Informatics in Medicine Unlocked 19 (2020) 100329

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

journal homepage: http://www.elsevier.com/locate/imu

New unified insights on deep learning in radiological and pathological images: Beyond quantitative performances to qualitative interpretation

Yoichi Hayashi *

Department of Computer Science, Meiji University, Kawasaki, 214-8571, Japan

ARTICLE INFO

Keywords:
Deep learning
Pathology
Rule extraction
Explainable AI
Black box
Interpretability

ABSTRACT

Deep learning (DL) has become the main focus of research in the field of artificial intelligence, despite its lack of explainability and interpretability. DL mainly involves automated feature extraction using deep neural networks (DNNs) that can classify radiological and pathological images. Convolutional neural network (CNNs) can be also applied to pathological image analysis, such as the detection of tumors and the quantification of cellular features. However, to our knowledge, no attempts have been made to identify interpretable signatures from CNN features, and few studies have examined the use of CNNs for cytopathology images. Therefore, the aim of the present paper is to provide new unified insights to aid the development of more interpretable CNN-based methods to classify radiological and pathological images and explain the reason for this classification in the form of *if-then* rules. We first describe the "black box" problem of shallow NNs, the concept of rule extraction, the renewed attack of the "black box" problem in DNN architectures, and the paradigm shift regarding the transparency of DL using rule extraction. Next, we review limitations of DL in pathology in regard to histopathology and cytopathology. We then investigate the discrimination of cytological features and explanations and review recent techniques for interpretable CNN-based methods in histopathology, as well as current approaches being taken to enhance the interpretability of CNN-based methods for radiological images. Finally, we provide new unified insights to extract qualitative interpretable rules for radiological and pathological images.

1. Introduction

Despite its lack of explainability and interpretability, deep learning (DL) is currently a main focus of research in the field of artificial intelligence (AI). In DL, appropriate feature models can be learned from data in the form of convolution filters or multi-dimensional embedding vectors [1]. In terms of exceeding human ability, DL has been the backbone of computer science. DL largely involves automated feature extraction using deep neural networks (DNNs) that can classify various images in radiology and pathology. A convolutional neural network (CNN) [2] has an input layer, an output layer, and a number of hidden layers, each corresponding to different image features. Hinton and Salakhutdinov [3] noted that with good data, unsupervised training of deep belief networks (DBNs) followed by a pass with backpropagation (BP) [4] can achieve better accuracy.

A key differentiating feature of DL [5] compared with other types of AI is its excellent quantitative performance (accuracy); that is, DNNs are not designed by humans [6]. Previous CNN-based methods are typically trained to process image pixels and predict disease labels. Despite its recognized human-level classification accuracy for several diseases, using DL as a diagnostic prediction mechanism is discouraged because of its lack of interpretability (qualitative interpretation) in regard to CNN-based features [7]. CNNs can be also applied to pathological image analysis, such as the detection of tumors and the quantification of cellular features. Diagnosing pathology slides is a complicated task requiring experienced pathologists. Unlike other types of medical images (e.g., radiological images), digital pathology (DP) slides are obtained at very high resolutions (>10 gigapixels) [8,9].

Moreover, pathologists and researchers who rely on nuclei/cell organization or tissue structure or morphology to characterize disease do not find CNNs interpretable, because traditional CNNs do not translate these features precisely. Observing feature activation in each layer of the

Abbreviations: DL, deep learning; ML, machine learning; CNN, convolutional neural network; DBN, deep belief network; DNN, deep neural network; Re-RX, Recursive-Rule eXtraction; Re-RX with J48graft, Recursive-Rule eXtraction algorithm with the J48graft decision tree; NN, Neural network; DT, decision tree; AI, artificial intelligence; DP, digital pathology; WSI, whole slide image; MRI, magnetic resonance imaging; 2D, two-dimensional.

* , Department of Computer Science, Meiji University, 1-1-1 Higashimita, Tama-ku, Kawasaki, 214-8571, Japan.
E-mail address: hayashiy@cs.meiji.ac.jp.

https://doi.org/10.1016/j.imu.2020.100329
Received 11 February 2020; Received in revised form 10 April 2020; Accepted 17 April 2020
Available online 23 April 2020

network can illuminate relationships between CNN- and pathology-driven features, but to our knowledge, no attempts have been made to identify conclusively interpretable signatures from CNN features [10]. These difficulties (the so-called "black box" problem) cause opaqueness in DL. Therefore, the reasons why "black box" types of machine learning (ML), such as CNNs, work well for classification tasks in pathology should be clarified. DL is an active research field, and the application of DL to histopathology is relatively new. However, few studies have examined the use of CNNs for cytopathology images [11]. Therefore, the aim of this paper is to provide new insights to aid the development of more interpretable CNN-based methods to classify pathological images and explain the reason for this classification in the form of *if-then* rules.

In this paper, we first describe the "black box" problem of shallow NNs, the concept of rule extraction, the renewed attack of the "black box" problem in DNN architectures, and the paradigm shift regarding the transparency of DL using rule extraction and DL-based rule extraction. Next, we review the limitations of DL in pathology and review recent techniques for interpretable CNN-based methods in histopathology, and examine the approaches currently being taken to enhance the interpretability of CNN-based methods for radiological images. Finally, we provide new insights to aid in the extraction of both qualitative interpretable rules for radiological images and qualitative interpretations for pathological images.

The main contribution of this paper is that it provides new insights for the qualitative interpretation of radiological and pathological images *following the same principle* and that can be *unified* from the perspective of rule extraction technology.

2. "Black box" problems of shallow NNs and rule extraction

The "black box" nature of shallow NNs (e.g., multi-layer perceptrons) with one hidden layer—i.e., they have no explicit declarative knowledge representation—is a major drawback. It is considerably difficult for shallow NNs to generate adequate explanation structures, which limits the full potential of such systems. However, a detailed characterization of classification strategies could contribute to their acceptance. Extracting symbolic rules is a natural way to elucidate the knowledge embedded within NNs. Since about 1990, many researchers have made extensive efforts to resolve the "black box" problem of trained NNs using rule extraction technology [12–19].

Rule extraction is not a new concept, being originally conceived by Gallant [20] for a shallow NN for the medical domain. Rule extraction [13] is a powerful and increasingly popular method of data mining that provides explainable and interpretable capabilities for models generated by shallow NNs. Extracted rules need to be simple and interpretable by humans, and to be able to discover highly accurate knowledge in the medical domain. Rule extraction technology has also been recognized as a technique that attempts to find a compromise between the two requirements of accuracy and interpretability by building a simple rule set that mimics how a well-performing complex model (a "black box") makes decisions for users [16].

Recently, as a promising means to address the "black box" problem, a rule extraction technology well balanced between accuracy and interpretability was proposed for shallow NNs [17,19]. In addition, Uehara et al. [21] reported an actual medical application in hepatology using rule extraction. Hayashi et al. [22] reported a rule extraction approach to explore the upper limit of hemoglobin during anemia treatment in patients with predialysis chronic kidney disease. Hayashi [23] also proposed a method to detect lower albuminuria levels and the early development of diabetic kidney disease using an AI-based rule extraction approach.

3. Renewed attack of the "black box" problem in deep neural network architectures

Algorithms in DL require only a labeled training set (e.g., pixel data and corresponding diagnosis labels from cell images). However, DL also has a trade-off, in that the learned features do not correspond to features typically understandable by humans, because these depend on complex interactions with other uninterpreted features [11].

In predictive models, interpretability is important, and thus, the "black box" nature of DL has been severely criticized in medical settings [6]. In regard to DL, explainability refers to the ease with which how and why a decision is made by ML can be explained, and as such, is imperative for DL techniques that can make life-altering decisions in the clinical setting. Explainable decisions for DL classifications that can be understood by humans would enable physicians to correct any decision made by an AI system [24].

Therefore, the "new black box" problem caused by highly complex DNN models generated by DL must be addressed. To resolve this "new black box" problem, transparency and interpretability are needed for *any type of* DNN; however, at present, various "black box" problems remain [25]. Much has been written, and substantial controversy still exists, about this "black box" problem, especially in the case of DNNs, in which the determination of output cannot be understood [25]. Unexpectedly, very little work has been achieved with respect to DNNs. Filling this gap may contribute to the real-world usability of DNNs.

Visualization methods such as saliency maps [26] and the gradient-weighted class activation mapping (Grad-CAM) technique [27] can be useful for determining which part of an image is being omitted by the classifier; however, such methods exclude all information about how relevant information is being used. Recently, the "black box" of DL on co-saliency has been investigated in the field of computer vision [28,29].

Knowing where a network is looking within the image does not tell the user what it is doing with that part of the image. In fact, saliency maps could be essentially the same for multiple classes [30]. Moreover, when applied to the classification of cancer images, the accuracy of visualization models in terms of explaining the decisions made by DL algorithms has not been evaluated quantitatively. Therefore, whether highly accurate image classification with DNNs is equivalent to the actual identification of cancer regions remains unclear [27].

4. A paradigm shift regarding the transparency of DL using rule extraction

As shown in Fig. 1, a paradigm shift regarding the transparency of DL using rule extraction has been clearly demonstrated for the MNIST data set [31], which is a difficult problem for rule extraction because the inputs (attributes) are 784 very low-level abstraction pixels in the images (unstructured data set) that have to be classified into 10-digit classes. The rules must therefore capture the "hidden" low-level abstraction learned by DL. Such image domains are notoriously difficult for symbolic reasoning [32]. As shown in the figure, starting with the "black box" nature of DL, they first achieved a low level of transparency (interpretability) [33], followed by a considerable level of transparency (interpretability) [34].

5. DL-based rule extraction

Regarding the attainment of a lower minimum for the empirical cost function, Erhan et al. [35] reported that the optimization process can be rendered more effective through unsupervised pre-training that can initialize a model to a point in the parameter space; the same difficulties are often experienced during the BP learning process. Therefore, the learning of input information from the feature space by the DBN during the supervised learning phase could allow the BP neural network (NN) to converge an objective function into a near good local optimum referred to as a DBN-NN, and this could be the rationale behind the enhancement made possible by a simple idea [36].

Generally, a large margin principle [37] can be applied to rating category (structured) data sets with large numbers of features (attributes), such as biomarkers or radiologists' readings. In fact, a new

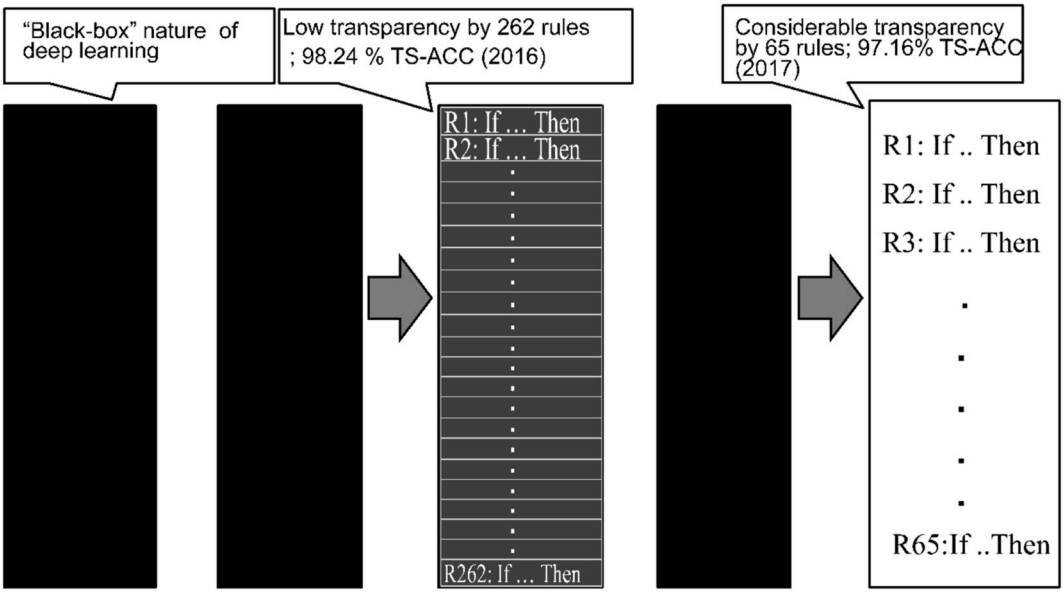

Fig. 1. A paradigm shift regarding the transparency of DL using rule extraction. for the MNIST data set.

method to extract accurate and interpretable classification rules for DBNs, known as DBN Re-RX with J48graft shown in Fig. 2, has recently been proposed [38]. This method was applied to a rating category (structured) data set [39], the Wisconsin Breast Cancer Data Set (WBCD), which is a small, high-abstraction data set with prior knowledge. After training the data set, a rule extraction method that could extract accurate and concise rules for DNNs trained by a DBN was proposed. The results suggested that the Re-RX family [18] could help fill the gap between the very high learning capability of DBNs and the very high interpretability of rule extraction algorithms such as Re-RX with J48graft [19]. Therefore, a better trade-off between predictive accuracy and interpretability can be achieved in not only rating category data sets, but also image data sets consisting of relatively high-level abstract features.

We can extend DBN Re-RX with J48graft to "CNN Re-RX" for high-level abstraction data sets using fully connected layer-first CNNs, in which the fully-connected layers are embedded before the first convolution layer [40] because the Re-RX family [18] uses decision trees (DTs) such as C4.5 [41] or J48graft [42]. In general, we can extract rules using pedagogical [13] approaches such as C4.5, J48graft, the Re-RX family, Trepan [43], and ALPA [16], regardless of the input and output layers in *any type of* DL for images with high-level abstraction attributes with prior knowledge. For more details, we will present a concrete method in another paper.

6. Limitations of deep learning in histopathology

CNN architectures, which show high performance compared with other conventional architectures, are widely accepted for use in both high- and low-resolution texture image analysis. Compared with radiological image processing and analysis, that for histopathology involves more difficult issues because histopathology images have a very muddled structure compared with other images [44].

DP concerns all aspects of the processing of digitized histopathology slides, including image analysis. DP has the potential to improve the accuracy, interpretability, and reproducibility, as well as the efficient throughput of diagnostic histology slides.

However, in contrast to other modalities in which entire images can

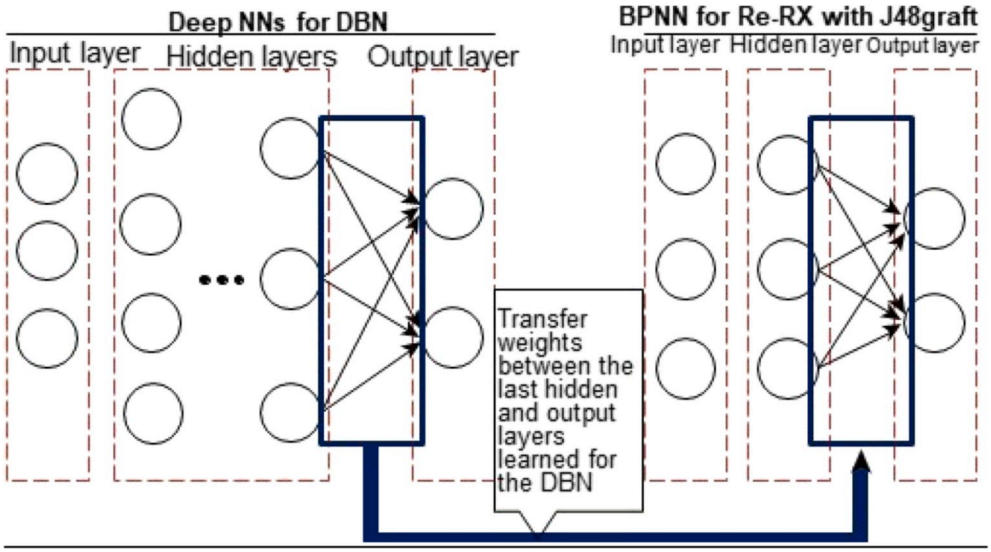

Fig. 2. Overview of DBN Re-RX with J48graft.

be used as inputs, such as magnetic resonance imaging (MRI), the classification of digitized histopathologic images presents a challenge in processing because of their extremely large size [44]. Thus, the size of digitized histopathology images and the lack of large detailed and consistently labeled data sets for ML training present ongoing challenges to its widespread application [9]. The development of AI algorithms and training of classifiers are therefore dependent on the correct labeling (ground truth) of digitized slides.

Because of the "black box" nature of DL, where results are generated with high accuracy, but with no specific medical-based reasons, sufficient expertise in computer science is required to interpret and apply DL to digitized histopathology in the clinical setting.

6.1. Detection and classification of cancer in whole slide breast histopathology images

The generalizability of algorithms for binary cancer compared with no cancer classification remains unclear for more clinically significant multi-class scenarios in which intermediate categories have different risk factors and treatment strategies. Gecer et al. [45] reported the use of a system that classifies whole slide images (WSIs) of breast biopsies into five diagnostic categories. However, this five-category classification formulation is a novel technique, and thus, the diagnostic process remains a "black box."

6.2. Dimensionality-reduction approach for visually interpreting the highest ranking feature in gliomas

Attempts to standardize the visual interpretation of malignant gliomas for tissue classification have resulted in the creation of a rule-based lexicon to improve the reproducibility of interpretation called the Visually Accessible Rembrandt Images feature set. However, such approaches are also limited because of their need for human visual interpretation and a priori feature selection, which innately distills a complex data set (>1 million voxels per image) into a small number of numeric descriptors [46].

Chang et al. [46] used a DL approach to classify individual somatic mutations of diffuse infiltrating gliomas. They applied a dimensionality-reduction approach to display visually the highest ranking feature of each mutation category. They introduced a new technique to visualize the imaging features most relevant to the classification of genetic mutation status using principal component analysis as a means of dimensionality reduction and disentanglement of the final feature vector layer.

7. Recent techniques for interpretable CNN-based methods in histopathology

7.1. Use of grading and scoring in pathology

The standard practice of microscopy for the diagnosis, grading, and staging of cancer has remained nearly unchanged over the past century [47]. Tumor stage and the Gleason score are still the most powerful prognostic predictors in virtually every large prostate cancer outcome study [48]. Despite its role in prognostication and patient management, Gleason scoring remains a subjective exercise carried out by pathologists, and suffers from suboptimal inter- and intra-observer variability.

One potential approach for improving the accuracy and consistency of Gleason grading can be found in the field of AI, where recent advances using DL have been successfully applied to imaging diagnostic tasks in fields such as dermatology [49], ophthalmology [50], radiology [51], and histopathology [45,52,53].

7.2. Beyond cancer detection and tumor grading to clinical-grade computational pathology

Most AI research in pathology still focuses on cancer detection and tumor grading; however, pathological diagnosis is a complex process involving the evaluation and judgment of various types of clinical data dealing with various organs and diseases, not simply a morphological diagnosis [1].

Recent studies [54–56] have referred to clinical grade through comparisons, usually under time or other types of constraints, with humans performing the same task. Campanella et al. [57] suggested that these comparisons offer little insight into how such systems can be used in actual clinical practice. Therefore, they developed a novel framework that leverages multiple instance learning (MIL) to train DNNs, which resulted in semantically rich, tile-level feature representations that could be used in a recurrent NN.

7.3. Hybrid approaches involving CNNs combined with handcrafted features for both detection and grading

Hybrid approaches involving CNNs combined with handcrafted features for both the detection and grading of prostate cancer have achieved accuracies ranging from 75% to 89%. However, the development of AI algorithms and classifier training depends on the correct labeling of digitized slides. Unfortunately, due to differences in training, experience, institutional guidelines, and protocols among pathologists, the grading of prostate cancer is known to be susceptible to both inter- and intra-observer variability [9].

7.4. High-level interpretable approaches in histopathology

Zhang et al. [7] developed a novel method for the automation of whole slide reading and pathological diagnosis via region-level tumor detection and pixel-level morphological analysis of nuclear anaplasia and architectural abnormalities, resulting in the establishment of slide-level diagnosis. Each process was powered by a DNN, where cascading progressively encodes enormous pixels into meaningful and compact representations. Beyond simply predicting diagnosis labels, their method involves interpretability mechanisms that decode learned representations into rich interpretable predictions that can be understood by pathologists [7].

8. Limitations of deep learning in cytopathology

Computed tomography (CT) has increasingly been used to screen for various types of cancer. If a CT examination identifies a suspicious lesion, pathological diagnosis is carried out to examine the lesion in greater detail. In this procedure, cytological diagnosis is performed by cytotechnologists and cytopathologists to determine a pathology.

Although most recent efforts in applying DL to pathology have focused mainly on histopathology, to the best of our knowledge, few studies have investigated the classification of cytopathology images [58,59]. In particular, the automated detection of cell nuclei remains extremely difficult because of the overlapping of cells, as does cytological diagnosis. Moreover, in cytological diagnosis, clearly classifying cells as benign or malignant remains difficult because of the existence of atypical cells [59].

Sanghvi et al. [60] recently developed a CNN algorithm that could accurately analyze WSIs of urine cytology cases. Compared with existing approaches, their algorithm used a much larger data set, exploited whole slide-level as opposed to only cell-level features, and utilized a cell gallery to display the output for easy end-user review.

An automated method for the classification of lung cells in cytological images was recently developed by Teramoto et al. [59], who previously proposed an automated method for the classification of lung cancer types using DL. In this method, they applied a deep CNN (DCNN)

architecture to classify cytological images into adenocarcinoma, squamous cell carcinoma, and small cell carcinoma. As a result, they obtained an approximate accuracy of 71%, which is comparable to that typically obtained by cytotechnologists and cytopathologists. However, it is difficult to understand the image features based only on the judgment results, because these CNN classifications are still considered "black boxes." Regardless, to calculate the activation maps for benign and malignant images and determine the areas focused on by the DCNN for classification, they employed Grad-CAM to produce visual explanations of the decisions made by CNN-based models [27].

9. Classification and rule extraction for cytopathology images with semantic features

When higher degree semantic features of cytopathology images are obtained, benign and malignant tumors can be discriminated [61]. Dey et al. [62] selected both qualitative subjective cytological features and objective quantitative morphometric data to diagnose lobular carcinoma in fine-needle aspiration (FNA) cytology of the breast. Subbaiah et al. [63] extracted the cytological features and morphometric data from FNA cytology of ductal carcinomas of the breast and fibroadenomas, and constructed a shallow NN to discriminate between benign and malignant cases, while Savala et al. [64] built a shallow NN to distinguish between follicular adenoma and follicular carcinoma of the thyroid. In those studies, the classification of FNA cytology features was conducted using BP [4].

Accurate and concise rules are extracted to explain how subjective cytological features identified by pathologists and objective quantitative features are classified into benign and malignant cases. Since the degree of feature abstraction is relatively high, there is no need to use DL; rather, highly accurate and concise rules can be extracted using Re-RX with J48graft (C4.5A) [42].

10. New insights to extract rules for radiological imaging to enhance the interpretability of CNN-based methods

We believe that the interpretable proof-of-concept DL system for clinical radiology developed by Wang et al. [65] enables the automatic scoring of radiological features, which allows radiologists to understand the elements of decision-making behind classification decisions; this could provide radiologists with guidance for detection and diagnosis. These distinctive features, called *semantic features* [67], can be used to create predictors of six classes of liver tumor samples.

Although a variety of high-level features are learned by the CNN, deep features are extracted from fully connected layers [68]. To utilize *deep features*, interpretable and accurate classification rules are extracted using DBN Re-RX with J48graft. DBN Re-RX with J48graft has been applied the WBCD [69]. The rationale behind this method is based on deep features [68] and the large margin principle [37] for shallow NNs. This method can be applied not only to ratings categories, but also to image data sets consisting of semantic features, as shown in Fig. 3.

The main contribution of this paper is that it provides new insights regarding how the "black box" of CNN training (colored black on the left-hand side of Fig. 3) can be converted to semantic features using Wang et al.'s DL system [65] (colored gray in the middle box in Fig. 3). Furthermore, the semantic "gray box" has been transparentized into a "white box" consisting of a series of interpretable classification rules to realize the qualitative interpretation of radiological images.

The new findings presented in this paper involve the extraction of interpretable rules for radiological images (a "black box"); however, stronger supporting evidence is needed. The data set in Ref. [65] was not publicly available, so the CheXpert data set was used [66]. The CheXpert data set was created from two-dimensional gray-scale images (2D CNN), a large data set that contains 224,316 chest radiographs of 65,240 patients. The proposed method achieved a precision (sensitivity) of 59.44% with a recall of 96.59% for the data set. Wang et al.'s method achieved a precision of 76.5% with a recall of 82.9% for a different data set [65] created from contrast-enhanced T1-weighted MRI (three-dimensional CNN). We believe that Wang et al.'s method extracts declarative features, which allow the authors to conduct the segmentation by hand as preprocessing. On the other hand, we could not create high-quality feature maps, as demonstrated in Wang et al.'s method. Therefore, we believe that the proposed method achieves much better precision and recall when the same data set (Wang et al.'s method) is given or publicly available.

Moreover, DBN Re-RX with J48graft can be extended to CNN Re-RX for high-level abstraction data sets (deep features) [25], irrespective of the input and output layers in *any type of* DL for high-level abstraction images with prior knowledge (semantic features) [25]. However, the provision of radiological images is often insufficient in a large number of abstraction data sets 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.

Previous work has confirmed that deeper NNs deliver better visual recognition performance than do shallower NNs when training data sets are kept constant [2]. However, in the study by Lee et al. [8], simply picking the deepest NN [67] was not sufficient. Their results also suggest that the highly accurate classification achieved by DL and the high

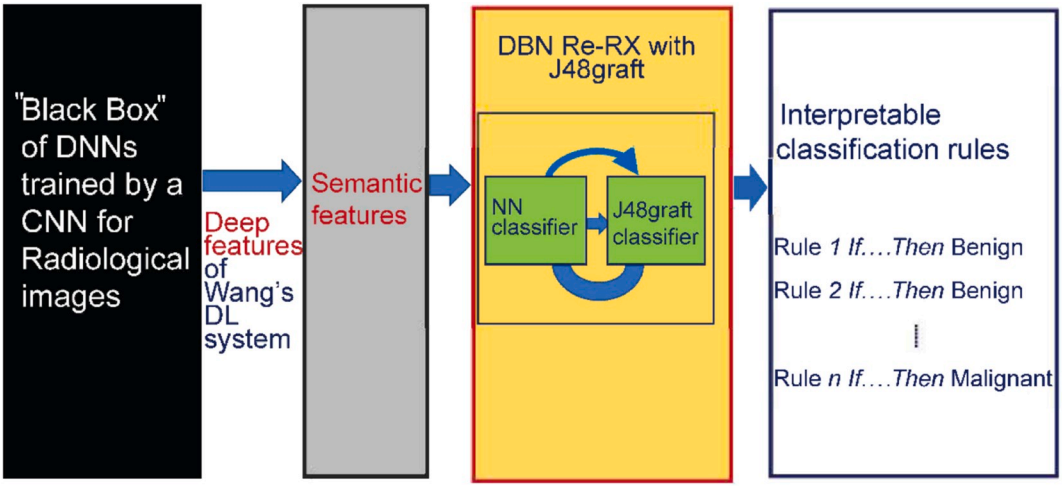

Fig. 3. Overview of interpretable classification rules for the "Black box" of DNNs trained by a CNN for radiological images using DBN Re-RX with J48graft, with semantic features obtained by Wang et al.'s DL system.

interpretability achieved by rule extraction should be used for different purposes in higher-level abstraction images in radiology [57].

Wang et al.'s [65] ideas can be adapted to wider interfaces with standardized reporting systems such as the Breast Imaging Reporting and Data System [70] and the Liver Imaging Reporting and Data System [71].

11. New insights into the qualitative interpretation of pathological images

In feature-based ML, image processing converts a digital image into a tissue-type map, and then segments individual glands and nuclei for feature extraction. Explicit predefined features, such as the ratio between segmented lumens and nuclei, are extracted, in addition to a multitude of cell morphological features. ML using DTs is trained to classify the image (e.g., cancer or not cancer) on basis of these explicit features. By contrast, DNNs are trained to extract and learn engineering features based on raw images in order to improve classification compared with CNNs [24].

As described in the Section 11.2, Campanella et al. [57] developed a novel framework that leverages MIL to train DNNs. However, DTs algorithmically converted from DNNs are very large and complex; therefore, these DTs are not suitable for rule extraction and the interpretation purposes of DNNs [24]. On the other hand, the characteristics of DTs based on explicit features can be well-balanced in terms of both accuracy and interpretability using rule extraction technologies proposed by the present author [15,17–19,38,61].

The rule extraction technologies proposed by the present author can easily explain medical diagnostic results in the form of interpretable and concise rules [20–22]. Qualitative DL [23,38,72] provides transparent diagnosis based on a common understanding of pathological abnormalities among pathologists, urologists, and physicians.

Since high-resolution pathological images such as WSIs contain a large number of pixels, deep features are generally very difficult to interpret. However, for the classification of endometrial lesions, classification rules can be extracted to discriminate between benign and malignant cases based on the following nine categories: proliferative and secretory phases, atrophy, hyperplasias without atypia, adenocarcinoma, endometrioid Ca, endometrioid Ca-grade I, serous Ca, and hyperplasia with atypia [73].

In the evaluation of thyroid cytomorphological characteristics, classification rules can also be extracted to classify benign, follicular neoplasms (undetermined behavior) and malignant cases based on seven and five cytomorphological characteristics, respectively [74]. We can show discrimination rules using based predictors for aggressive grade (Gleason $\geq$7 vs. $\leq$6) prostate cancers and discrimination rules using based predictors for indolent grade (Gleason 6 vs. benign) [75].

In this line, furthermore, hybrid ideas of cell-level, slide-level [60], and morphological, spatial, and text features [76] will be promising to create higher-degree semantic features. That is, if the semantic features described in Section 11 can be obtained, each attribute will have a much higher degree of abstraction. It may also be possible to classify the semantic features of other types of pathological images.

As previously described by Teramoto et al. [59], CNN classifications remain a "black box," and thus, it is difficult to understand the pathological image features based on only the judgment result. Nevertheless, in that study, Grad-CAM was used to produce visual explanations of the decisions made to calculate/render the activation maps for the determination of benign and malignant images. These activation maps are indirect interpretation methods [46] and remain a blanket solution for interpretation.

As recently reported [7,9,52,53,57,59], better interpretations of pathological images are independent from the number of pixels in high-resolution pathological images. We therefore believe that the key point to overcoming barriers is to increase low-level abstraction of the deep features obtained by CNN-based methods to higher-level abstraction semantic features with prior knowledge.

Accordingly, the "obviousness score," which was described in a previous study [53], was used with the aims of providing deeper insights into the subjective perception of tasks with and without assistance and building on the objective measurements of accuracy and efficiency. Despite considerable inter-reader variability, the scores suggested that the perceived ease of image review increased with algorithm assistance, specifically for micrometastases.

The widespread use of these approaches for pathological images and rule extraction technologies has considerable potential to enhance health care delivery by enabling the use of qualitative AI pathways, but only after the system is optimized to achieve appropriate and acceptable results for pathologists and physicians.

12. Concluding remarks

As CNN-based methods have been more widely adopted in histopathology than in cytopathology, the usefulness of CNN-based methods depends not only on quantitative performance, such as accuracy, but also on the ability to integrate the existing diagnostic workflow. By specifically focusing on radiological and pathological images, this paper aims to open the door to the qualitative interpretation of images using DL and deliver the shallow NN-based and DL-based qualitative interpretation of radiological and pathological images.

The qualitative interpretation of pathological images has been described only in limited situations in the literature. However, we believe that cell-level, slide-level, and morphological features will be hybridized in the near future to increase the level of abstraction of semantic features and realize the qualitative interpretation of pathological images using CNN-based models.

As described in Sections 10 and 11, insights for the qualitative interpretation of radiological and pathological images follow the same principle and can be unified from the viewpoint of rule extraction technology.

Although most AI research in pathology still focuses on cancer detection and tumor grading, pathological diagnosis is not simply a morphological diagnosis. Therefore, a complex process consisting of various types of clinical data that deal with various organs and diseases should be evaluated.

Pathologists, urologists, and physicians should strive to understand this burgeoning science, and to acknowledge that ML requires collaboration with rule extraction professionals and computer scientists to optimize the data shown to CNNs, develop highly qualitative AI-based decision support applications, and improve patient care.

Declarations of interests

None.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Funding

This work was supported in part by the Japan Society for the Promotion of Science through a Grant-in-Aid for Scientific Research (C) (18K11481).

Data availability statement

The WBCD Data Set is available from http://archive.ics.uci.edu/ml/

datasets/Breast+Cancer+Wisconsin+%28Original%29.

Role of the funding source

The funding source had no involvement in study design, data curation, investigation, resource, writing, or decision to submit the article for publication.

Declaration of competing interest

The author has no conflicts of interests to declare.

Acknowledgements

Not applicable.

References

[1] Chang H, Jung CK, Woo JI, Lee S, Cho J, Kim SW, et al. Artificial intelligence in pathology. J Pathl Trans Med 2019;53:1–12.
[2] LeCun Y, Boser B, Denker JS, et al. Handwritten digit recognition with a back-propagation network. In: Touretzky DS, editor. Advances in Neural Information Processing Systems, vol. 2. Cambridge, MA: MIT Press; 1989. p. 396–404.
[3] Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. Science 2006;313:504–7.
[4] Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature 1986;323(6088):533–6.
[5] LeCun Y, Bengio Y, Hinton G. Deep learning, Nat 2015;521:436–44.
[6] Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med 2019;25:44–56.
[7] Zhang Z, Chen P, McGough M, et al. Pathologist-level interpretable whole-slide cancer diagnosis with deep learning. Nat Mach Intell 2019;1:236–45.
[8] 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.
[9] Goldenberg SL, Nir G, Salcudean SE. A new era: artificial intelligence and machine learning in prostate cancer. Nat Rev Urol 2019;16:391–403.
[10] Srivastava A, Kulkarni C, Huang K, Parwani A, Mallick P, Machiraju R. Imitating pathologist based assessment with interpretable and context based neural network modeling of histology images. Biomed Inf Insights 2018;10:1–7.
[11] Landau MS, Pantanowitz L. Artificial intelligence in cytopathology: a review of the literature and overview of commercial landscape. J Am Soc Cytopathol 2019;8:230–41.
[12] Hayashi Y. A neural expert system with automated extraction of fuzzy if–then rules and its application to medical diagnosis. In: Lippmann RP, Moody JE, Touretzky DS, editors. Advances in Neural Information Processing Systems, vol. 3. Los Altos, CA: Morgan Kaufmann; 1991. p. 578–84.
[13] Andrews R, Diederich J, Tickele A. Survey and critiques of techniques for extracting rules from trained artificial neural networks. Knowl Base Syst 1995;8:373–89.
[14] Setiono R, Baesens B, Mues C. Recursive neural network rule extraction for data with mixed attributes. IEEE Trans Neural Network 2008;19:299–307.
[15] Hayashi Y, Yukita S. Rule extraction using recursive-rule extraction algorithm with J48graft with sampling selection techniques for the diagnosis of type 2 diabetes mellitus in the Pima Indian Dataset. Informat Med Unlocked 2016;2:92–104.
[16] Fortuny EJD, Martens D. Active learning-based pedagogical rule extraction. IEEE Trans Neural Netw Learn Syst 2015;26:2664–77.
[17] Hayashi Y, Oisi T. High accuracy-priority rule extraction for reconciling accuracy an interpretability in credit scoring. New Generat Comput 2018;36(4):393–418. https://doi.org/10.1007/s00354-018-0043-5.
[18] Hayashi Y. Application of rule extraction algorithm family based on the Re-RX algorithm to financial credit risk assessment from pareto optimal perspective. Operat Res Perspect 2016;3:32–42.
[19] Hayashi Y. Synergy effects between the grafting and the subdivision in the Re-RX with J48graft for the diagnosis of thyroid disease. Knowl Base Syst 2017;131:70–182.
[20] Gallant SI. Connectionist expert systems. Commun ACM 1988;31:152–69.
[21] Uehara D, Hayashi Y, Seki Y, Kakizaki S, Horiguchi N, Hashizume H, et al. The non-invasive prediction steatohepatitis in Japanese patients with morbid obesity by artificial intelligence using rule extraction technology. World J Hepatol 2018;10(12):934–43. https://doi.org/10.4254/wjh.v10.i12.934.
[22] Hayashi Y, Nakajima K, Nakajima K. A rule extraction approach to explore the upper limit of hemoglobin during anemia treatment in patients with predialysis chronic kidney disease. Informat Med Unlocked 2019;17:100262.
[23] Hayashi Y. Detection of lower albuminuria levels and early development of diabetic kidney disease using an artificial intelligence-based rule extraction approach. Diagnostics 2019;9:133. https://doi.org/10.3390/diagnostics9040133.
[24] Akatsuka J, Yamamoto Y, Sekine T, Numata Y, Morikawa H, Tsutsumi K, Yanagi M, Endo Y, Takeda H, Hayashi T, Ueki M, Tamiya G, Maeda I, Fukumoto M, Shimizu A, Tsuzuki T, Kimura G, Kondo Y. Illuminating clues of cancer buried in prostate MR image: deep learning and expert approaches. Biomolecules 2019;9:673. https://doi.org/10.3390/biom9110673.
[25] 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.
[26] Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. In: Fleet D, Pajdla T, Schiele B, et al, eds. Proceedings of the European Conference on Computer Vision, Zurich, Switzerland. September 6–12, 2014.
[27] Selvaraju RR, Cogswell M, Das A, et al. Grad-CAM: visual explanations from deep networks via gradient-based localization. ICCV 2017:618–26. https://doi.org/10.1109/iccv.2017.74. https://search.crossref.org/?q=Selvaraju+RR%2C+Cogswell+M%2C+Das+A%2C+et+al.+Grad-CAM%3A+visual+explanations+from+deep+networks+via+gradient-based+localization.+ICCV.+2017.+p.+618-626.
[28] Zhang D, Meng D, Han J. Co-saliency detection via a self-paced multiple-instance learning framework. IEEE Trans Pattern Anal Mach Intell 2017;39(5):865–78.
[29] Wang C, Dong S, Zhao X, Papanastasiou G, Zhang H, Yang G. SaliencyGAN: deep learning semisupervised salient object detection in the fog of IoT. IEEE Trans Indust Informat 2020;16(4):2667–76.
[30] Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 2019;1:206–15.
[31] LeCun Y, Cortes C, Burges C. The mnist database of handwritten digits. 2012, Available electronically at, http://yann.lecun.com/exdb/mnist; 1998.
[32] Tran SN, Garcez d'Avila AS. Deep logic networks: inserting and extracting knowledge from deep belief networks. IEEE Trans Neural Netw Learn Syst 2016;29:246–58. https://doi.org/10.1109/TNNLS.2016.2603784.
[33] Bologna G, Hayashi Y. A rule extraction study on neural network trained by deep learning. In: Proceedings of the International Joint Conference on Neural Networks. Vancouver): IJCNN 2016); 2016. p. 668–75.
[34] Bologna G, Hayashi Y. Characterization of symbolic rules embedded in deep DIMLP networks: a challenge to transparency of deep learning. J Artif Intell Soft Comput Res 2017;7:265–86. https://doi.org/10.1515/jaiscr-2017-0019.
[35] Erhan D, Bengio Y, Courville A, Manzagol PA, Vincent A. Why does unsupervised pre-training help deep learning? J Mach Learn Res 2010;11:625–60.
[36] Abdel-Zaher AM, Eldeib AM. Breast cancer classification using deep belief networks. Expert Syst Appl 2016;46:139–44.
[37] Vapnik VN. The Nature of Statistical Learning Theory. NY: Springer-Verlag New York, Inc.; 1995.
[38] Hayashi Y. Use of a deep belief network for small high-level abstraction data sets using artificial intelligence with rule extraction. Neural Comput 2018;30(12):3309–33.
[39] Luo C, Wu D, Wu D. A deep learning approach for credit scoring using credit default swaps. Eng Appl Artif Intell 2017;65:406–20.
[40] Liu K, Kang G, Zhang N, Hou B. Breast cancer classification base on fully-connected layer first convolutional neural networks. IEEE Access 2018;6:23722–32. https://doi.org/10.1109/ACCESS.2018.2817593.
[41] Quinlan JR. C4.5: Programs for Machine Learning. San Mateo, CA: Morgan Kaufmann; 1993.
[42] 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. p. 702–7.
[43] 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, vol. 8. Cambridge, MA: MIT Press; 1996. p. 24–30.
[44] Liu Y, Kohlberger T, Norouzi M, Dahl GE, Smith JL, Mohtashamian A, et al. Artificial intelligence-based breast cancer nodal metastasis detection insights: into the black box for pathologists. Arch Pathol Lab Med 2019;143:859–68. https://doi.org/10.5858/arpa.2018-0147-OA.
[45] Gecer B, Aksy S, Mercan E, Shapiro LG, Weaver DL, Elmore JG. Detection and classification of cancer in whole slide breast histopathology images using deep convolutional networks. Pattern Recogn 2018;84:345–56.
[46] Chang P, Grinband J, Weinberg BD, Bardis M, Khy M, et al. Deep-learning convolutional neural networks accurately classify genetic mutations in gliomas. Am J Neuroradiol 2018;39:1201–7.
[47] Hajdu SI. Microscopic contributions of pioneer pathologists. Ann Clin Lab Sci 2011;41:201–6.
[48] Epstein JI, Zelefsky M, Sjoberg DD, Nelson JB, Egevad L, Magi-Galluzzi C, et al. A contemporary prostate cancer grading system: a validated alternative to the Gleason score. Eur Urol 2016;69:428–35.
[49] Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017;542:115–8.
[50] Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. J Am Med Assoc 2016;316:2402–10.
[51] Kermany DS, Goldbaum M, Cai W, Valentim CCS, Liang H, Baxter SL, et al. Identifying medical diagnoses and treatable diseases by image- based deep learning. Cell 2018;172:1122–31. e9.
[52] Bejnordi BE, Veta M, van Diest PJ, van Ginneken B, Karssemeijer N, Litjens G, van der Laar JAWM. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. J Am Med Assoc 2017;318:2199–210.
[53] Steiner DF, MacDonald R, Liu Y, Truszkowski P, Hipp JD, Gammage C, et al. Impact of deep learning assistance on the histopathologic review of lymph nodes for metastatic breast cancer. Am J Surg Pathol 2018;42(2):1636–46. https://doi.org/10.1097/PAS.0000000000001151.

[54] Jiménez G, Racoceanu D. Deep learning for semantic segmentation vs. classification in computational pathology: application to mitosis analysis in breast cancer grading. Front Bioeng Biotechnol 2019;7:145.
[55] García G, Colomer A, Naranjo V. First-stage prostate cancer identification on histopathological images: hand-driven versus automatic learning. Entropy 2019; 21:356. https://doi.org/10.3390/e21040356.
[56] Schelb P, Kohl S, Radtke JP, Wiesenfarth M, Kickingereder P, Bickelhaupt S, et al. Classification of cancer at prostate MRI: deep learning versus clinical PI-RADS assessment. Radiology 2019;293. 607-617.
[57] Campanella G, Hanna MG, Geneslaw L, Miraflor A, Silva VWK, Busam KJ, et al. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat Med 2019;25:1301–9.
[58] Hashimoto Y, Ohno I, Imaoka H, Takahashi H, Mitsunaga S, Sasaki M, et al. Preliminary results of computer aided diagnosis (CAD) performances using deep learning in EUS-FNA cytology of pancreatic cancer. Gastrointest Endosc 2018;87 (6). AB434–AB434.
[59] Teramoto A, Yamada A, Kiriyama Y, Tsukamoto T, Yan K, Zhang L, Imaizumi K, Saito K, Fujita H. Automated classification of benign and malignant cells from lung cytological images using deep convolutional neural network. Informat Med Unlocked 2019;16:100205.
[60] Sanghvi AB, Allen EZ, Callenberg KM, Pantanowitz L. Performance of an artificial intelligence algorithm for reporting urine cytopathology. Canc Cytopathol 2019; 127:658–66.
[61] 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. Informat Med Unlocked 2015;1:9–16.
[62] Dey P, Logasundaram R, Joshi K. Artificial neural network in diagnosis of lobular carcinoma of breast in fine-needle aspiration cytology. Diagn Cytopathol 2011;41 (2):102–6.
[63] Subbaiah RM, Dey P, Nijhawan R. Artificial neural network in breast lesions from fine-needle aspiration cytology smear. Diagn Cytopathol 2013;42(3):218–24.
[64] Savala R, Dey P, Gupta N. Artificial neural network model to distinguish follicular adenoma from follicular carcinoma on fine needle aspiration of thyroid. Diagn Cytopathol 2018;46:244–9.
[65] 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.
[66] Irvin J, et al. CheXpert t: a large chest radiograph dataset with uncertainty labels and expert comparison. In: Proceedings of 32rd AAAI conference on Artificial Intelligence (AAAI-19). vol. 1. Honolulu: Hawaii—Jan; 2019. p. 591–7. 27-Feb.
[67] Paul R, Schabath M, Balagurunathan Y, et al. Explaining deep features using radiologist defined semantic features and traditional quantitative features. Tomography 2019;5(1):192–200.
[68] Giryes R, Sapiro G, Bronstein AM. Deep neural networks with random Gaussian weights: a universal classification strategy? IEEE Trans Signal Process 2015;64(13): 3444–57.
[69] Setiono R. Extracting rules from pruned neural networks for breast cancer diagnosis. Artif Intell Med 1996;8:37–51.
[70] Obenauer S, Hermann K, Grabbe E. Applications and literature review of the BI-RADS classification. Eur Radiol 2005;15:1027–36.
[71] Mitchell DG, Bruix J, Sherman M, Sirlin CB. LI-RADS (liver imaging reporting and data system): summary, discussion, and consensus of the LI-RADS management working group and future directions. Hepatology 2015;61:1056–65.
[72] Hayashi Y. Toward the transparency of deep learning in radiological imaging: beyond quantitative to qualitative artificial intelligence. J Med Artif Intell 2019;2: 19. https://doi.org/10.21037/jmai.2019.09.06.
[73] Pouliakis A, Margari C, Margari N, Chrelias C, Zygouris D, Meristoudis C, Panayiotides I, Karakitsos P. Using classification and regression trees, liquid-based cytology and nuclear morphometry for the discrimination of endometrial lesions. Diagn Cytopathol 2013;42(7):582–91.
[74] Margari N, Mastorakis E, Pouliakis A, Gouloumi A-R, Eleftherios Asimis E, Konstantoudakis S, Ieromonachou P, Panayiotides IG. Classification and regression trees for the evaluation of thyroid cytomorphological characteristics: a study based on liquid based cytology specimens from thyroid fine needle aspirations. Diagn Cytopathol 2018;46:670–81.
[75] Li Q, Lu H, Choi J, Gage K, Feuerlein S, Pow-Sang JM, Gillies R, Balagurunathan Y. Radiological semantics discriminate clinically significant grade prostate cancer. Canc Imag 2019;19:81.
[76] Yu C, Chen H, Li Y, Peng Y, Li J, Yang F. Breast cancer classification images based on hybrid features. Multimed Tool Appl 2019;78:21325–45.