** argues that leaving explanation generation to human analysts can be dangerous since, depending on their background knowledge about the data and its domain, different explanations may be deduced.

The handcrafting of the classification patterns of the activation maps is another major limitation in this study. Until newer methods that might enable explicit automated reasoning for DL predictions by extraction of symbolic rules are developed, the human reader identified features might help in comprehending the model properties and functions. Research studies offering supporting evidence for explainability of such algorithms can assist in devising strategies to make AI based applications safe and accountable (18,19).

Conclusions

In this study we describe a method of explain the functioning of a DL network used in the characterization of lung nodule by manual analysis of the class activation maps and

corroborating it with the domain knowledge of radiologists. We derived a decision tree with reasonable accuracy that can predict the model output by analyzing the radiologist described features on the clinical attribution maps.

References

1. Syed, A. B. & Zoga, A. C. Artificial Intelligence in Radiology: Current Technology and Future Directions. *Semin Musculoskelet Radiol* **22**, 540–545 (2018).
2. Tang, A. *et al.* Canadian Association of Radiologists White Paper on Artificial Intelligence in Radiology. *Can Assoc Radiol J* **69**, 120–135 (2018).
3. Ardila, D. *et al.* End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. *Nat. Med.* **25**, 954–961 (2019).
4. O'Sullivan, S. *et al.* Legal, regulatory, and ethical frameworks for development of standards in artificial intelligence (AI) and autonomous robotic surgery. *Int J Med Robot* **15**, e1968 (2019).
5. National Lung Screening Trial Research Team *et al.* Reduced lung-cancer mortality with low-dose computed tomographic screening. *N. Engl. J. Med.* **365**, 395–409 (2011).
6. Armato, S. G. *et al.* The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): A Completed Reference Database of Lung Nodules on CT Scans. *Med Phys* **38**, 915–931 (2011).
7. McNitt-Gray, M. F. *et al.* The Lung Image Database Consortium (LIDC) Data Collection Process for Nodule Detection and Annotation. *Acad Radiol* **14**, 1464–1474 (2007).
8. Lin, T.-Y. *et al.* Feature Pyramid Networks for Object Detection. *arXiv:1612.03144 [cs]* (2016).
9. Liao, F., Liang, M., Li, Z., Hu, X. & Song, S. Evaluate the Malignancy of Pulmonary Nodules Using the 3D Deep Leaky Noisy-or Network. (2017). doi:10.1109/TNNLS.2019.2892409
10. Zeiler, M. D. & Fergus, R. Visualizing and Understanding Convolutional Networks. *arXiv:1311.2901 [cs]* (2013).
11. Lipton, Z. C. The Mythos of Model Interpretability. *arXiv:1606.03490 [cs, stat]* (2016).
12. Doran, D., Schulz, S. & Besold, T. R. What Does Explainable AI Really Mean? A New Conceptualization of Perspectives. *arXiv:1710.00794 [cs]* (2017).
13. Girvin, F. & Ko, J. P. Pulmonary Nodules: Detection, Assessment, and CAD. *American Journal of Roentgenology* **191**, 1057–1069 (2008).

14. Viggiano, R. W., Swensen, S. J. & Rosenow, E. C. Evaluation and management of solitary and multiple pulmonary nodules. *Clin. Chest Med.* **13**, 83–95 (1992).
15. Takashima, S. *et al.* Small Solitary Pulmonary Nodules (≤1 cm) Detected at Population-Based CT Screening for Lung Cancer: Reliable High-Resolution CT Features of Benign Lesions. *American Journal of Roentgenology* **180**, 955–964 (2003).
16. Naidich, D. P. *et al.* Recommendations for the management of subsolid pulmonary nodules detected at CT: a statement from the Fleischner Society. *Radiology* **266**, 304–317 (2013).
17. MacMahon, H. *et al.* Guidelines for Management of Incidental Pulmonary Nodules Detected on CT Images: From the Fleischner Society 2017. *Radiology* **284**, 228–243 (2017).
18. Pesapane, F., Volonté, C., Codari, M. & Sardanelli, F. Artificial intelligence as a medical device in radiology: ethical and regulatory issues in Europe and the United States. *Insights Imaging* **9**, 745–753 (2018).
19. Beig, N. *et al.* Perinodular and Intranodular Radiomic Features on Lung CT Images Distinguish Adenocarcinomas from Granulomas. *Radiology* **290**, 783–792 (2019).