Journal of Medical Imaging

MedicalImaging.SPIEDigitalLibrary.org

Comparison of deep learning and human observer performance for detection and characterization of simulated lesions

Ruben De Man
Grace J. Gang
Xin Li
Ge Wang

Comparison of deep learning and human observer performance for detection and characterization of simulated lesions

Ruben De Man,^{a,*} Grace J. Gang,^b Xin Li,^c and Ge Wang^d

^aStony Brook University, Department of Biochemistry and Cell Biology, Stony Brook, New York, United States

^bJohns Hopkins University, Department of Biomedical Engineering, Baltimore, Maryland, United States

^cGE Global Research, Radiation Imaging Sciences, Niskayuna, New York, United States

^dRensselaer Polytechnic Institute, Department of Biomedical Engineering, Troy, New York, United States

Abstract. Detection and characterization of abnormalities in clinical imaging are of utmost importance for patient diagnosis and treatment. We present a comparison of convolutional neural network (CNN) and human observer performance on a simulated lesion detection and characterization task. We apply both conventional performance metrics, including accuracy and nonconventional metrics such as lift charts to perform qualitative and quantitative comparisons of each type of observer. It is determined that the CNN generally outperforms the human observers, particularly at high noise levels. However, high noise correlation reduces the relative performance of the CNN, and human observer performance is comparable to CNN under these conditions. These findings extend into the field of diagnostic radiology, where the adoption of deep learning is starting to become widespread. Consideration of the applications for which deep learning is most effective is of critical importance to this development. © 2019 Society of Photo-Optical Instrumentation Engineers (SPIE) [DOI: [10.1117/1.JMI.6.2.025503](https://doi.org/10.1117/1.JMI.6.2.025503)]

Keywords: artificial intelligence; image analysis; detection; noise; image quality.

Paper 19007R received Jan. 3, 2019; accepted for publication May 30, 2019; published online Jun. 21, 2019.

1 Introduction

In the field of radiology, accurate detection and characterization of suspicious lesions or tumors are critical for effective and timely treatment of patients.¹ The information gathered from patient images provides oncologists invaluable information needed to successfully diagnose or treat patients with abnormalities. For example, the differing histopathology of benign and malignant tumors has long been one of the most important clinical markers in medicine.² In particular, the differing morphological characteristics of those tumors can be used by histopathologists to determine the type of tumor and its potential development.³ One of the fundamental challenges in extracting this information from patient scans is the presence of noise, the reduction of which has been a key focus in imaging research.^{4,5} With the advent of machine learning and its widespread adaptation in radiology, there have been numerous studies applying this technology to the task of characterizing tumors in noisy images as malignant or benign.^{6–8} In this study, we are interested in comparing the performance of convolutional neural networks (CNNs) with human observers for a lesion detection and characterization task and, in particular, the impact of image noise and image noise structure on their relative performance.

A variety of factors affect lesion detectability and characterization. These include the experience, interest level, and the vision capability of the observer, as well as the properties of the images. Lesion shape, size, contrast, and anatomical clutter have been shown to have a direct effect on lesion detectability.⁹ Noise characteristics such as standard deviation and correlation also have measurable effects on detectability and characterizability,

although the mechanisms by which they do so are still being studied.^{10–12} The performance of model and human observers under varying noise conditions has been compared in the literature.^{13–15} It has also been shown that CNNs can function as anthropomorphic model observers for task-based image quality evaluations.^{16,17} To our knowledge, this is the first systematic comparison between CNN and human observer performance for a lesion detection and classification task as a function of noise level and structure.

On a high level, the purpose of this study is to compare the performance of human observers and deep learning algorithms on basic detection and characterization tasks under the set conditions. It is generally believed that deep learning performance has become superior to human observer performance. In this study, we compare relatively simple CNNs with relatively inexperienced human observers, with the understanding that both CNNs and humans could be more specialized and be further trained in the future. We are interested in understanding their relative performance for basic detection tasks and more complex classification tasks in the presence of confounding structures or noise.

2 Methods

2.1 Image Generation

Grayscale images of size 128 by 128 were generated. The background level was zero. In a subset of the images, a lesion was inserted. Lesions were either circular-, triangular-, or star-shaped. These geometric shapes were chosen for the ease of generating a large set of artificial lesions, the clear morphological distinction between the three lesion types, and the ease of

*Address all correspondence to Ruben De Man, E-mail: rubendeman@gmail.com

training lay human observers. Finally, random noise was added to make the lesions less visible and make the detection and characterization tasks challenging. These resulting images were intended to be generic and not mimic any particular imaging modality. However, we were specifically interested in studying the impact of correlated noise as it occurs in CT images or in any line integral-based imaging modality.

Circle-, triangle-, and star-shaped lesions were generated using basic geometric principles. The nominal dimensions of the shapes were adjusted so that their mean areas were the same, with a nominal radius of 18 pixels for the circular lesions. To make the tasks more realistic, random perturbations of the lesion-to-background contrast as well as in the shape and size of the lesions were performed. For example, the star corners are not uniformly distributed over 360 deg and have varying radii. For all shapes, the uniform deviation of the corner points from nominal was $[-10 \text{ deg}; 10 \text{ deg}]$ for the angles and $[-10\%; 10\%]$ for the radii. In addition, all lesions were blurred using a two-dimensional Gaussian smoothing kernel with a standard deviation of 2 pixels. The lesion intensities were uniformly randomly distributed from 80 to 120 (i.e., with a mean of 100).

Random noise with different correlation levels and standard deviations was added to the images. Uncorrelated noise was

generated by an independent and identically distributed Gaussian process. Correlated noise was derived by spatially convolving the uncorrelated noise with two-dimensional Gaussian blur kernels (with spatial standard deviation 2.5 and 10) and rescaling the result to achieve the desired noise standard deviation. A total of 155 noise realizations were performed for each of the four classes (circle, triangle, star, and no lesion), each of the three noise correlation levels, and each of the 10 noise standard deviations, to create a large image database suitable for deep learning, resulting in a total of 18,600 images (Fig. 1). To account for the change in detectability due to noise correlation, the noise standard deviation was scaled by 50, 25, and 10 to ensure approximately equal detectability (as depicted in Fig. 1). These scale factors were determined empirically so that the lesions in images with the same noise standard deviation levels were equally detectable, regardless of noise correlation level. Example images with star lesions and with various noise levels and noise correlations are shown in Fig. 2.

2.2 Convolutional Neural Network

A CNN was implemented for this study. A total of 18,000 images described above were used to train the CNN. An additional 600 images were generated to test the CNN. All training and testing were completed in MATLAB using the Machine Learning Toolbox. An Nvidia GTX 1050 was used for all graphical computations. The topology of our CNN was validated with other similar experiments found in the literature.^{18,19} It contained four convolutional layers and two fully connected layers and was structured as follows (also illustrated in Fig. 3):

- first convolutional layer with filter size 3×3 and 64 feature maps,
- second convolutional layer with filter size 3×3 and 96 feature maps,
- third convolutional layer with filter size 3×3 and 128 feature maps,
- fourth convolutional layer with filter size 3×3 and 256 feature maps,
- fully connected layer with 128 (low noise) or 2000 (high noise) hidden units,
- dropout layer (high noise),
- fully connected layer with four hidden units, and
- Softmax layer.

An rectified linear unit activation function was applied to all convolutional and fully connected layers. Max pooling layers were applied for each convolutional layer, as in Ref. 6. The low noise levels (1 to 5) and high noise levels (6 to 10) were trained separately to optimize output from the CNN. The architectures used differed only in the number of hidden units and the presence of a dropout layer, as described above. The same two network architectures were used for each of the three noise correlation levels. Hence, six CNNs were independently trained and tested. A total of 3000 images were used to training and 100 images were used for testing each CNN. Unique validation sets of various sizes were used to optimize the network hyperparameters prior to final training and testing.

A learning rate of 0.001 and a mini batch size of 80 were used to optimize the output from the CNN. In addition, L2 regularization was applied to the network weights and biases to

No lesion	Circle	Star	Triangle
Noise correlation: 3 levels [0, 2.5, 10]			
Standard deviation: 10 levels [10 : 100] x [50, 25, 10]			
150 realizations (training) + 5 realizations (testing)			

Fig. 1 Schematic setup for image generation for CNN, resulting in 18,600 images.

Fig. 2 Examples of images with varying noise correlation and standard deviation. Noise level increases from left to right and noise correlation increases from top to bottom.

Fig. 3 Architecture of CNN used in this study.

prevent overfitting of the model. The CNN was trained for 100 epochs at a constant learning rate.

2.3 Human Observer Study

A Java program was developed for human observers to perform the detection and characterization tasks and to compute the accuracy thereof. The Java console randomly displays an image from the library and prompts the user to label it. The user interacts with a panel of options consisting of different shapes or “none,” indicating the absence of a lesion. The user then iterates through a random sample of N images and selects the perceived shape from the panel of options. Feedback is given throughout the process in the form of the real-time detection and characterization accuracies. The Java console is depicted in Fig. 4.

After the user closes the program or selects “quit,” the user input is transmitted to a secure internet server.

Five individuals (including some of the authors) were asked to participate in the human observer study. All human observers were experienced scientific researchers, including some with decades of experience. However, no radiologists were included.

Fig. 4 Java graphical user interface for human observer studies.

Human observers were trained prior to assessing the images by reviewing a brief collection of 20 sample images considered to be representative of the entire library combined with careful instructions. The observers were asked to classify each and were subsequently made aware of the actual classification, thereby allowing them to understand the typical composition of the image library. Observers were also given basic information about the nature of the images, including the types of lesions and appearance of the lesions. For the purposes of this study, completion of this basic training was determined to be sufficient to warrant their participation for this simple phantom lesion detection and characterization task. On completion of the training module, the human observers were asked to evaluate the library. This was done using the Java program described above. For each observer, a unique library consisting of 240 images of varying noise standard deviation and correlation was generated. The composition of these human observer libraries were the same as the one used to train and test the CNN.

2.4 Performance Metrics

All performance metrics were calculated based on the single task consisting of four options: “none,” “circle,” “triangle,” or “star.” For analyzing the accuracy of both the deep learning study and the human observer study, we quantitatively defined detection accuracy (DA) as

$$DA = \frac{tp + tn}{tp + tn + fp + fn} = \frac{tp + tn}{t},$$

Fig. 5 Plot of DA versus noise level for human observers.

where tp , tn , fp , fn , and t refer to the number of true positive, true negative, false positive, false negatives images and total number of images, all defined in terms of “lesion present” or “lesion not present.”

However, to eliminate bias due to the unequal proportion of images with (75%) and without (25%) lesions, we instead used the following weighted detection accuracy (DA_w):

$$DA_{\text{weighted}} = \frac{\frac{tp}{3} + tn}{\frac{tp}{3} + tn + fp + \frac{fn}{3}}.$$

A lift chart²⁰ was produced in lieu of an receiver operating characteristic curve since neither the CNN nor the human studies used a variable threshold based on which a positive or negative was defined. The parametric definition of the lift chart can be given as

$$x = \frac{tp/3 + fp}{tp/3 + tn + fp + fn/3}, \quad y = tp/3,$$

where the tp and fn were again divided by 3 to eliminate bias.

This chart plots each classification task performed by a specific observer. By plotting the threshold of the probabilistic classifier (x) against the number of true positives (y), a convex curve arises. The plot of the most effective observer will have a greater area under the curve, hence providing a meaningful alternative method of evaluation for detection efficacy. We plot an (x , y) pair for each noise level for both human observers and CNN.

For characterization accuracy (CA), we defined the four-class CA as

$$CA = \frac{tc}{t},$$

Fig. 6 Noise level versus detectability and CA plots. Noise correlation increases from top to bottom. DAw plots are on the left, and CA is plotted on the right. Black represents CNN performance, and gray represents human observer performance for detection or characterization.

where t_c refers to the total number of correctly classified images and t refers to the total number of images. The underlying metrics have been widely adopted in similar studies.^{21,22}

To compare the efficacy of the human observer and CNN, DAw and CA were plotted as a function of noise level, and for each of the three correlation levels.

To supplement this and gain a better understanding of classification errors that reduced the CA, confusion matrices were also developed for each classification task.

3 Results

The results of the human observer study are presented in Fig. 5, where the DA is plotted against noise level. Boxplots are used to illustrate the extent of variation between trials performed by different observers at different noise correlation levels.

Human versus CNN comparisons of detection and CA as a function of noise level are shown in Fig. 6.

The lift curve comparing human and CNN performance is plotted in Fig. 7. One should note both the distribution of the markers and the relative distance from the diagonal line, which represent the upper limit on performance.

The confusion matrices corresponding to the zero, low, and high noise correlation CNN output are shown in Fig. 8.

4 Discussion

The results of the human observer study by themselves give some meaningful insight into the effect of noise level and correlation on detectability and characterizability. As expected, the trained CNN performed better than the human observers for both the detection and the characterization tasks at nearly every noise level. Interestingly, the CNN outperformed the human observers decisively for the low noise correlation level, whereas for the high noise correlation level their performance was more similar. The confusion matrix plots give additional information about the CNN classification decisions. As expected, Fig. 7 shows that the CNN is most effective at differentiating between “none” and the other classes. It struggles most with differentiating “star” shapes from “circle” and “triangle” shapes, which can likely be attributed to the fact that the “star” shares important

identifying features with other classes including sharp edges and radial symmetry.

The lift curves allow for comparison between the human observer and the CNN performance. As seen in Fig. 8, the points corresponding to the CNN are situated closer to the upper bound line, indicating greater overall performance. Generally, the points corresponding to lower noise levels are also situated closer to the upper bound line. The CNN points also are more clustered together due to the consistent and robust nature of the CNN decision-making mechanism. Within the human observer data points, there is a notable distribution indicating a stronger dependence on noise level, which may be due to varying skills, capabilities, and evaluative strategies.

Our findings show that the CNNs outperform human observers for the simulated lesion detection and the classification tasks. In particular, it is shown that the CNN has the capacity to perform significantly better than other observers under conditions of low noise correlation, especially at low signal-to-noise ratios. The improved CNN performance at low noise correlation levels might be attributed to the size of the filters used by the network. The small 3×3 size of the filter could contribute to information loss in conditions of high noise correlation, thereby reducing its accuracy. Further work using different filter sizes and network architectures is needed to confirm this possibility.

This study has the following limitations. There is room for further improving the CNN, by using more advanced architectures such as using generative adversarial networks or encoder-decoder network structures, and by using even more extensive training strategies. On the other hand, the human observers can also be improved by using experienced radiologists and more elaborate training. It is also possible that the results would differ when using real medical images, as the anatomical clutter and varying lesion distribution may alter the performance of both the CNN and the human observer.

Moreover, it would have been preferable to use the same testing images for the CNN and the human observers. However, since the human observer study was limited to a random sub-selection of images and the CNN was trained on a much larger dataset, a perfect match between the two was impossible. Since

Fig. 7 Lift curves for human observers and CNN. A line is drawn to indicate the theoretical maximum performance for an observer. There are 10 points corresponding to the 10 noise levels for each observer (some are overlapping).

Fig. 8 Confusion matrices show both detection and characterization mistakes by the CNN. Green values represent the percentage of correct decisions and red values represent the percentage of incorrect decisions.

both testing libraries were generated to have the exact same characteristics, the differences were assumed to be insignificant.

These findings have important implications for clinical imaging. As artificial intelligence becomes more widespread, it is certain that this technology will find a place in the fields of radiology and diagnostic imaging. The results of this study indicate that applying deep learning technologies may be more appropriate for clinical images that have low noise correlation. In addition, deep learning is particularly useful for tasks involving low signal-to-noise ratio, which human observers are more likely to struggle with. In other words, a CNN is capable of seeing things where the human visual system is not capable of seeing anything. On the other hand, in the case of high noise correlation, a decision needs to be made whether a structure is true or whether it is artificially created by correlated noise. In this case, the relative performance between humans and CNN is less clear.

5 Conclusions

A series of experiments have been designed to compare the performance of standard neural networks with human observers in simple lesion detection and classification tasks. Generally, the CNN outperforms human observers. However, in the case of high noise correlation, their performance is more similar. We conclude that a neural network is more robust to uncorrelated noise than human observers, and the neural network and the human observers are confused by correlated noise and their performance becomes more similar.

Disclosures

The authors have no relevant conflicts of interest to disclose.

References

1. International Early Lung Cancer Action Program Investigators, "Survival of patients with stage I lung cancer detected on CT screening," *N Engl. J. Med.* **355**(17), 1763–1771 (2006).
2. T. Malati, "Tumour markers: an overview," *Indian J. Clin. Biochem.* **22**(2), 17–31 (2007).
3. H. V. Slooten et al., "Morphologic characteristics of benign and malignant adrenocortical tumors," *Cancer* **55**(4), 766–773 (1985).
4. S. Richard and J. H. Siewerdsen, "Cascaded systems analysis of noise reduction algorithms in dual-energy imaging," *Med. Phys.* **35**(2), 586–601 (2008).
5. R. De Man et al., "Upper-bound on dose reduction in CT reconstruction for nodule detection," *IEEE Access* **4**, 4247–4253 (2016).
6. D. Bardou, K. Zhang, and S. M. Ahmad, "Classification of breast cancer based on histology images using convolutional neural networks," *IEEE Access* **6**, 24680–24693 (2018).
7. K. Suzuki et al., "Computer-aided diagnostic scheme for distinction between benign and malignant nodules in thoracic low-dose CT by use of massive training artificial neural network," *IEEE Trans. Med. Imaging* **24**(9), 1138–1150 (2005).
8. D. R. Chen, R. F. Chang, and Y. L. Huang, "Computer-aided diagnosis applied to US of solid breast nodules by using neural networks," *Radiology* **213**(2), 407–412 (1999).
9. M. P. Eckstein and J. S. Whiting, "Lesion detection in structured noise," *Acad. Radiol.* **2**(3), 249–253 (1995).
10. K. J. Myers et al., "Effect of noise correlation on detectability of disk signals in medical imaging," *J. Opt. Soc. Am. A* **2**(10), 1752–1759 (1985).
11. A. E. Burgess, "Evaluation of detection model performance in power-law noise," *Proc. SPIE* **4324**, 123–132 (2001).
12. C. J. Kotre, "The effect of background structure on the detection of low contrast objects in mammography," *Br. J. Radiol.* **71**(851), 1162–1167 (1998).

13. S. Richard, J. H. Siewerdsen, and D. J. Tward, "NEQ and task in dual-energy imaging: from cascaded systems analysis to human observer performance," *Proc. SPIE* **6913**, 691311 (2008).
14. L. Yu et al., "Correlation between a 2D channelized Hotelling observer and human observers in a low-contrast detection task with multislice reading in CT," *Med. Phys.* **44**(8), 3990–3999 (2017).
15. J. B. Solomon and E. Samei, "Correlation between human detection accuracy and observer model-based image quality metrics in computed tomography," *J. Med. Imaging* **3**(3), 035506 (2016).
16. F. K. Kopp et al., "CNN as model observer in a liver lesion detection task for x-ray computed tomography: a phantom study," *Med. Phys.* **45**(10), 4439–4447 (2018).
17. F. Massanes and J. G. Brankov, "Evaluation of CNN as anthropomorphic model observer," *Proc. SPIE* **10136**, 101360Q (2017).
18. H. R. Roth et al., "Detection of sclerotic spine metastases via random aggregation of deep convolutional neural network classifications," in *Recent Advances in Computational Methods and Clinical Applications for Spine Imaging*, Springer, Cham, pp. 3–12 (2015).
19. J. Arevalo et al., "Representation learning for mammography mass lesion classification with convolutional neural networks," *Comput. Methods Progr. Biomed.* **127**, 248–257 (2016).
20. M. Vuk and T. Curk, "ROC curve, lift chart and calibration plot," *Metodoloski zvezki* **3**(1), 89–108 (2006).
21. S. Richard et al., "Predictive models for observer performance in CT: applications in protocol optimization," *Proc. SPIE* **7961**, 79610H (2011).
22. P. Prakash et al., "Task-based modeling and optimization of a cone-beam CT scanner for musculoskeletal imaging," *Med. Phys.* **38**(10), 5612–5629 (2011).

Ruben De Man is an undergraduate student in the Department of Biochemistry and Cell Biology at Stony Brook University. His current

research interests include machine learning for cancer screening and diagnosis, MRI-based studies on the effects of chemotherapy on the brain, and the identification of novel clinical markers for mortality risk due to advanced cancer.

Grace J. Gang obtained her PhD in Biomedical Engineering at the University of Toronto in 2014. Currently, she is Research Associate Faculty at Johns Hopkins University. Her main research interests involve image quality modeling, assessment, and optimization for advanced technologies in CT and cone-beam CT. Her work spans a wide range of topics including detector physics, image reconstruction, post-processing, and statistical decision theory. She has won numerous awards for her work.

Xin Li is a research scientist at Radiation Physics Lab at GE Global Research, focusing on providing solutions for various industrial NDT problems using x-ray and CT inspection. She graduated from Johns Hopkins University with a PhD of electrical and computer engineering in 2016, focusing on patient-specific image reconstruction optimization for SPECT imaging. Her expertise includes image quality evaluation, imaging system and imaging protocol design and optimization, image simulation, image analysis, etc.

Ge Wang is Clark & Crossan Endowed Chair Professor and director of Biomedical Imaging Center, Rensselaer Polytechnic Institute, United States. In 1991, he published the first spiral cone-beam/multi-slice CT algorithm and then systematically contributed +100 papers to this important area of CT research. Overall, he published 460+ journal papers covering diverse imaging-related topics. In 2016, he has published the first perspective on deep-learning-based tomographic imaging. He is fellow of IEEE, SPIE, OSA, AIMBE, AAPM, and AAAS.