AJR FOCUS ON:

Artificial Intelligence for Medical Image Analysis: A Guide for Authors and Reviewers

Joseph R. England[1]
Phillip M. Cheng[2]

OBJECTIVE. The purpose of this article is to highlight best practices for writing and reviewing articles on artificial intelligence for medical image analysis.

CONCLUSION. Artificial intelligence is in the early phases of application to medical imaging, and patient safety demands a commitment to sound methods and avoidance of rhetorical and overly optimistic claims. Adherence to best practices should elevate the quality of articles submitted to and published by clinical journals.

Remarkable strides have been made in artificial intelligence (AI), a subfield of computer science devoted to creating systems to perform tasks ordinarily requiring human intelligence. The advent of large datasets has spurred advances in machine learning systems that can learn from patterns in data rather than from explicit rules. Breakthrough performance gains in machine learning for computer vision have led to reports of systems with expert or near-expert performance in medical imaging tasks, such as identification of pulmonary tuberculosis, detection of hip fractures, and estimation of pediatric bone age [1–10]. The number and rate of AI manuscript submissions to clinical journals can only be expected to increase, and reviewers need to know how to evaluate these manuscripts (Appendix 1). The purpose of this article is to present best practices for writing and reviewing articles on AI for medical image analysis. The emphasis is on content and methods best suited for submission to a clinical journal.

Content and Venue

A manuscript on the application of AI to medical imaging could be submitted to a computer science, biomedical engineering, or clinical journal. This choice should be guided by the topic and content of the study. Articles that show fundamental technical and methodologic advances in AI, generalizable beyond medical imaging applications, are well suited to a computer science journal. Articles analyzing technical issues and innovations specific to medical datasets may be best suited to a biomedical engineering journal. The audience of a clinical journal is likely to be more interested in articles showing new or improved applications of AI to clinical problems and is likely to have the necessary medical expertise to determine whether the proposed solutions are convincing. Accordingly, articles submitted to clinical journals should discuss research aimed at solving practical clinical issues and should present the research in a manner that is accessible to physicians and biomedical researchers.

Purpose

The purpose of an article on AI in biomedical imaging sets the expectations of the reader and the level of evidence needed to accept the conclusions of a study. Possible aims include proof of technical feasibility, expert-level performance, or real-world clinical performance (Tables 1 and 2).

Technical Feasibility

A study that aims to evaluate technical feasibility should present a system with promising performance despite suboptimal training data or computing power. Compelling articles in this vein could include the first successful attempt to apply a type of algorithm to a specific medical imaging task. Because collection of high-quality data is resource intensive, feasibility studies often have small or limited datasets and are used to explore whether committing more resources to data collection and algorithm refinement is worthwhile.

Keywords: artificial intelligence, deep learning, machine learning, technology assessment

doi.org/10.2214/AJR.18.20490

Received August 2, 2018; accepted after revision September 4, 2018.

[1]Department of Radiological Sciences, David Geffen School of Medicine at UCLA, Los Angeles, CA.

[2]Department of Radiology, Keck School of Medicine of USC, 1441 Eastlake Ave, Ste 2315B, Los Angeles, CA 90033. Address correspondence to P. M. Cheng (Phillip.Cheng@med.usc.edu).

AJR 2019; 212:1–7

0361–803X/19/2123–1

TABLE 1: Summary of Possible Study Design Options Listed by Study Aim

Component	Technical Feasibility	Expert-Level Performance	Real-World Clinical Performance
Learning approach	Supervised learning or unsupervised learning	Supervised learning	Supervised learning
Data collection			
Cohort	Retrospective or prospective	Retrospective or prospective	Retrospective or prospective[a]
Sampling	Convenience or consecutive	Convenience or consecutive[a]	Consecutive
Subsets	Training and test or training, validation, and test	Training and test or training, validation, and test[b]	Training and test; training, validation, and test[b]; or test only[b]
Data labels	Radiology reports or expert consensus or reference standard	Expert consensus or reference standard	Expert consensus or reference standard

[a]Preferred.
[b]Most common.

Human- or Expert-Level Performance

A study that aims to evaluate human- or expert-level performance must compare the performance of a system and the performance of experts based on either a benchmark set of images (a test set) or previously reported metrics of expert performance. Performance comparison on a large and robust dataset is typically required to justify a claim of expert-level performance.

Clinical Performance

Proof of real-world clinical performance involves testing a system on a large dataset that accurately and fully reflects the expected variability of images in the intended usage setting. With few exceptions, such a test set should be multiinstitutional and consist of images collected consecutively according to explicit eligibility criteria and including all relevant variations in patient demographics and disease state [11]. The highest-quality study would be a randomized controlled trial in which a large number of patients undergoing imaging for a specific clinical indication are randomized to have the images interpreted by a machine learning algorithm or a radiologist.

Image Analysis Task

Machine learning can be used to perform several types of image analysis tasks, including classification, regression, localization, and segmentation.

Classification

Classification algorithms assign images to categories. The simplest example is a binary classification algorithm with exactly two categories, such as disease being present or disease being absent. Multiclass or multinomial classification algorithms categorize an image into one of three or more categories.

Regression

Regression algorithms assign a number to an image [12]. An example of a regression task in medical imaging is estimation of pediatric bone age from hand radiographs [10].

Localization

Localization algorithms return the location of a particular object in an image, either as the location of a center pixel or voxel or as a bounding box around the object [13].

Segmentation

Segmentation algorithms classify each pixel or voxel in an image as part of or not part of a particular object [14–16]. An example of a segmentation task in medical imaging is measurement of left ventricular volume on cardiac MR images [17].

Learning Approach

There are two general approaches to a machine learning problem: supervised learning and unsupervised learning.

Supervised Learning

In supervised learning, a machine learning algorithm is presented with data (images in the case of image analysis problems) and ground truth labels [18, 19]. The labels are the correct answers that an algorithm is intended to learn when confronted with certain data. For example, a labeled dataset of CT brain images may include the label "acute hemorrhage" or "no acute hemorrhage." Training proceeds in iterations of three steps: first, the algorithm outputs answers for each item in the training dataset; second, the answers and ground truth labels are used to compute a cost measurement of how wrong the algorithm is; and third, the cost is used as feedback to try to improve the algorithm so that the answers in the next iteration will be better. The learning is said to be supervised because ground truth labels are available to correct the algorithm at each iteration.

Unsupervised Learning

In unsupervised learning, a machine learning algorithm is presented with unlabeled data and learns to group the data by similarities and differences [18]. For example, an algorithm presented with a set of brain CT and brain MR images may learn to assign these images into groups based solely on patterns in the pixel data without reference to any ground truth labels. Although both learning approaches are possible, supervised learning is much more commonly used for image analysis problems and is emphasized in this article.

Data Collection and Processing

Machine learning for image analysis typically requires a large quantity of image data. Researchers can use public datasets or collect new data. Because public datasets may not adequately represent a target patient population, new data may be preferable for proving expert-level or clinical performance. When new data are used, articles should describe in detail how the images were collected, processed, and divided into subsets.

Collection

A description of data collection should include whether the collection was retrospective or prospective, the time period and geographic locations of sampling, whether sampling was consecutive or convenience, whether explicit inclusion or exclusion criteria were used, the file types used for saving image data, and, if DICOM files were not used, the window settings used in the conversion from DICOM to other file types [20].

TABLE 2: Elements of an Artificial Intelligence Research Article With Example Language From Published Articles

Element	Language
Purpose	
Technical feasibility	"The purpose of this pilot study is to determine whether a deep convolutional neural network can be trained with limited image data to detect high-grade small-bowel obstruction patterns on supine abdominal radiographs." [44]
Expert-level performance	"Here we.. present the first large scale study where a deep learning system achieves human-level performance on a common and important radiological task." [9]
Image analysis task	
Classification	"The pneumonia detection task is a binary classification problem, where the input is a frontal frontal-view chest X-ray..and the output is a binary label..indicating the absence or presence of pneumonia, respectively." [45]
Data collection and processing	
Collection	"A total of 4037 consecutive clinical gray-scale abdominal radiographs from 3270 examinations performed on 1346 distinct patients (764 male and 582 female) from January to June 2016 were retrospectively obtained." [44]
Processing	"Overall our preprocessing method includes binary image segmentation as a first step and then the analysis of connected components for the postprocessing of segmentation results." [10]
Data labels	
Radiology reports	"Each study [in the training set] was manually labeled as normal or abnormal by board-certified radiologists from the Stanford Hospital at the time of clinical radiographic interpretation." [46]
Expert consensus	"We collected a test set of 420 frontal chest X-rays. Annotations were obtained independently from four practicing radiologists at Stanford University.. The radiologists had 4, 7, 25, and 28 years of experience, and one of the radiologists is a sub-specialty fellowship trained thoracic radiologist." [45]
Model training	"CheXNet is a 121-layer Dense Convolutional Network (DenseNet). The weights of the network are randomly initialized and trained end-to-end using Adam with standard parameters [and] an initial learning rate of 0.01.." [45]
Evaluation of performance	"On the test datasets, ROC curves and AUCs were determined. Contingency tables, accuracy, sensitivity, and specificity were determined from the optimal threshold by the Youden index." [8]
Visualization	"Sample images from the test set with corresponding superimposed saliency maps [show that] the most sensitive regions [for prediction of pediatric bone age] corresponded to the proximal interphalangeal joints, the metacarpal-phalangeal joints, and the carpal bones." [47]

Processing

Image processing before model training can be as simple as windowing or cropping or as complicated as using a separately trained segmentation algorithm to segment out parts of the images [10]. Authors should include sufficient detail to ensure that their work can be replicated.

Division Into Subsets

The simplest division of data is a training set for training the algorithm and a separate test set used only for final performance testing. If there are sufficient data, an intermediate validation set is often used to fine-tune the training process. Most of the data are usually used for training, and the rest is used for validation and testing (e.g., a 70%, 15%, 15% split of training, validation, and test sets). Authors should detail which data subsets were used and how data were assigned to each subset.

Data Labels

Obtaining high-quality data labels is one of the most difficult aspects of developing any high-performance machine learning model. In medical imaging, this problem is compounded by the limited availability of qualified experts to provide accurate image labels. For this reason, data labeling efforts are often concentrated on the test set. Training high-performance models is usually still possible with occasional mislabeled images in the training set, but even a few mislabeled images in the test set can lead to drastically different conclusions. For example, in a test set of 100 images, if a single image is mislabeled, accuracy will be misrepresented by 1%. Possible methods of obtaining labels include radiology reports, expert consensus, reference standard imaging or laboratory examinations, and surgical or pathologic confirmation.

Radiology Reports

The use of labels from radiology reports alone may be sufficient for the training dataset but should be discouraged in the test dataset for anything but feasibility studies. Radiology reports generally represent the interpretation of a single radiologist. The radiologist could have erred or been biased by additional clinical information or follow-up imaging when giving the interpretation.

Expert Consensus

Consensus by three or more experts who independently relabel the images partially solves labeling errors and biases and is a common method of labeling medical image test sets. Advantages of this method include a controlled context for interpretation and the feasibility of calculating a lower limit of label accuracy based on an assumption that agreement implies accuracy [21]. The disadvantage is that expert labor is typically a scarce resource and may be impractical for application to large training sets. Some research groups have addressed this problem by developing methods for flagging a small percentage of the training set that is likely to be mislabeled [9]. Authors who use experts for image labeling or relabeling should report the level of expertise for each expert reviewer (e.g., subspecialty training, years of clinical experience), which images were interpreted (e.g., test set only, test set and validation set, etc.), and measures of observer agreement among the experts.

Reference Standards

Reference standards such as follow-up imaging, surgical confirmation, and laboratory

American Journal of Roentgenology

or pathologic diagnosis are ideal but can be difficult to obtain and standardize for large datasets. However, future work proving real-world clinical performance in image analysis tasks will likely require this level of label accuracy.

Model Training

Training machine learning algorithms from large volumes of image data requires computing power and optimization of hyperparameters [22]. This section of an article can easily become heavy with technical jargon. For biomedical audiences, any technical aspects of model training that might confuse or detract from readability should be summarized or deferred to a supplemental section for interested readers.

Hardware and Software

Authors should provide a detailed description of the computer hardware and software used in the training process. Specific items to mention include the amount and type of random access memory, processor type, graphics card type if applicable, and machine learning libraries or software.

Hyperparameters

A machine learning algorithm specifies mathematic operations that will be performed on the input data to arrive at an output. These operations involve numeric parameters. An algorithm can contain anywhere from a few to hundreds of millions of parameters, and training the algorithm is an automated iterative process by which the parameters are gradually altered to improve output accuracy. The training process itself may have many options, called hyperparameters to distinguish from parameters, that specify how training proceeds but are not operative in the final trained model. Examples of hyperparameters include learning rate (i.e., the degree to which the parameters change during each iteration), regularization variables (used to prevent overfitting), and the number of total iterations. Hyperparameters can be optimized manually or systematically by means of grid search or random search [23, 24]. Authors should include the basic methods used to select and optimize hyperparameters, but detailed explanation, if necessary, would be best suited to a supplemental section.

Evaluation of Performance

Performance testing is the most important step in assessing technical feasibility, expert-level performance, and the real-world clinical performance of an AI algorithm. To ensure that reported performance is not overly optimistic, testing should be performed with a test dataset that was separate from the training dataset during the training process. The best measures of performance differ by image analysis task. In this article we focus on classification, the most frequent task in published medical imaging AI work thus far.

Classification

The simplest measure of how well a classification algorithm performs is accuracy: the fraction of examples in the test set that it predicts correctly. But accuracy alone can mislead because it depends on prevalence. For example, a binary classifier trained to diagnose a rare condition present on only 0.1% of images would achieve 99.9% accuracy if it always predicted disease not present. Physicians use a variety of statistics to understand diagnostic tests, including the binary contingency table containing true-positive, true-negative, false-positive, and false-negative rates and derivatives of these measures, such as sensitivity, specificity, positive predictive value, negative predictive value, and likelihood ratio. These statistics are partially redundant and involve tradeoffs that make system optimization and comparison difficult. For direct comparison, a single summary measure may be desirable. Several commonly reported summary measures are the $F1$ score, Youden J index, and ROC AUC.

The $F1$ score is the harmonic mean of positive predictive value (also known as precision) and sensitivity (also known as recall) and can range between 1 (perfect classification) and 0 [25]. It is calculated as follows:

$$F1score = 2 \times \frac{PPV \times sensitivity}{PPV + sensitivity}.$$

An advantage of the $F1$ score is that it summarizes the information in the binary contingency table in one number. It assumes, however, that the number of true-negative results is not important and assigns an equal cost to false-negative and false-positive results [26].

The Youden J index [27] can also be used to summarize the performance of a binary classifier. It is calculated as follows:

$$J = sensitivity + specificity - 1.$$

The Youden index assumes that false-negative and false-positive classifications are equally undesirable. For instances in which information on pretest probability and the costs of false results is available, a method exists to weight the Youden index accordingly [28].

The ROC curve is useful when a binary classifier outputs a numeric value to which a threshold can be applied for determining a category. The ROC curve is a plot of sensitivity versus false-positive rate (1 – specificity) for each set of unique binary contingency tables possible with different thresholds for classifier output [28]. The ROC AUC can be used as a single metric to summarize performance across the entire operating range of a classifier, not just at one threshold. There are established methods for comparing ROC curves between different diagnostic tests [29–32]. A disadvantage is that the entire operating range of a classifier or diagnostic test is rarely of interest, and the region where sensitivity and specificity are more balanced is often more useful. For this reason, contingency table statistics such as sensitivity and specificity should still be reported for a chosen classifier threshold. A common method for choosing an optimal threshold is to maximize the Youden index, though this entails the limitations of the Youden index described earlier. For systems that require an explicit threshold, authors should report the threshold used and defend the use of this threshold in the discussion section of the article.

For multiclass classification, statistics can be reported for each separate class. Summary measures can be calculated by averaging the relevant statistic in each class, either as a macroaverage (each class weighted equally) or as a microaverage (each class weighted by prevalence in the test set) [26]. In some multiclass problems certain errors will be more serious than others (e.g., misclassifying a malignancy as a benign condition versus confusing two benign entities), and custom metrics may be needed to weigh errors appropriately.

Regression, Localization, Segmentation

Detailed discussion of metrics for nonclassification machine learning tasks is beyond the scope of this article. Examples of evaluation metrics for regression tasks are mean absolute error, mean squared error, and root-mean-square error [10, 12, 33]. Evaluation metrics for localization and segmentation tasks include intersection-over-union, warping error, Rand error, pixel error, and the Dice similarity coefficient [13, 15, 16, 34–36].

Reporting Measures of Performance

Different evaluation methods have different advantages and disadvantages. As a rule,

American Journal of Roentgenology

authors should provide as many metrics as necessary to describe strengths and weaknesses of a given algorithm, and reviewers should be alert to selective reporting of metrics that might give a biased portrayal of performance. To assist reviewers, authors should cite prior studies using the same or similar performance measures and literature on the advantages and disadvantages of those measures. When possible, evaluation metrics should be reported with confidence intervals. Comparisons between algorithms or between algorithms and humans should include either confidence intervals or measurements of statistical significance.

Visualization

With potentially millions of parameters in a machine learning model, it can be difficult to understand what the model is seeing in an image. This black-box nature creates a challenge for validating AI algorithms. It is important to show that a high-performance machine learning model is actually detecting the relevant region of an image and not overfitting to unimportant findings (see later, Overfitting). This is especially true for classification and regression tasks because the output of the algorithms in these cases is a label (i.e., answer) for what is in the image without any supporting evidence. An in-depth discussion of visualization methods is beyond the scope of this article, but some commonly used examples include occlusion maps, saliency maps, and class activation maps [37–39]. Localization and segmentation algorithms may be inherently more understandable because the output is an image.

Pitfalls and Biases

Authors and reviewers should be aware of several common pitfalls and biases that may arise in AI research.

Overfitting

Overfitting occurs when an algorithm becomes so accurate on a limited dataset that its predictions are not well generalized to new examples. Most often the algorithm is overfitted to the training dataset, but overfitting can also occur on a validation or test set if the investigators test the performance of so many different models that one performs extremely well by chance (see later, P-Hacking Bias). Overfitting can also occur when there are features in an image that are superficially related to a disease state but do not actually represent the disease. For example, one study [40] showed that algorithms trained to detect pneumonia sometimes rely on information in the corners of the image, such as radiopaque markers for left and right.

The importance of confounders is contextual and open to interpretation because radiologists also sometimes use other information in an image to support a conclusion or change their degree of suspicion. For example, an algorithm trained to detect acute fractures that is found to rely partially on the presence of a radiopaque marker of maximal tenderness may be legitimately fitted to the relevant evidence of fracture because radiologists also look at these markers during routine interpretation. Authors and reviewers should attempt to imagine any common confounding factors that may lead to overfitting. Authors can either provide evidence against such overfitting by providing visualizations that include common confounders or can explain in the discussion section why fitting to a specific confounder may be acceptable when considered in context.

Data Snooping Bias

Data snooping bias occurs when the test set directly or indirectly influences the training process [41]. For example, if the training set contains some images identical or nearly identical to images in the test set, the algorithm will appear to achieve high performance on the test set when in fact it has merely memorized some images in the training set. Authors can defend against the possibility of data snooping by double-checking their method of data collection and by considering any context-specific issues that might lead to a false appearance of high performance on the test set.

Spectrum Bias

Spectrum bias occurs when the dataset does not appropriately represent the range of possible patients and disease manifestations for the image analysis task at hand [11]. This issue is most critical to identify in the test dataset because the conclusions of a study are drawn from performance on the test set and not from the training data. Authors should report all relevant clinical and demographic information necessary for thoughtful criticism of their results and conclusions [42].

Straw Man Bias

Straw man bias occurs when authors claim expert-level performance but the standard of comparison does not represent real expertise in the relevant image analysis task. Straw man bias can be difficult to identify because there are legitimate contexts in which the standard of comparison need not be an expert. For example, if an algorithm is proposed for rendering preliminary reports during off-hour times, the use of radiology residents as a standard of comparison would be reasonable. Because this is a contextual issue, authors should qualify the strength of their conclusions and provide convincing support for the standard of comparison they used.

P-Hacking Bias

P-hacking bias occurs when investigators run a sufficient number of statistical tests that one is successful purely by chance. In AI research, a similar bias can occur if investigators train many algorithms using different hyperparameters, test them all against the test set, and selectively publish the best results. To avoid this, only a limited number of models should ever be tested against the test set, and the criteria for selecting which models are tested should be explicit. If investigators choose to fine-tune performance by training many models with different hyperparameters, then the data should be divided into three sets with training performed by use of the training set, optimization of hyperparameters by use of the validation set, and final performance evaluation on the test set.

Preprints and Open-Source Code

Clinical journals often require that references come from peer-reviewed medical literature. However, the computer science research community embraces the use of archived preprints, and many breakthrough articles are published as conference proceedings and preprints [43]. Selective citation of these papers may improve the quality of an article. Authors should also cite any open-source code that was used in their research so that others can reproduce their work if necessary. Submission of research code and datasets for publication or open access is especially welcome when possible.

Summary

AI is in the early phases of application to medical imaging, and patient safety demands a commitment to sound research methods. Clinical journals are well positioned to ensure that AI articles are held to a high standard. This guide is intended to help authors

and reviewers meet this standard and to serve as a basis for editorial boards or imaging societies seeking to create formal research reporting guidelines.

References

1. LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. *Proc IEEE* 1998; 86:2278–2324
2. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. *Adv Neural Inf Process Syst* 2012; 25:1097–1105
3. Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions. *Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit* 2015; 2015:1–9
4. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. *Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit* 2016; 2016:2818–2826
5. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv website. arxiv.org/abs/1409.1556. Last revised April 10, 2015. Accessed October 11, 2018
6. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. *Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit* 2016; 2016:770–778
7. Huang G, Liu Z, Weinberger KQ, van der Maaten L. Densely connected convolutional networks. *Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit* 2017; 2017:1–3
8. Lakhani P, Sundaram B. Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. *Radiology* 2017; 284:574–582
9. Gale W, Oakden-Rayner L, Carneiro G, Bradley AP, Palmer LJ. Detecting hip fractures with radiologist-level performance using deep neural networks. arXiv website. arxiv.org/abs/1711.06504. November 17, 2017. Accessed October 11, 2018
10. Iglovikov V, Rakhlin A, Kalinin A, Shvets A. Pediatric bone age assessment using deep convolutional neural networks. arXiv website. arxiv.org/abs/1712.05053. Last revised June 19, 2018. Accessed October 11, 2018
11. Park SH, Han K. Methodologic guide for evaluating clinical performance and effect of artificial intelligence technology for medical diagnosis and prediction. *Radiology* 2018; 286:800–809
12. Lathuilière S, Mesejo P, Alameda-Pineda X, Horaud R. A comprehensive analysis of deep regression. arXiv website. arxiv.org/abs/1803.08450. March 22, 2018. Accessed October 11, 2018
13. Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: unified, real-time object detection. *Proc IEEE Comput Soc Conf Comput Vis Pattern Recognit* 2016; 2016:779–788
14. Yuheng S, Hao Y. Image segmentation algorithms overview. arXiv website. arxiv.org/abs/1707.02051. Last revised July 7, 2017. Accessed October 11, 2018
15. Garcia-Garcia A, Orts-Escolano S, Oprea S, Villena-Martinez V, Garcia-Rodriguez J. A review on deep learning techniques applied to semantic segmentation. arXiv website. arxiv.org/abs/1704.06857. Last revised April 22, 2017. Accessed October 11, 2018
16. Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells WM, Frangi AF, eds. *Medical image computing and computer-assisted intervention: MICCAI 2015*. Berlin, Germany: Springer, 2015:234–241
17. Lieman-Sifry J, Le M, Lau F, Sall S, Golden D. FastVentricle: cardiac segmentation with ENet. In: Pop M, Wright GA, eds. *International Conference on Functional Imaging and Modeling of the Heart*. Berlin, Germany: Springer, 2017:127–138
18. Erickson BJ, Korfiatis P, Akkus Z, Kline TL. Machine learning for medical imaging. *RadioGraphics* 2017; 37:505–515
19. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature* 2015; 521:436–444
20. Hulley SB, Cummings SR, Browner WS, Grady DG, Newman TB. *Designing clinical research*. Philadelphia, PA: Lippincott Williams & Wilkins, 2011
21. Kundel HL, Polansky M. Measurement of observer agreement. *Radiology* 2003; 228:303–308
22. Chartrand G, Cheng PM, Vorontsov E, et al. Deep learning: a primer for radiologists. *RadioGraphics* 2017; 37:2113–2131
23. Bengio Y. Practical recommendations for gradient-based training of deep architectures. In: Montavon G, Orr GB, Müller K-R, eds. *Neural networks: tricks of the trade*, 2nd ed. Berlin, Germany: Springer, 2012:437–478
24. Bergstra J, Bengio Y. Random search for hyper-parameter optimization. *J Mach Learn Res* 2012; 13:281–305
25. Powers DM. What the F-measure doesn't measure: features, flaws, fallacies and fixes. arXiv website. arxiv.org/abs/1503.06410. March 22, 2015. Accessed October 11, 2018
26. Sokolova M, Lapalme G. A systematic analysis of performance measures for classification tasks. *Inf Process Manage* 2009; 45:427–437
27. Youden WJ. Index for rating diagnostic tests. *Cancer* 1950; 3:32–35
28. Obuchowski NA. Receiver operating characteristic curves and their use in radiology. *Radiology* 2003; 229:3–8
29. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 1988; 44:837–845
30. Hanley JA, McNeil BJ. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology* 1983; 148:839–843
31. Metz CE. Basic principles of ROC analysis. *Semin Nucl Med* 1978; 8:283–298
32. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* 1982; 143:29–36
33. Baccianella S, Esuli A, Sebastiani F. Evaluation measures for ordinal regression. In: *2009 Ninth international conference on intelligent systems design and applications*. Piscataway, NJ: IEEE, 2009:283–287
34. Rahman MA, Wang Y. Optimizing intersection-over-union in deep neural networks for image segmentation. In: Bebis G, Boyle R, Parvin B, et al., eds. *Advances in visual computing*. Berlin, Germany: Springer, 2016:234–244
35. Redmon J, Farhadi A. YOLO9000: better, faster, stronger. arXiv website. arxiv.org/abs/1612.08242. December 25, 2015. Accessed October 11, 2018
36. Zou KH, Warfield SK, Bharatha A, et al. Statistical validation of image segmentation quality based on a spatial overlap index: scientific reports. *Acad Radiol* 2004; 11:178–189
37. Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. In: Fleet D, Pajdla T, Schiele B, Tuytelaars T, eds. *Computer vision: ECCV 2014*. Berlin, Germany: Springer, 2014:818–833
38. Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: visualising image classification models and saliency maps. arXiv website. arxiv.org/abs/1312.6034. Last revised April 19, 2014. Accessed October 11, 2018
39. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. In: *2017 IEEE international conference on computer vision*. Piscataway, NJ: IEEE, 2017:618–626
40. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Confounding variables can degrade generalization performance of radiological deep learning models. arXiv website. arxiv.org/abs/1807.00431. Last revised July 13, 2018. Accessed October 11, 2018
41. Abu-Mostafa YS, Magdon-Ismail M, Lin HT. *Learning from data*. New York, NY: AMLBook: 2012
42. Bossuyt PM, Reitsma JB, Bruns DE, et al. Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *AJR* 2003; 181:51–55
43. Oakden-Rayner L, Beam AL, Palmer LJ. Medical journals should embrace preprints to address the reproducibility crisis. *Int J Epidemiol* 2018 Jun 3 [Epub ahead of print]
44. Cheng PM, Tejura TK, Tran KN, Whang G. Detec-

tion of high-grade small bowel obstruction on conventional radiography with convolutional neural networks. *Abdom Radiol (NY)* 2018; 43:1120–1127

45. Rajpurkar P, Irvin J, Zhu K, et al. CheXNet: radiologist-level pneumonia detection on chest x-rays with deep learning. arXiv website. arxiv.org/abs/1711.05225. Last revised December 25, 2017. Accessed October 11, 2018
46. Rajpurkar P, Irvin J, Bagul A, et al. MURA dataset: towards radiologist-level abnormality detection in musculoskeletal radiographs. arXiv website. arxiv.org/abs/1712.06957. Last revised May 22, 2018. Accessed October 11, 2018
47. Larson DB, Chen MC, Lungren MP, Halabi SS, Stence NV, Langlotz CP. Performance of a deep-learning neural network model in assessing skeletal maturity on pediatric hand radiographs. *Radiology* 2018; 287:313–322

APPENDIX 1: Questions a Reviewer Should Ask

Is the study aimed at solving a practical clinical issue?
Is the writing accessible to physicians and biomedical researchers?
Is the purpose of the study clearly stated?
Do the authors clearly explain data collection, processing, and division methods?
Do the data appropriately represent the range of possible patients and disease manifestations?
Are the data labels (if applicable) of sufficient quality to support the claimed performance of the algorithm or algorithms?
Do the authors report a sufficient number and type of performance measures to accurately represent strengths and weaknesses of the algorithms?
Are performance measures reported with confidence intervals?
If expert-level performance is claimed, does the standard of comparison meet an appropriate level of expertise?
Are comparisons with human performance reported with confidence intervals or *p* values?
Do the authors provide graphics that show the algorithm is detecting the relevant regions of the images and not overfitting to unrelated features?
If performance measures require an explicit threshold, do the authors provide convincing support for the threshold used?
Do the authors appropriately qualify the strength of their conclusions and discuss limitations in their methods?
Do the authors discuss directions for future research?

American Journal of Roentgenology