Deep learning: Opening a third eye to myocardial perfusion imaging

Tomoe Hagio, Ph.D.,[a] and Venkatesh L. Murthy, M.D., Ph.D.[b]

[a] INVIA Medical Imaging Solutions, Ann Arbor, MI
[b] Division of Cardiovascular Medicine, Department of Internal Medicine, University of Michigan, Ann Arbor, MI

Received Mar 8, 2022; accepted Mar 9, 2022
doi:10.1007/s12350-022-02959-w

See related article, pp. 3300–3310

Deep learning (DL) methods are increasingly used in both research and practice of medicine for acquisition, segmentation, analysis, and interpretation of medical imaging studies, including nuclear cardiology.[1–3] A very large study from our group[3] recently reported the development and initial validation of a DL model which applies ''virtual'' attenuation correction to SPECT myocardial perfusion imaging (MPI) data and improves detection of coronary artery disease (CAD). One common question in response to hearing those findings was to pose the question of how a computer could ''see'' attenuation artifacts which may have been difficult to identify for human readers or standard quantification approaches. The study by Yeung and colleagues in the current issue of the *Journal of Nuclear Cardiology* may offer some insights into a related question. Specifically, this study expands the literature showing that DL algorithms can identify patient characteristics and predict clinical risk factors beyond what physicians could conventionally extract from images alone. For example, DL has been applied to other medical imaging tasks to predict mortality risk from fundoscopy images[4] and for quantifying bloo°d flow from optical images.[5]

Yeung *et al.* employ a two-step transfer learning approach to train their DL algorithm using polar maps as inputs, rest perfusion, stress perfusion, and myocardial flow reserve (MFR), to predict 19 outputs in each model: mean MFR in 17 segments, impaired MFR (< 2.0), and cardiovascular risk factors (sex, positive smoking status, hypertension, dyslipidemia, and diabetes mellitus). The first step was to use a publicly available pre-trained DL model, ResNet-50,[6] previously trained by others using ImageNet data which include photographic images and associated object classes.[7] Example images from ImageNet are shown in Figure 1. The pre-trained ResNet-50 model was fine-tuned on the weights in the later layers of the model and its associated hyperparameters to identify impaired regional MFR. This type of transfer learning approach has been useful, enabling training with only a fraction of the data required by traditional approaches. The second step of their training method was to further tune the model to classify individual cardiovascular factors. With a relatively small dataset (N = 851 for training and tuning the model and a test set of N = 93), transfer learning facilitated training while avoiding overfitting. Training against MFR seems circular but may have been a necessary step to further train the model to detect risk factors, which was the goal of this study.

In validating DL, it is often difficult to determine which results are uniquely enabled by the DL approach and which might have been feasible with more conventional approaches. This study compared DL results to conventional statistical methods (logistic regression models) for identifying cardiovascular risk factors. Both conventional models and DL were able to determine sex with the high accuracy using the rest polar map alone: the area under the curve (AUC) = 0.81 for DL and 0.85 for the logistic regression model. Both methods were also able to identify diabetes status, although marginally, with AUC of 0.65 for DL and 0.61 for logistic model using the reserve polar map as input. Interestingly, DL was also able to identify smoking status, whereas conventional logistic model was not able to;

Reprint requests: Tomoe Hagio, Ph.D., INVIA Medical Imaging Solutions, 3025 Boardwalk St, Suite 200, Ann Arbor, MI 48108; *thagio@inviasolutions.com*

J Nucl Cardiol 2022;29:3311–4.
1071-3581/$34.00

Figure 1. Example images from ImageNet.[7].

using rest polar map as the input, DL yielded AUC of 0.71 with the accuracy of 86%. Although the study was only validated on small population, the paper demonstrated the feasibility of DL application in MPI and showed valuable insight into how DL may enhance our ability to identify complex attributes from MPI images that have difficult to discern morphological features.

Nonetheless, it is important to be cautious in our interpretation of DL results. Like many other recent DL studies in nuclear cardiology,[8–10] this study was developed and validated using only small populations from a single center. Further evaluation with much larger populations from multiple centers is likely to show lower real-world performance in larger, more heterogeneous applications. Further, broader testing may identify patterns and patients more likely to systematically experience pitfalls with DL. Importantly, training and/or evaluating DL algorithms with diverse populations and with images acquired with a variety of protocols, scanners, and settings are necessary as it has been well established that DL algorithms may suffer from poor or unpredictable performance on unfamiliar datasets.[11–13] Even when trained with large amounts of data, DL may not be generalizable across different institutions as shown in Ref. [14] in which a DL algorithm for chest radiography performed significantly worse on external data compared to internal data (data from the same institution as training data). Furthermore, differences in disease prevalence in the training and test data may affect DL performance. A recent systematic review and meta-analysis compared DL performance to health-care professionals in detecting diseases from medical imaging.[15] Although equivalent diagnostic performance was shown overall, few studies presented externally validated results. Consequently, although many of DL results in the medical imaging literature may demonstrate a high diagnostic accuracy in initial studies, DL performance may have been biased and overestimated.[16]

Training DL can also be challenging even with large datasets if the gold standard is ill-defined. Finding high-quality, unbiased gold standards are often difficult. Although human readers are often used as the truth standard in many studies, e.g., disease diagnosis and image segmentation, this approach is expensive and time consuming. With the increased number of images used in developing and validating DL algorithms, this may not always be feasible. Further, individual human readers are also imperfect and there is considerable variability between readers. As such, DL approaches in other cardiac imaging modalities were able to exceed the performance of individual readers when trained against multiple readers.[17] However, having multiple readers' score each study can be prohibitive for large datasets. Fortunately, in this study, the gold standards were straightforward, albeit perhaps more interesting for our understanding of DL's strengths and weaknesses than for direct clinical application.

For PET/SPECT MPI, detection of CAD using invasive coronary angiography (ICA) is often used as the ground truth. While this is appealing in many ways, it raises many potential problems as well. First, angiographic findings will not identify many etiologies for hypoperfusion, including microvascular disease.[18] Perhaps even more problematic, very few patients are referred to angiography after stress testing—8.8% in one large study.[19] Further, there are major biases in referral for angiography in which women and minorities may be less likely to be referred.[20] The present study immediately demonstrates why this is a problem as DL can clearly be influenced by sex and likely also by race.[21] Consequently, DL algorithms trained to interpret MPI using angiography as a gold standard may be unfair to women and minorities and risk recapitulating existing disparities in care of these groups.

Overall, while it is not entirely clear that the results of this study by Yeung et al. are directly translatable to clinical practice, they do offer critical insights into the potential strengths and weaknesses of DL approaches in nuclear cardiology. While these algorithms offer great potential for improving accuracy of nuclear cardiology examinations, it will be essential to approach these in a manner which does not incorporate and perpetuate sex and race biases and subsequent health disparities. In other words, if DL algorithms potentially represent a third eye for interpretation of nuclear cardiology studies, we must ensure they are not blind to existing biases in care.

Disclosures

Dr. Murthy has received research grants and speaking honoraria from Siemens Healthineers. He owns stock in General Electric and Cardinal Health and stock options in Ionetix. He serves as a scientific advisor for Ionetix. He receives non-financial research support from INVIA Medical Imaging Solutions. Dr. Hagio is a full-time employee of INVIA Medical Imaging Solutions.

References

1. Betancur J, Commandeur F, Motlagh M, et al. Deep learning for prediction of obstructive disease from fast myocardial perfusion SPECT: a multicenter study. JACC Cardiovasc Imaging 2018;11:1654-63. https://doi.org/10.1016/j.jcmg.2018.01.020.
2. Otaki Y, Singh A, Kavanagh P, et al. Clinical deployment of explainable artificial intelligence of SPECT for diagnosis of coronary artery disease. JACC Cardiovasc Imaging 2021. https://doi.org/10.1016/J.JCMG.2021.04.030.
3. Hagio T, Poitrasson A, Jonathan R, et al. ‘‘ Virtual ’’ attenuation correction: improving stress myocardial perfusion SPECT imaging using deep learning. Eur J Nucl Med Mol Imaging In Press: 2022. https://doi.org/10.1007/s00259-022-05735-7.
4. Zhu Z, Shi D, Guankai P, et al. Retinal age gap as a predictive biomarker for mortality risk. Br J Ophthalmol 2022. https://doi.org/10.1136/BJOPHTHALMOL-2021-319807.
5. Braaf B, Donner S, Uribe-Patarroyo N, et al. A Neural network approach to quantify blood flow from retinal OCT intensity time-series measurements. Sci Rep 2020;101:1-13. https://doi.org/10.1038/s41598-020-66158-8.
6. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit 2016-December; 2016. p. 770–778. https://doi.org/10.1109/CVPR.2016.90
7. Russakovsky O, Deng J, Su H, et al. ImageNet Large Scale visual recognition challenge. Int J Comput Vis 2015;115:211-52. https://doi.org/10.1007/S11263-015-0816-Y/FIGURES/16.
8. Yang J, Shi L, Wang R, et al. Direct attenuation correction using deep learning for cardiac SPECT: a feasibility study. J Nucl Med 2021. https://doi.org/10.2967/jnumed.120.256396.
9. Shao W, Pomper MG, Du Y. A learned reconstruction network for SPECT imaging. IEEE Trans Radiat plasma Med Sci 2021;5:26-34. https://doi.org/10.1109/TRPMS.2020.2994041.
10. Kafouris PP, Koutagiar IP, Georgakopoulos AT, et al. Fluorine-18 fluorodeoxyglucose positron emission tomography-based textural features for prediction of event prone carotid atherosclerotic plaques. J Nucl Cardiol 2019;285:1861-71. https://doi.org/10.1007/S12350-019-01943-1.
11. Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks. In: 2nd international conference on learning representations, ICLR 2014: conference track proceedings. ICLR; 2014
12. Antun V, Renna F, Poon C, et al. On instabilities of deep learning in image reconstruction and the potential costs of AI. Proc Natl Acad Sci 2020. https://doi.org/10.1073/pnas.1907377117.
13. Serre T. Deep learning: the good, the bad, and the ugly. Annu Rev Vis Sci 2019;5:399-426. https://doi.org/10.1146/annurev-vision-091718-014951.
14. Zech JR, Badgeley MA, Liu M, et al. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. PLoS Med 2018. https://doi.org/10.1371/JOURNAL.PMED.1002683.
15. Liu X, Faes L, Kale AU, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. Lancet Digit Heal 2019;1:e271-97. https://doi.org/10.1016/S2589-7500(19)30123-2/ATTACHMENT/49AB10F2-3AA4-4101-A155-11A5ED6772BC/MMC1.PDF.
16. Aggarwal R, Sounderajah V, Martin G, et al. (2021) Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. NPJ Digit Med 2021;4:1-23. https://doi.org/10.1038/s41746-021-00438-z.
17. Howard JP, Stowell CC, Cole GD, et al. Automated left ventricular dimension assessment using artificial intelligence developed and validated by a UK-wide collaborative. Circ Cardiovasc Imaging 2021;14:405-15. https://doi.org/10.1161/CIRCIMAGING.120.011951.

18. Murthy VL, Bateman TM, Beanlands RS, et al. Clinical quantification of myocardial blood flow using PET: joint position paper of the SNMMI Cardiovascular Council and the ASNC. J Nucl Cardiol 2018;25:269-97. https://doi.org/10.1007/S12350-017-1110-X.
19. Mudrick DW, Cowper PA, Shah BR, et al. Downstream procedures and outcomes after stress testing for suspected coronary artery disease in the United States. Am Heart J 2012;163:454-61. https://doi.org/10.1016/J.AHJ.2011.11.022.
20. Chulman EAS, Erlin EAB, Arless IH, et al. The effect of race and sex on physicians' recommendations for cardiac catheterization. N Engl J Med 2008;340:618-26. https://doi.org/10.1056/NEJM199902253400806.
21. Puyol-Antón E, Ruijsink B, Mariscal Harana J, et al. Fairness in cardiac magnetic resonance imaging: assessing sex and racial bias in deep learning-based segmentation. medRxiv 2021. https://doi.org/10.1101/2021.07.19.21260749.