AJR FOCUS ON:

Implementing Machine Learning in Radiology Practice and Research

Marc Kohli[1]
Luciano M. Prevedello[2]
Ross W. Filice[3]
J. Raymond Geis[4]

OBJECTIVE. The purposes of this article are to describe concepts that radiologists should understand to evaluate machine learning projects, including common algorithms, supervised as opposed to unsupervised techniques, statistical pitfalls, and data considerations for training and evaluation, and to briefly describe ethical dilemmas and legal risk.

CONCLUSION. Machine learning includes a broad class of computer programs that improve with experience. The complexity of creating, training, and monitoring machine learning indicates that the success of the algorithms will require radiologist involvement for years to come, leading to engagement rather than replacement.

Keywords: artificial intelligence, imaging, informatics, machine learning, statistics

DOI:10.2214/AJR.16.17224

Received August 22, 2016; accepted after revision November 22, 2016.

[1]Department of Radiology and Biomedical Imaging, University of California, San Francisco, 505 Parnassus Ave, M-391, San Francisco, CA 94143. Address correspondence to M. Kohli (marc.kohli@ucsf.edu).

[2]Department of Radiology, The Ohio State University Wexner Medical Center, Columbus, OH.

[3]Department of Radiology, MedStar Georgetown University Hospital, Washington, DC.

[4]Department of Radiology, University of Colorado School of Medicine, Fort Collins, CO.

AJR 2017; 208:754–760

0361–803X/17/2084–754

It is difficult to ignore the growing interest in machine learning (ML). ML algorithms generate public excitement because they include playing games against humans [1], self-driving cars, and identifying the characteristics of a great selfie [2]. In radiology, they may help identify schizophrenia with brain MRI [3] and identify genetic markers in glioblastoma [4]. This article introduces radiologists to ML and describes considerations for initiating and evaluating ML projects.

Machine Learning Overview

ML comprises a broad class of statistical analysis algorithms that iteratively improve in response to training data to build models for autonomous predictions. In other words, computer program performance improves automatically with experience [5]. The goal of an ML algorithm is to develop a mathematic model that fits the data. Once this model fits known data, it can be used to predict the labels of new data. Because radiology is inherently a data interpretation profession—in extracting features from images and applying a large knowledge base to interpret those features—it provides ripe opportunities to apply these tools to improve practice.

Many ML algorithms are both small (hundreds of lines of code) and widely applicable in that base algorithms can be applied to disparate domains. As a result, it is fairly easy to start an ML project—if users have the appropriate foundation. Although these models may be straightforward to build, many topics outside the typical knowledge of either academic or practicing radiologists affect whether their models produce accurate, useful results.

ML algorithms start with a set of available inputs and desired outputs. Common inputs in radiology are image data and report text. Output takes the form of a set of conditions and associated probabilities. As an example, we are training a network called MageNet to identify animals in images. If we feed the network a picture of a domestic dog (input), it returns the following list (output): domestic dog, 92%; wolf, 7%; fox, 0.2%; horse, 0.01% (Fig. 1A). With a picture of a lion as an input, the network returns a different set of outputs: domestic cat, 70%; lion, 10%; leopard, 5%; cheetah, 2% (Fig. 1B). In the second example, the network incorrectly classified the lion as a domestic cat. The work of a network is performed in the hidden layers. Hidden layers are sets of equations with several numeric weights that operate with input data and output statistical probabilities, also known as nodes. The numeric weights inside the hidden nodes are called hyperparameters. Trained hidden layers function as feature detectors by detecting the best weight and pattern of activation of the nodes for a specific outcome. Backpropagation is used to help readjust the values within a network. This is done in two steps: forward and backward. Training data (images with known labels) are fed through the algorithm in the forward phase of the computation. In the

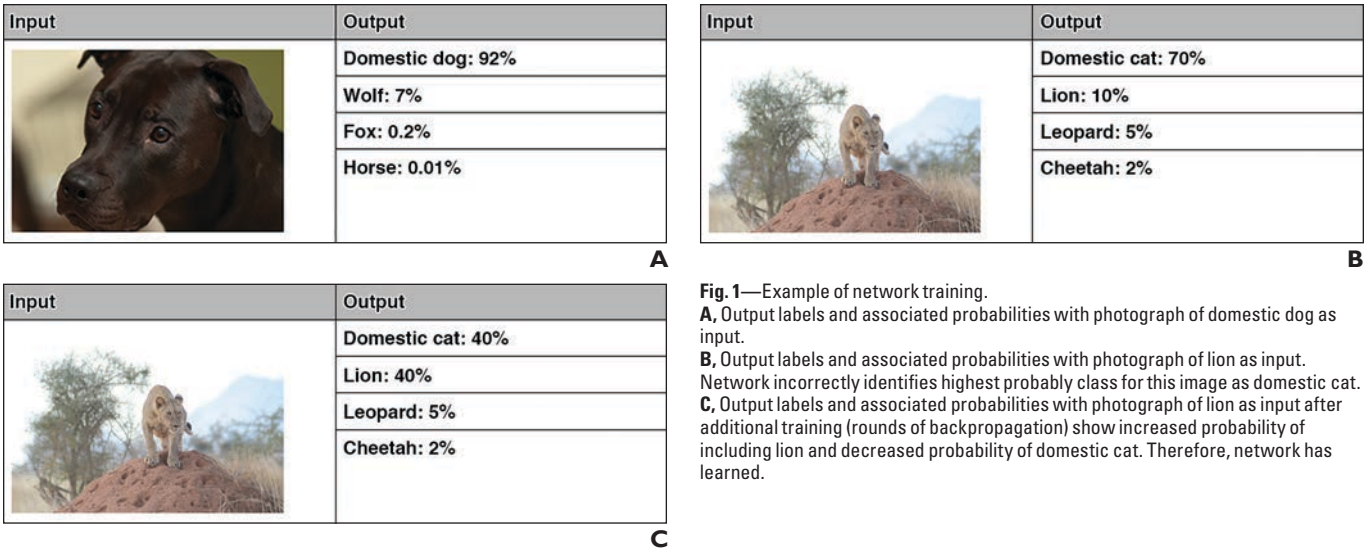

Fig. 1—Example of network training.
A, Output labels and associated probabilities with photograph of domestic dog as input.
B, Output labels and associated probabilities with photograph of lion as input. Network incorrectly identifies highest probably class for this image as domestic cat.
C, Output labels and associated probabilities with photograph of lion as input after additional training (rounds of backpropagation) show increased probability of including lion and decreased probability of domestic cat. Therefore, network has learned.

backward step the differences between the computed output and the label are used to adjust the hyperparameters within the hidden nodes. This second step is also known as backpropagation. Through several iterations of forward propagation and backpropagation, the network learns how to best approximate the desired outcome. Figure 1C shows the network output for the same image of a lion after several rounds of backpropagation.

The classic concept of three layers (input, hidden, and output) of ML algorithms is displayed graphically in Figure 2A. Deep learning combines multiple hidden layers of neural networks (or other classifiers) to tackle advanced classification tasks. In Figure 2B, each hidden vertical box indicates a classic neural network. In this example, the deep neural network contains 32 nodes (four per layer, eight layers). In addition, each node has several hyperparameters, which are adjusted through backpropagation. A network trained to distinguish two classes requires fewer nodes and layers than one attempting to differentiate 10 or 100 classes.

Two specific types of ML algorithms are support vector machines (SVMs) and convolutional neural networks (CNNs). SVMs are useful for taking a large number of features and discriminating inputs into one of two classes. Figure 3 shows an SVM that has two classes: diamonds and circles. These inputs have two features, X and Y, and are plotted in 2D space. Figure 3A shows several potential lines that can be used to differentiate the two classes. SVMs, once trained, show the line that provides the greatest margin of separation (Fig. 3B). Although this example is limited to two features, this concept can be extrapolated to a larger number of features (or dimensions) whereby the line of separation becomes an irregular plane known as a hyperplane. Because of the large number of features that can be combined mathematically, SVMs have been found useful for image processing.

CNNs are a specific type of neural network that have features useful for image analysis. Convolutions are mathematic transformations (similar to a basic filter in a photograph editing application) that are applied to pixel data. Some typical convolutions, such as embossing

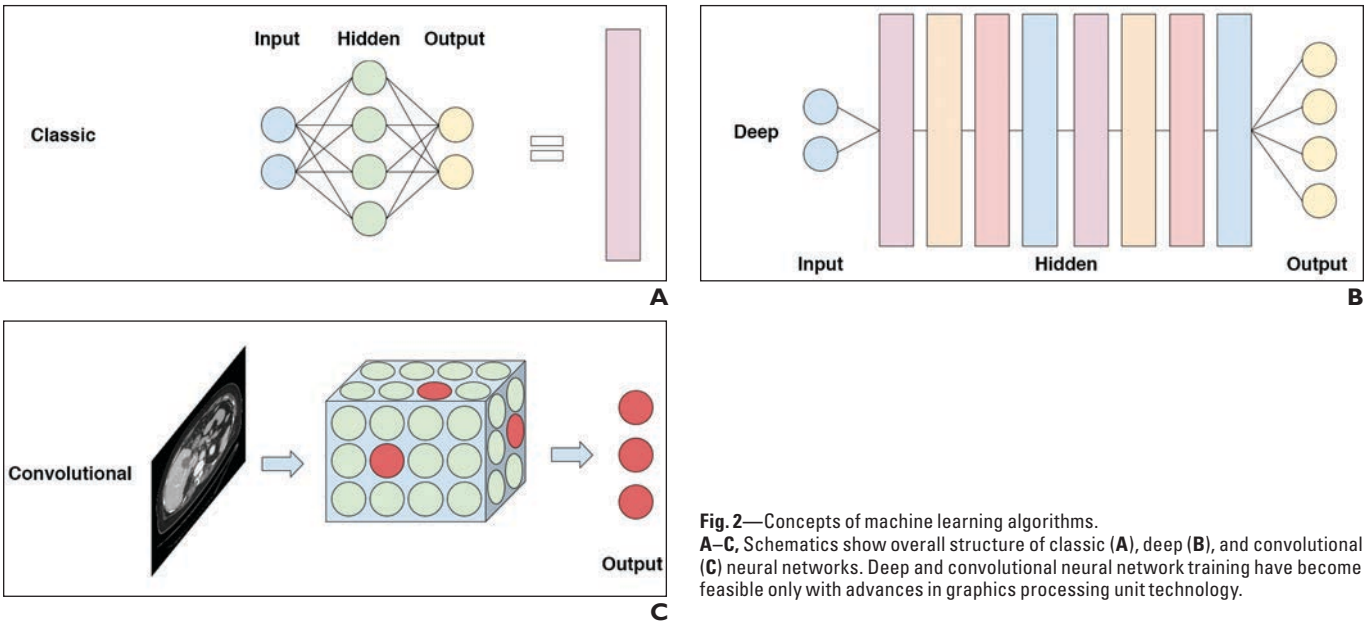

Fig. 2—Concepts of machine learning algorithms.
A–C, Schematics show overall structure of classic (**A**), deep (**B**), and convolutional (**C**) neural networks. Deep and convolutional neural network training have become feasible only with advances in graphics processing unit technology.

and outline detection, are depicted in Figure 4A. In addition to using known convolution techniques, some CNNs apply random transformations to the images they process, evaluate the outcome, and adjust the parameters for each convolution. As the model is repeatedly trained, individual convolutions begin to identify a specific portion of the image. Hundreds of these classifiers can be linked together to identify more complex structures within each image. Figure 4B shows an overlay of numerous image convolutions on a CT scan. Swirls surround the kidneys and the gallbladder on the image. Each of these convolutions adds a dimension to the dataset. However, rather than each convolution being tightly linked to the one before or after, CNNs find clusters of locally connected neurons and weight each cluster for a given input. For example, a cluster of neurons may help identify the kidney or liver within an image. Traditional image analysis pipelines rely heavily on the performance of the previous steps, making classification adjustments a complicated and arduous process. Deep learning and CNN, on the other hand, exploit spatially and structurally associated features and tune their performance automatically. This makes algorithm creation less complicated and easier to adjust. Deep learning and CNN can be used to automatically preprocess images, as for lesion segmentation, a major advantage over manual processes. The result is that in complicated situations such as image analysis, these multilevel algorithms improve faster and, once trained, require less computing power.

The following factors are to be considered in the choice of an algorithm: the type of pattern, the number of hyperparameters included, the resources (time and computing power) available for the project, and the data on which to train and then examine. ML data patterns include identifying clusters and other structures, grouping into two or more classifications, identifying outliers, and predicting values. Hyperparameters are variables within an algorithm that may be adjusted when the model is tuned to optimize results. Increasing the number of hyperparameters often increases algorithm accuracy, although too many can contribute to overfitting the data.

Critical resources for ML projects include computing power, time, and money. Time and processing power are inversely proportional—with more processing power, models take less time to train. Some projects run easily on desktop computers, but others require dedicated hardware. Advances in graphics processing unit (GPU) computing greatly accelerate ML training. GPU-based computing takes advantage of parallel processing, which accelerates throughput similar to parallel acquisitions in MRI. Training the model is computationally expensive. Once it is fully trained, however, it is usually efficient and requires relatively little computing power. Thus, many copies can be deployed to evaluate new images with readily available computers, such as PACS workstations and even smartphones and tablets.

It may be useful to train several types of algorithms with the same data, identifying advantages and disadvantages of each. Results from disparate algorithms may be combined or used in a majority rules scenario. Numerous ML algorithms are readily available, each having advantages and disadvantages. Algorithm selection is a broad topic, described elsewhere [6].

Supervised Versus Unsupervised Machine Learning

Most ML relevant to radiology is supervised. In supervised ML, data are labeled before the model is trained. For example, in training a project to identify a specific brain tumor type, the label would be tumor pathologic results or genomic information. These labels, also known as ground truth, can be as specific or general as needed to answer the question. The ML algorithm is exposed to enough of these labeled data to allow them to morph into a model designed to answer the question of interest. Because of the large number of well-labeled images required to train models, curating these datasets is often laborious and expensive.

Thoughtful project selection is critical to successful ML projects. Successful projects have clearly defined outcomes for which meaningful ground truth can be easily established. A researcher seeking to use ML to detect pulmonary embolism on chest CT studies must determine whether an existing radiology report is a suitable label or whether follow-up imaging of the patient at a fixed time point is more appropriate. This is less of an issue with focused, tightly defined projects, such as trying to differentiate brain tumor subtypes, but is an important and incompletely studied topic for situations such as screening examinations, in which studies originally judged normal in time turn out to have false-negative findings.

In unsupervised ML, unlabeled data are exposed to the algorithm with the goal of generating labels that will meaningfully organize the data. This is typically done by identifying useful clusters of data based on one or more dimensions. Compared with supervised techniques, unsupervised learning sometimes requires much larger training datasets. Unsupervised learning is useful in identifying meaningful clustering labels that can then be used in supervised training to develop a useful ML algorithm. This blend of supervised and unsupervised learning is known as semisupervised.

Statistics and Machine Learning

Statistics play a major role in ML algorithms. Statistical theories and techniques for

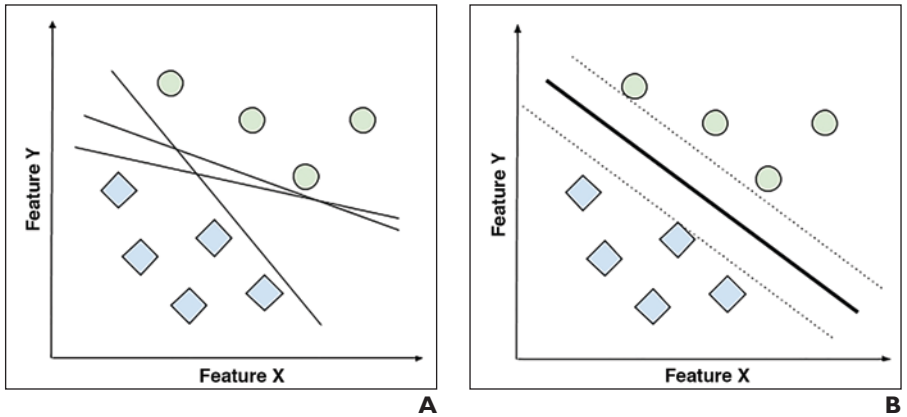

Fig. 3—Graphical depiction of simple support vector machine (SVM) used to discriminate between two labeled classes (circles and diamonds), each of which has two features X and Y.
A, Graph obtained with cartesian coordinates shows several potential lines that can be used to differentiate classes.
B, Graph obtained with trained SVM shows line that separates two classes with greatest margin. This can be extrapolated to hundreds of features, and line becomes known as hyperplane.

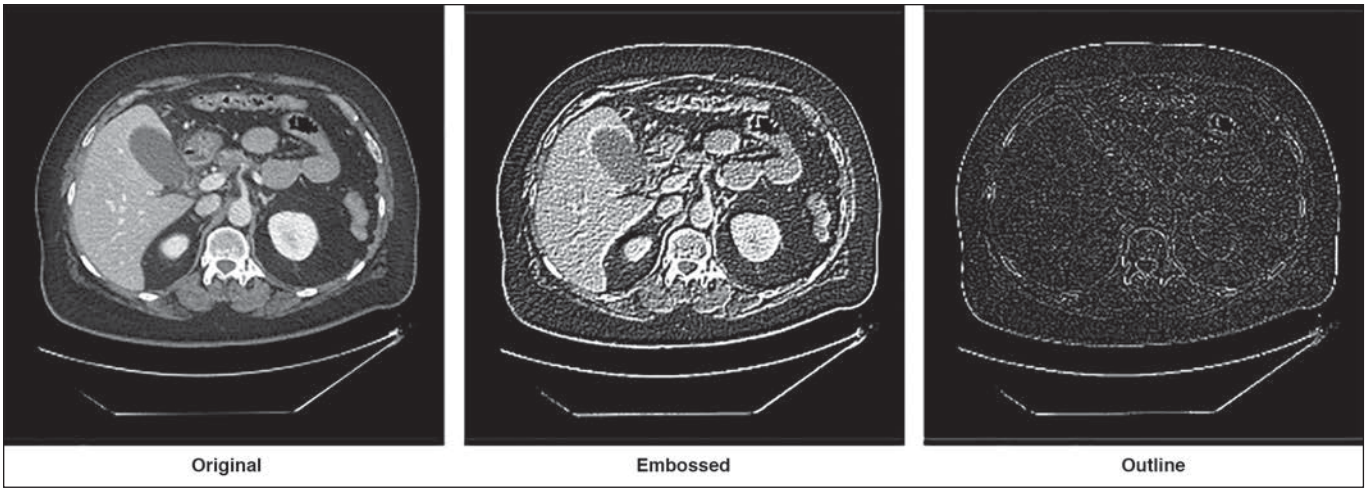

Fig. 4—Examples of image convolutions.
A, Common 3 × 3 matrices including embossing and outline detection.
B, Source image overlaid with several classifiers from convolutional neural network deep learning system. Several classifiers appear to have highlighted structures, such as kidneys and gallbladder.

ML are largely different from those traditionally used in radiology research. The basis of ML is statistical inference, in which characteristics of a population are extrapolated on the basis of sampling only a portion of the total group. Each ML algorithm contains a statistical model, or set of assumptions about the observed (training) data, which is equally applied to new data, on which predictions are made. Statistical models fall into three broad categories: parametric, nonparametric, or semiparametric, which has some aspects of the other two. Although detailed discussion of each of these models is outside the scope of this article, understanding when to use each type is critical to success. For example, imaging processing and pattern recognition often entail nonparametric models with high dimensions, because images cannot be described by a general distribution function. Thus, the more precisely the image is sampled, the more complex the estimation becomes. In contrast, a gaussian curve follows a function that is definable by sampling, and more sampling makes the estimate less complex. In this situation, parametric models, which embody assumptions about the entire data population and with which most radiologists are most familiar, are best.

Commonly used statistical templates include frequentist, or classic; bayesian; and other less familiar templates, such as Akaike information criterion. Frequentist inference is the traditional approach with CIs and *p* values. It is most applicable when the data can be described with a function such as a gaussian distribution. It can be used to predict a future event and typically gives that prediction as an exact number. Bayesian inference is commonly used in ML, because it is both appropriate to many common situations and is fairly easy to model with computer code. Bayesian inference delivers sequential improvement in predictions, a core ML goal. Rather than providing an exact prediction, bayesian inference gives a probability distribution for a future event. It is based on the probability distributions of prior events, and it updates that probability as more training data become available. It can be thwarted, however, by skewed training data. Akaike information criterion and similar approaches may be appropriate in extremely complex settings, such as high dimensions or populations without defined boundaries, some of which occur with true big data and potentially apply to less tightly defined image interpretation models. These brief descriptions illustrate the need for care when choosing and building algorithms.

For any statistical model, one must make correct assumptions about the data and have methods of verifying those assumptions. Common errors include assuming random samples when the data are not truly random; incorrectly assuming normality in the population; and assuming data regularity when in reality the data change inconsistently or erratically. The risks of inappropriate statistical methods are all too apparent in the work [7] that exposed software bugs and inappropriate methods that call into question the results of nearly 40,000 research studies.

Another important consideration in dealing with complex statistical models or deep learning projects is to recognize prediction errors due to bias or variance. Bias relates to the degree to which predictions made with the model differ from the correct value. Variance is the degree to which predictions for a given point vary between different instances of the model (Fig. 5).

The ideal model minimizes both bias and variance. Often it depends on training data sample size and the number of hyperparameter attributes. Bias increases and variance decreases with increased sample size up to the point at which they stabilize. Models with high bias may require more hyperparameters, or the model may have to be redesigned. Models with high variance may benefit from fewer hyperparameters or increased sample size, depending on the situation.

No ML algorithm is 100% accurate in prediction of future events. Every ML algorithm contains an error measure, which reflects the number of cases incorrectly described by the algorithm. An important question to answer beforehand is: What is the acceptable error

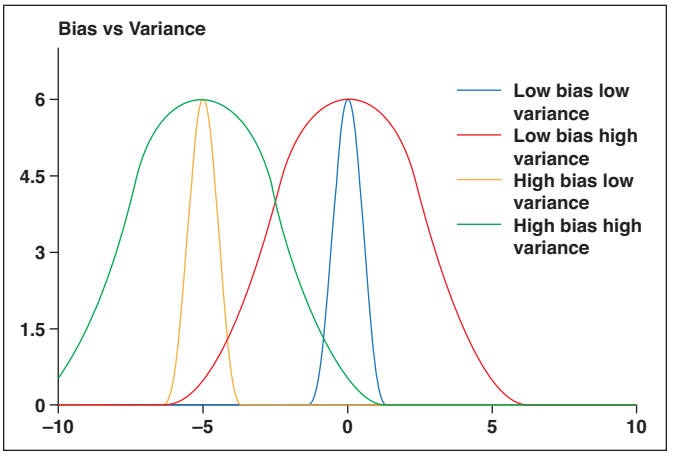

Fig. 5—Graph shows example of four hypothetical models with variable degrees of bias and variance with expected prediction of zero.

measure? Is it 60%, 95%? The error measure goal helps to determine how many training data, and what types of algorithms, to use. Unrealistic expectations of error rates can destroy a project, either because the rate is unobtainable or it requires impractical amounts of computer and time resources.

Medical Image Data for Machine Learning

At the outset of an ML project, data are divided into three sets: training, test, and validation. The training dataset is sent through the algorithm repeatedly to establish values for each hyperparameter. After the hyperparameters stabilize, the test dataset is sent through the model, and the accuracy of the predictions or classifications is evaluated. At this point, the trainer decides whether the model is fully trained or adjusts the algorithm architecture to repeat training. After several iterative cycles of training and testing, the algorithm is fed validation data for final evaluation. Application of ML to radiology involves both medical knowledge and pixel data.

Data characteristics for ML include not only volume and accuracy but also velocity and dimensions. Data velocity indicates how quickly data can be delivered for processing. Velocity is a spectrum ranging from batch to real-time processing and has implications for how ML systems are implemented. Dimensions are independent features of a datum. For example, for a person, dimensions may include name, height, weight, address, and hair color. A lesion on a medical image may have anywhere between a few and thousands of potential dimensions. Common dimensions, such as pixel value, lesion diameter, and volume, are supplemented with other computer-generated observations, such as texture analysis [8]. Each DICOM header field can be a dimension, as can clinical, demographic, and even social media or GPS data. Because medical image pixel data are usually highly dimensional, complex, and variable, they are difficult to represent with formulas and thus are problematic to model. When data have numerous dimensions, dimensionality reduction programs are often used to reduce dimensionality to a practical number for available computer power [9]. This introduces additional risk of overfitting and adds complexity to algorithm design and hyperparameter selection.

Medical image data lack standards for ground truth labeling. Sometimes radiology reports are accurate sources, but other scenarios require pathologic results, clinical follow-up, genomic data, and comparison with other imaging studies. Both normal and abnormal anatomic features typically are highly dimensional and have sizable individual variations. A lesion or abnormal feature may hide within these variations. Pattern recognition for complex, high-dimensionality images are generally trained on large datasets, but such datasets, particularly with appropriate labels, are rare. To produce such sets can be expensive and time-consuming because labeling is difficult, and preprocessing of images must typically be performed to provide useful inputs to the ML algorithm. Newer deep learning and CNN techniques can help by incorporating the image-preprocessing step into the algorithm itself, saving manual labor and potentially leading to the discovery of preprocessing techniques that perform better in the subsequent neural network layers. These techniques require either tightly focused and well defined data or extremely large datasets.

Transfer Learning

Although the history of neural networks and ML goes back decades, more recent advancements have ignited interest in the computer science community and the general news media. One advancement, particularly interesting to radiology, is the ImageNet recognition challenge. The ImageNet collaboration maintains a large dataset (currently 14 million images) that are labeled with nouns related to the content of each image [10]. In addition to providing annotated images, ImageNet sponsors annual events at which computer science groups from around the world submit trained algorithms in an attempt to classify images from a subset of the ImageNet data for higher and higher levels of accuracy. In 2012, Krizhevsky et al. [11] reported dramatic improvement in recognition by using a deep CNN (AlexNet) with 60 million hyperparameters spread across 650,000 nodes trained on 1.2 million images. Researchers in computer science have found that as deep CNNs are trained, the first layer of neurons begins to recognize characteristics such as shape and color, which has led researchers to evaluate whether these trained networks can be generalized to other tasks in a process called transfer learning [12]. This is especially interesting for radi-

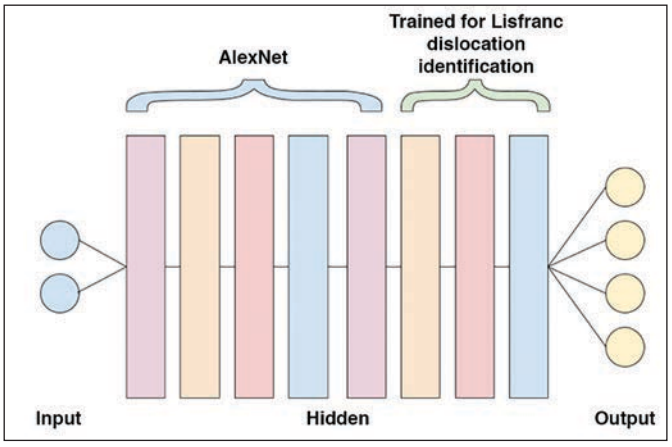

Fig. 6—Graphical representation of network trained to recognize Lisfranc dislocation by use of AlexNet as foundation for transfer learning.

ology applications, given the lack of massive annotated datasets with millions of images. For this process, researchers take a trained network and add layers that are trained in relation to a specific task, such as identification of Lisfranc dislocations (Fig. 6). Limitations of this type of technique include resolution and bit-depth constraints and the assumption that classifiers trained on photographic images are relevant to medical images. AlexNet limited down-sampled ImageNet images to 256 × 256 (8-bit red-green-blue color), such that any transfer learning project must similarly downsample images.

Data Augmentation

Data augmentation is an important strategy for maximizing the usefulness of a well-curated image training dataset. ML algorithms generally benefit from large well-curated image datasets, which are hard to come by in the clinical radiology world because of both the large amount of tedious work and the high degree of expertise required. Use of image morphing techniques, in which the appearance of an object of interest is slightly modified, can increase an image training dataset by many multiples. Examples include skewing the image slightly, modifying the contrast or resolution slightly, flipping or rotating the image, adjusting zoom, and changing the location of a finding within an image. These strategies essentially present a slightly different appearance of the same finding but can make an ML algorithm much more robust with a relatively small curated dataset—important in radiology, in which it is difficult to acquire good data.

Computer-Aided Detection Versus Machine Learning

It is worthwhile to distinguish ML from traditional computer-aided detection (CAD) algorithms. Traditional CAD algorithms are mathematic models that identify the presence or absence of image features known to be associated with a disease state. One such example is a microcalcification on a mammogram. With traditional CAD, the developer identifies a feature explicitly and attempts to determine the presence or absence of that feature within a set of images. In contrast, ML techniques focus on a particular labeled outcome (ductal adenocarcinoma), and in the process of training, clusters of nodes evolve into algorithms for identifying features. The power and promise of the ML approach over traditional CAD is that useful features can exist that are not currently known or are beyond the limit of human detection.

Existing Research in Medical Imaging Machine Learning

Because of the challenges, effective radiology ML projects thus far have focused on tightly defined projects amenable to available training datasets. Summers [13] provided an in-depth review of state-of-the-art automated interpretation of abdominopelvic CT scans. He described numerous techniques for segmentation and analysis of organs and tissues, including traditional CAD and ML approaches. The review is an excellent source for learning more about the existing research landscape.

In an early application of ML, Yao et al. [14] used texture analysis and trained SVMs to differentiate normal and abnormal lung and to differentiate airspace opacity from fibrotic changes. Especially with transfer learning in which input resolution is limited, techniques for identifying an ROI out of a larger image set will be valuable as preprocessors.

Another area of promise is to use ML to detect phenotypes that may not be readily apparent to an interpreting radiologist. Korfiatis et al. [4] used SVMs to predict O6-methylguanine methyltransferase (MGMT) gene promoter methylation in glioblastoma multiforme tumors on the basis of intensity variations on MR images. The SVM had an AUC of 0.85 even when trained on only 155 cases. Another highlight of that study was the large amount of preprocessing required for ML even for very small datasets.

Deep learning has been applied to other facets of radiology, including early detection of breast cancer with mammography and ultrasound, classification of chest radiographs, and differentiating malignant from benign nodules on chest CT images [15–17].

Ethical Dilemmas and Legal Risk

Because applications of ML in radiology are new and few, there are more questions than answers when it comes to ethics. We present topics for further discussion rather than give specific advice. Although deep learning works, it is often difficult to elucidate what is happening in the multiple hidden layers. Many commercially deployed models evolve as they are exposed to new data. For example, voice assistants developed by companies such as Apple and Google are constantly being improved as more people use the services and enlarge the pool of voice data available for training. The risk and benefit of training with live data will have to be established scientifically and constantly monitored for unintended consequences.

At present the algorithms are being applied in highly controlled situations. But what happens when they are in wide distribution? Should individual institutions be allowed to augment validated models with additional training to develop setting-specific predictions? Or should ML products being used in practice be subject to code freezes, in which the model is not changed while in live practice? If models are allowed to evolve, how will they be regulated by the U.S. Food and Drug Administration (FDA)? If an entity develops a proprietary ML model and describes it as doing a particular task, such as identifying a specific pathologic entity, how should the model, and the associated claim, be verified?

When ML or deep learning is used to drive subsequent action, additional ethical questions arise. Is the definition of the best algorithm that which improves the life span of each patient, no matter what the cost? If cost is factored into the algorithm, it is easy to see any number of untoward and unethical machine-generated outcomes. Suppose an unsupervised algorithm shows the pattern that for disease X a person with a high income with insurance Y usually receives an expensive treatment and does well but that a person with different insurance has access to only a less successful treatment. The algorithm would no doubt incorporate those different recommendations. At the least, these variations must be transparent, and the supervising physician alerted to what is happening.

The FDA has not issued rules about test datasets, transparency, or verification procedures. It will probably evaluate models and associated test datasets on a case by case basis. How this will evolve is unclear at present. In addition, regulation that created the FDA was enacted before the availability of ML, and existing laws regarding devices are difficult to apply to ML algorithms.

Liability issues may become more challenging as ML algorithms become more widespread. Currently most experts envision ML supporting a radiologist who would make the final decision or interpretation, thus retaining the primary responsibility or liability for any outcome. However, as ML algorithms become more advanced, this liability may shift toward the algorithm, companies, developers, and

those responsible for training the algorithm, which will pose challenges in medical malpractice and liability in radiology.

Conclusion

ML encompasses many powerful tools with the potential to dramatically increase the information radiologists extract from images. It is no exaggeration to suggest the tools will change radiology as dramatically as the advent of cross-sectional imaging did. We believe that owing to the narrow scope of existing applications of ML and the complexity of creating and training ML models, the possibility that radiologists will be replaced by machines is at best far in the future. Successful application of ML to the radiology domain will require that radiologists extend their knowledge of statistics and data science to supervise and correctly interpret ML-derived results.

Resources for Further Study

Educational resources and online courses on ML are available at Andrej Karpathy's blog; the NVIDIA Deep Learning Institute website; the Stanford University machine learning course at Coursera; the online course CS231n: Convolutional Neural Networks for Visual Recognition; and the deeplearning4j course titled Convolutional Networks.

References

1. Metz C. In two moves, AlphaGo and Lee Sedol redefined the future. *Wired* www.wired.com/2016/03/two-moves-alphago-lee-sedol-redefined-future. March 16, 2016. Accessed August 4, 2016
2. Karpathy A. What a deep neural network thinks about your #selfie. Andrej Karpathy blog. karpathy.github.io/2015/10/25/selfie. October 25, 2015. Accessed August 4, 2016
3. Lu X, Yang Y, Wu F, et al. Discriminative analysis of schizophrenia using support vector machine and recursive feature elimination on structural MRI images. *Medicine (Baltimore)* 2016; 95:e3973
4. Korfiatis P, Kline TL, Coufalova L, et al. MRI texture features as biomarkers to predict MGMT methylation status in glioblastomas. *Med Phys* 2016; 43:2835–2844
5. Jordan MI, Mitchell TM. Machine learning: trends, perspectives, and prospects. *Science* 2015; 349:255–260
6. Ericson G, Franks L, Gronlund CJ, Rohrer B. Machine learning algorithm cheat sheet for Microsoft Azure machine learning studio. Microsoft Azure website. azure.microsoft.com/en-us/documentation/articles/machine-learning-algorithm-cheat-sheet. December 14, 2016. Accessed August 4, 2016
7. Eklund A, Nichols TE, Knutsson H. Cluster failure: why fMRI inferences for spatial extent have inflated false-positive rates. *Proc Natl Acad Sci USA* 2016; 113:7900–7905
8. Depeursinge A, Foncubierta-Rodriguez A, Van De Ville D, Müller H. Three-dimensional solid texture analysis in biomedical imaging: review and opportunities. *Med Image Anal* 2014; 18:176–196
9. Fodor IK. A survey of dimension reduction techniques. U.S. Department of Energy, Office of Scientific and Technical Information website. www.osti.gov/scitech/servlets/purl/15002155. May 9, 2002. Accessed August 4, 2016
10. Deng J, Dong W, Socher R, et al. ImageNet: a large-scale hierarchical image database. In: Essa I, Kang SB, Pollefeys M, eds. *2009 IEEE Conference on Computer Vision and Pattern Recognition*. Piscataway, NJ: IEEE, 2009:248–255
11. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In: Pereira F, Burges CJ, Bottou L, Weinberger KQ, eds. *Advances in neural information processing systems*. Vol. 25. Red Hook, NY: Curran 2012:1097–1105
12. Yosinski J, Clune J, Bengio Y, Lipson H. How transferable are features in deep neural networks? In: Ghahramani Z, Welling MW, Cortes C, Lawrence ND, Weinberger KQ, eds. *Advances in neural information processing systems*. Vol 27. Red Hook, NY: Curran, 2014:3320–3328
13. Summers RM. Progress in fully automated abdominal CT interpretation. *AJR* 2016; 207:67–79
14. Yao J, Dwyer A, Summers RM, Mollura DJ. Computer-aided diagnosis of pulmonary infections using texture analysis and support vector machine classification. *Acad Radiol* 2011; 18:306–314
15. Wang J, Yang X, Cai H, et al. Discrimination of breast cancer with microcalcifications on mammography by deep learning. *Sci Rep* 2016; 6:27327
16. Rajkomar A, Lingam S, Taylor AG, et al. High-throughput classification of radiographs using deep convolutional neural networks. *J Digit Imaging*; 2016 Oct 11 [Epub ahead of print]
17. Cheng JZ, Ni D, Chou YH, et al. Computer-aided diagnosis with deep learning architecture: applications to breast lesions in US images and pulmonary nodules in CT scans. *Sci Rep* 2016; 6:24454