Editorial Commentary

Toward the transparency of deep learning in radiological imaging: beyond quantitative to qualitative artificial intelligence

Yoichi Hayashi

Department of Computer Science, Meiji University, Kawasaki, Japan

Correspondence to: Yoichi Hayashi. Department of Computer Science, Meiji University, 1-1-1 Higashimita, Tama-ku, Kawasaki 214-8571, Japan. Email: hayashiy@cs.meiji.ac.jp.

Comment on: Wang CJ, Hamm CA, Savic LJ, *et al.* Deep learning for liver tumor diagnosis part II: convolutional neural network interpretation using radiologic imaging features. Eur Radiol 2019;29:3348-57.

Received: 11 August 2019; Accepted: 17 September 2019; Published: 30 September 2019.

doi: 10.21037/jmai.2019.09.06

View this article at: http://dx.doi.org/10.21037/jmai.2019.09.06

In the near future, nearly every type of clinician, from paramedics to certificated medical specialists, will be expected to utilize artificial intelligence (AI) technology, and deep learning (DL) in particular (1). In terms of exceeding human ability, DL has been the backbone of computer science. DL mostly involves automated feature extraction using deep neural networks (DNNs), which can aid in the classification and discrimination of medical images, including mammograms, skin lesions, pathological slides, radiological images, and retinal fundus photographs.

A key differentiating feature of DL (1) compared with other types of AI is its autodidactic quality; neural networks (NNs) are not designed by humans (2). NNs, which are considered to have a "black box" nature, have been made more transparent by techniques first applied to shallow NNs. An explanation to NN responses is through using propositional rules (3). Andrews *et al.* (4) developed a taxonomy describing the general features of all rule extraction methods. AI can analyze massive images and patient data which a single radiologist could not.

However, as a set of parameters, DNNs learn to produce an output on their own. Known inputs and algorithms of AI programs start the process, while the resulting parameters are hard to interpret (2). These "black box" problems lead to opaqueness in DL. The aim of this editorial commentary is to help realize the transparency of "black box" machine learning for radiologic imaging. To achieve this, a renewed attack is undertaken on the "black box" problem and the limitations of DL for radiological imaging, and an attempt is made to reveal a paradigm shift in radiological imaging in which diagnostic accuracy is surpassed to achieve interpretability, which is highly important for predictive models.

As the "black box" nature of DL in medicine has been strongly criticized, especially in the radiology field, the new "black box" problem caused by highly complex DNNs must be addressed, and for any solution, transparency and interpretability are needed. However, at present, a number of "black box" problems remain in relation to DNNs (5). A large body of research has been conducted on the "black box" of algorithms, and this topic continues to generate substantial controversy.

Especially in the case of DNNs, it is not possible to understand the determination of output. In contrast to computer vision tasks, DL in the field of radiology remains considerably limited in terms of its interpretability and transparency. Owing to the "black box" nature of DL, where results can achieve high accuracy, but with no specific medical-based reasoning, effectively interpreting and applying DL to radiological images in the clinical setting requires sufficient expertise in computer science. Owing to this, the results from DL can be hard to interpret in clinical, limiting their ability to be used in medical decision-making (5).

Some researchers have highlighted the importance of accuracy over interpretability; however, it is the view of the present paper that improved transparency of DL would encourage the universal accept of such methods for medical imaging (5).

Context also plays a part. If the life-and-death decisions made by systems only provided trivial improvements in

accuracy over human, then greater transparency would be warranted compared with decisions made with near-perfect accuracy or for lower stakes (6). Especially in the medical field in which accountability is of crucial importance and could lead to severe legal consequences, DL is usually insufficient when used for prediction. In addition, in terms of outcome prediction, predictive radiotherapy based on DL could still be a long way off. First, radiation oncologists must acquire the ability to understand predictions using DL algorithms; whereas their interpretation frequently remains not easy as these are still thought "black boxes" (6).

The development of an interpretable proof-of-concept DL system for clinical radiology was recently reported by Wang *et al.* (7). Their prototype allows the automatic identification, mapping, and scoring of radiological features, thereby allowing radiologists to understand the elements of decision-making behind classification decisions. This concept of interpretability is the first major contribution of Wang *et al.*'s work.

DL algorithms could substantially improve the clinical workflow of diagnosis, prognosis, and treatment. However, transparency is vital in this process. Indeed, clinicians would be unlikely to accept automated decisions of diagnosis without measuring evidence to justify the predictions (7).

A general relationship between the misclassification of a lesion entity and the misidentification of radiological features was observed in Wang *et al.*'s model; this could present the transparency necessity to identify how and when a convolutional neural network (CNN) model fails. If a model were to predict nonexistent imaging features, clinicians would realize that it had likely made a mistake (7).

Considering the fact that researchers and clinicians should be made aware of such nonexistent imaging features, the nature of radiological images should be recognized as being considerably different from that of computer vision images.

For example, the earliest work using a CNN in the field of medical imaging had the same limitations as those seen for detecting diabetic retinopathy (DR) (8). A fundamental drawback inherent to DNNs is that, in regard to DR, the NN is not provided with any explicit definitions of the features to explain the medical diagnosis. The quality of the image is judged by graders using the rubric in the Grading Instructions, while the severity is graded based on the International Clinical Diabetic Retinopathy scale; thus, the diagnostic process is a "black box" (8).

As a more constructive way to resolve this "black box" nature, specific features can be generated from six classes of liver tumor samples by analyzing magnetic resonance imaging (MRI) scans; this could provide radiologists with guidance for detection and diagnosis. These distinctive features, called semantic features (9), can be used to create predictors of liver tumors. Thus, the second major contribution of Wang *et al.*'s work was the proposal of a concrete method to capture from a dataset the semantic features of six classes of liver tumor samples.

In an approach known as radiomics, quantitative information can be generated from liver tumors using MRI scans, and then analyzed using machine learning or high-dimensional data analysis and categorized into different groups. Traditional quantitative features can be used for the creation of biomarkers for tumor prognosis, analysis, and prediction.

However, even in the case that such a method enables the accurate assignment of instances to groups, it cannot provide users with the reasoning underlying that assignment. Therefore, systems and/or algorithms that are able to provide insights into these underlying reasons are needed (10).

It is the view of the present paper that traditional quantitative features generated from radiological images using MRI scans and computer vision images can be used only for classification tasks. After a machine learning algorithm has been trained, it still remains difficult to understand why it provides a particular response to the training dataset, and this can be a disadvantage, especially in the medical setting, because the main task of state-of-the-art machine learning algorithms such as DL is to achieve very high accuracy in classifying datasets, for example, into six separate classes of liver tumors. Machine learning algorithms such as DL are currently incapable of explaining their classification results.

By contrast, rule extraction (4), a newer branch of machine learning that utilizes AI, focuses on how the entire dataset is classified. In rule extraction, the rules are typically expressed as the most popular and comprehensible symbolic descriptions: "if (conditions 1) & (conditions 2), … & (condition n), then (target class)". Rule extraction algorithms in the medical field require a sufficient number of cases and their final diagnoses as a supervised signal (specific class of liver tumor) for learning.

In recent years, CNNs have been used effectively in regard to liver tumor analysis (11). However, data are currently scarce in the medical imaging field, so transfer learning has been used as an alternative to constructing a new model.

CNN convolution layers contain representations of edge gradients and textures after learning. When propagated through fully connected layers, a variety of high-level features are learned by the CNN. Then, deep features (the outputs of units in the layer) are extracted from fully connected layers and denoted by the number of the feature from the learning tool (12).

To utilize deep features, the present author recently devised a new method known as deep belief network (DBN) Re-RX with J48graft (13) to extract interpretable and accurate classification rules from DBNs (14). This method was applied to three small, high-abstraction, rating category (semantic or structured) datasets with prior knowledge, i.e., semantic features (9): the Wisconsin Breast Cancer Dataset (WBCD), the Mammographic Mass dataset, and the Dermatology dataset. After these three datasets were trained, a rule extraction method capable of extracting accurate and concise rules for DNNs trained by a DBN was proposed. The rationale behind this method is based on deep features (12) and the large margin principle (15) for shallow NNs.

The results indicated that the Re-RX family (16) could help bridge the divide between the high learning capability of DBNs and the high interpretability of rule extraction algorithms such as Re-RX with J48graft (10,17). This could lead to a better trade-off between predictive accuracy and interpretability. This method can be applied to not only ratings categories, but also image datasets consisting of semantic features. Although traditional quantitative features do not provide sufficient high-level abstraction for input attributes, semantic features, which include prior knowledge graded and/or rated by radiologists, can be useful for input features, as demonstrated in the present author's work (13).

In addition, applying fully connected layer-first CNNs that the fully connected layers are imbedded before the first convolution layer, DBN Re-RX with J48graft can be extended to CNN Re-RX for high-level abstraction datasets (deep features) (18) for the Re-RX family (16) uses decision trees like C4.5 (19) and J48graft (C4.5A) (20).

Generally, irrespective of the input and output layers in any type of DL for high-level abstraction images with prior knowledge (semantic features) (5), rules can be extracted using pedagogical (4) ways such as C4.5, J48graft, the Re-RX family, Trepan (21), and ALPA (22).

However, the provision of radiological images is often insufficient in a large number of abstraction datasets with prior knowledge (semantic features). This difficulty may be avoided by noticing the high-level abstraction of attributes (semantic features) related to the radiological images. According to the present author, since semantic features could bridge the gap between a priori knowledge and rule extraction, the most important point in realizing the transparency of DL in radiological images is the use of the high-level abstraction of attributes (deep and/or semantic features) related to medical images with known knowledge graded and/or rated by radiologists, not the fact that driven characteristics depend on filter responses solicited from massive training data; this suffers from a shortage of direct human interpretability. Wang *et al.*'s work thus provided important insights to generate useful semantic features in rule extraction.

Furthermore, Lee *et al.* (23) shared opinions on ways for optimizing DL model performance. Previous work has proved that deeper NNs deliver better visual recognition performance than shallower NNs when training datasets are kept constant (1). Simply picking the deepest NN (12), whereas, was not the answer in the study by Lee *et al.* (23). By using preprocessing and network optimization techniques, they were able to achieve much greater performance gains compared with the small incremental improvements gained using deeper and more complex NNs (1). Their results suggest that application-specific customization techniques are more effective than the choosing the underlying CNN architecture to improve performance. Their results also suggest that high accuracy classification by DL and high interpretability by rule extraction should be used for different purposes in radiological imaging.

The strengths of Wang *et al.*'s study (7) were that it provided an effective technique for interrogating concerned portions of an existing CNN and offered a rationale for classification through analyzing relevant features. Their ideas can be adapted to wider interfaces with standardized reporting systems like the Breast Imaging Reporting and Data System (24).

On the other hand, the weakness of Wang *et al.*'s study is that their method does not use deep features concretely in comparison with the method proposed by Paul *et al.* (9). By contrast, they attempted to relate and explain deep features with respect to quantitative and semantic features.

Although the computational power of CNNs is typically supplied to create quantitative features by so-called future engineering using many graphical processing units, the works by Wang *et al.* (7) and Paul *et al.* (9) are expected to increase the degree of abstraction of semantic and deep features used for rule extraction and to accelerate

the explainable AI boom in medical imaging beyond quantitative to qualitative AI.

Acknowledgments

Funding: None.

Footnote

Provenance and Peer Review: This article was commissioned by the editorial office, *Journal of Medical Artificial Intelligence*. The article did not undergo external peer review.

Conflicts of Interest: The author has completed the ICMJE uniform disclosure form (available at http://dx.doi.org/10.21037/jmai.2019.09.06). The author has no conflicts of interests to declare.

Ethical Statement: The author is accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

References

1. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436-44.
2. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019;25:44-56.
3. Bologna G. A simple convolutional neural network with rule extraction. Appl Sci 2019;9:2411.
4. Andrews R, Diederich J, Tickle AB. Survey and critique of techniques for extracting rules from trained artificial neural networks. Knowl Based Syst 1995;8:373-89.
5. Hayashi Y. The right direction needed to develop white-box deep learning in radiology, pathology, and ophthalmology: a short review. Front Robot AI 2019;6:24.
6. Shickel B, Tighe PJ, Bihorac A, et al. Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. IEEE J Biomed Health Inform 2018;22:1589-604.
7. Wang CJ, Hamm CA, Savic LJ, et al. Deep learning for liver tumor diagnosis part II: convolutional neural network interpretation using radiologic imaging features. Eur Radiol 2019;29:3348-57.
8. Gulshan V, Peng L, Coram M, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA 2016;316:2402-10.
9. Paul R, Schabath M, Balagurunathan Y, et al. Explaining deep features using radiologist-defined semantic features and traditional quantitative features. Tomography 2019;5:192-200.
10. Hayashi Y, Nakano S. Use of a recursive-rule eXtraction algorithm with J48graft to achieve highly accurate and concise rule extraction from a large breast cancer dataset. Inform Med Unlocked 2015;1:9-16.
11. Hamm CA, Wang CJ, Savic LJ, et al. Deep learning for liver tumor diagnosis part I: development of a convolutional neural network classifier for multi-phasic MRI. Eur Radiol 2019;29:3338-47.
12. Giryes R, Sapiro G, Bronstein AM. Deep neural networks with random gaussian weights: a universal classification strategy? IEEE Trans Signal Process 2016;64:3444-57.
13. Hayashi Y. Use of a deep belief network for small high-level abstraction data sets using artificial intelligence with rule extraction. Neural Comput 2018:1-18.
14. Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. Science 2006;313:504-7.
15. Vapnik VN. The nature of statistical learning theory. New York: Springer, 1995.
16. Hayashi Y. Application of a rule extraction algorithm family based on the Re-RX algorithm to financial credit risk assessment from a Pareto optimal perspective. Operations Research Perspectives 2016;3:32-42.
17. Hayashi Y. Synergy effects between the grafting and the subdivision in Re-RX with J48graft for the diagnosis of thyroid disease. Knowl Based Syst 2017;131:170-82.
18. Liu K, Kang G, Zhang N, et al. Breast cancer classification based on fully-connected layer first convolutional neural networks. IEEE Access 2018;6:23722-32.
19. Quinlan JR. C4.5: programs for machine learning. San Mateo: Morgan Kaufmann, 1993.
20. Webb GI. Decision tree grafting from the all-tests-but-one

partition. In: Proceedings of the 16th International Joint Conference on Artificial Intelligence. San Mateo: Morgan Kaufmann, 1999;2:702-7.
21. Craven JM, Shavlik J. Extracting tree-structured representations of trained networks. In: Touretzky DS, Mozer MC, Hasselmo ME. editors. Advances in neural information processing systems. Cambridge: MIT Press, 1996:24-30.
22. de Fortuny EJ, Martens D. Active learning-based pedagogical rule extraction. IEEE Trans Neural Netw Learn Syst 2015;26:2664-77.
23. Lee H, Yune S, Mansouri M, et al. An explainable deep-learning algorithm for the detection of acute intracranial haemorrhage from small datasets. Nat Biomed Eng 2019;3:173-82.
24. Obenauer S, Hermann KP, Grabbe E. Applications and literature review of the BI-RADS classification. Eur Radiol 2005;15:1027-36.

doi: 10.21037/jmai.2019.09.06
Cite this article as: Hayashi Y. Toward the transparency of deep learning in radiological imaging: beyond quantitative to qualitative artificial intelligence. J Med Artif Intell 2019;2:19.

