Machine Learning in Nuclear Medicine: Part 2—Neural Networks and Clinical Aspects

Katherine Zukotynski*[1], Vincent Gaudet*[2], Carlos F. Uribe[3], Sulantha Mathotaarachchi[4], Kenneth C. Smith[5], Pedro Rosa-Neto[4], François Bénard[3,6], and Sandra E. Black[7]

[1]Departments of Medicine and Radiology, McMaster University, Hamilton, Ontario, Canada; [2]Department of Electrical and Computer Engineering, University of Waterloo, Waterloo, Ontario, Canada; [3]PET Functional Imaging, BC Cancer, Vancouver, British Columbia, Canada; [4]Translational Neuroimaging Lab, McGill University, Montreal, Quebec, Canada; [5]Department of Electrical and Computer Engineering, University of Toronto, Toronto, Ontario, Canada; [6]Department of Radiology, University of British Columbia, Vancouver, British Columbia, Canada; and [7]Department of Medicine (Neurology), Sunnybrook Health Sciences Centre, University of Toronto, Toronto, Ontario, Canada

Learning Objectives: On successful completion of this activity, participants should (1) be familiar with neural networks and (2) have an understanding of where they can be helpful in clinical practice.

Financial Disclosure: Dr. Neto is an unpaid consultant/advisor for Cerveau Radiopharmaceuticals. The authors of this article have indicated no other relevant relationships that could be perceived as a real or apparent conflict of interest.

CME Credit: SNMMI is accredited by the Accreditation Council for Continuing Medical Education (ACCME) to sponsor continuing education for physicians. SNMMI designates each *JNM* continuing education article for a maximum of 2.0 AMA PRA Category 1 Credits. Physicians should claim only credit commensurate with the extent of their participation in the activity. For CE credit, SAM, and other credit types, participants can access this activity through the SNMMI website (http://www.snmmilearningcenter.org) through January 2024.

This article is the second part in our machine learning series. Part 1 provided a general overview of machine learning in nuclear medicine. Part 2 focuses on neural networks. We start with an example illustrating how neural networks work and a discussion of potential applications. Recognizing that there is a spectrum of applications, we focus on recent publications in the areas of image reconstruction, low-dose PET, disease detection, and models used for diagnosis and outcome prediction. Finally, since the way machine learning algorithms are reported in the literature is extremely variable, we conclude with a call to arms regarding the need for standardized reporting of design and outcome metrics and we propose a basic checklist our community might follow going forward.

Key Words: machine learning; nuclear medicine; neural networks

J Nucl Med 2021; 62:22–29
DOI: 10.2967/jnumed.119.231837

Part 1 in our series on machine learning (ML) in nuclear medicine (*1*) provided a general overview of ML algorithms and their basic components. Although applications of ML algorithms such as random forests (*2–4*) and support vector machines (*5,6*) continue to proliferate, sophisticated ML algorithms such as artificial neural networks (ANNs) are becoming ubiquitous. Further, ANNs using radiomic data are increasingly common in nuclear medicine applications. The term *radiomic data* typically refers to quantitative data from medical images, such as texture values enabling assessment of tumor heterogeneity, extracted either manually or using a computer-based approach (*7*). Part 2 provides an expanded explanation of ANNs, one of the most powerful ML models used today. After a brief review of the ANN concepts introduced in part 1, we illustrate how ANNs work using an example and follow this with a brief discussion of clinical applications.

Received Apr. 26, 2020; revision accepted Aug. 13, 2020.
For correspondence or reprints contact: Katherine Zukotynski, Departments of Medicine and Radiology, McMaster University, 1200 Main St. W., Room 1P11, Hamilton, ON L8N 3Z5, Canada.
E-mail: katherine.zukotynski@utoronto.ca
*Contributed equally to this work.
Published online Sep. 25, 2020.

BRIEF REVIEW FROM PART 1

ANNs are advanced ML algorithms (Fig. 1) typically used in classification (discrete-output) or regression (analog-output) applications. Although ANNs have existed for decades, they have only recently become common in medical imaging, in part due to technologic advances as well as access to large datasets for training. Data input to an ANN is processed in steps, where each step consists of a layer of neurons. A neuron is a computational unit that produces a weighted summation of input data, applies a bias, and computes a nonlinear transformation of the result. The output data from each layer pass to the next layer until the final layer produces the output result. The architectural design of an ANN describes the relationship between the various neurons and layers. ANNs are typically supervised, using tagged data to learn weights and biases, and can be simple, including only a few layers and a single output, or complex. More complex ANNs generally have greater capabilities but at higher computational cost. Complex ANNs are used for deep learning. Designing an ANN of optimal complexity to solve a specific task and obtaining access to sufficient high-quality input data are challenging. Today, ANNs are among the most common ML algorithms used in nuclear medicine. and understanding how they work is key.

In this text, we try to convey the structure and purpose of an ANN. To illustrate, the next section starts with an example of a simple ANN (with a single layer) that could detect a handwritten letter from an input image. We then discuss more complex ANNs and their applications in nuclear medicine. For reference, common

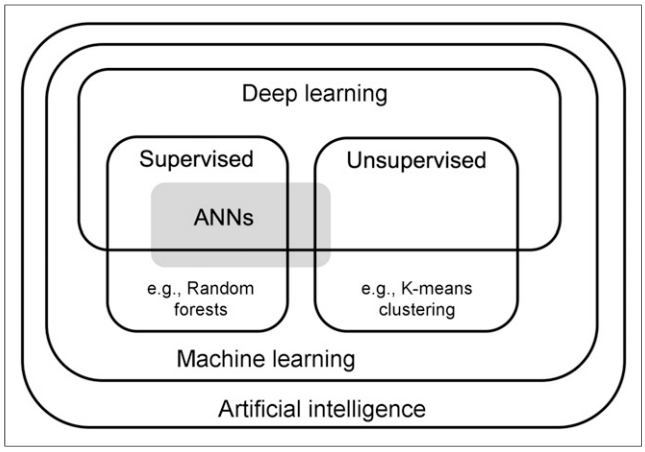

FIGURE 1. Venn diagram depicting relationship of ML, AI, and deep learning. Simple ML algorithms such as random forests and K-means clustering are shown. Complex ML algorithms such as ANNs extend beyond supervised deep learning (where they are primarily used). Algorithms that are neither supervised nor unsupervised—for example, reinforcement learning—are not shown.

ML terms are summarized in Table 1. Further, when we write "nuclear medicine," please note that we implicitly include PET, PET/CT, and PET/MRI.

ANNS: UNDERSTANDING BY EXAMPLE

In this example, we design and train an ANN with only a single layer to recognize a handwritten image of the letter X (Fig. 2). The input to the ANN is an 8 × 8 (2-dimensional) black-and-white image or 64-pixel array/vector (a_1–a_{64}), where each pixel has the value −1 (white) or 1 (black) (Figs. 2A and 2B). The output (f) is an analog number between 0 and 1 that reflects the likelihood that the input is letter X. For example, if the input is an image of X, the output should be 1; if the input is another letter, the output should be 0. Our ANN is trained using a dataset including several input images that have been tagged as representing X or not (Fig. 2B). The ANN output is calculated by multiplying each input by a corresponding weight (w_1–w_{64}), adding the products together (assuming all biases are 0), and passing the result through a nonlinear function (here, a sigmoid function) called an activation function (Fig. 2C).

ANNs can have many types of activation functions, including a sigmoid function and a rectified linear unit (ReLU) (Fig. 2D). Each activation function constrains the output in some manner; for example, the sigmoid function constrains outputs to be between 0 and 1, and the ReLU zeroes out negative numbers (Fig. 2E). These nonlinear functions are key for optimizing ANN performance.

Before the ANN can process new images, it must learn the values for the weights through training. To do this, the algorithm uses a cost (or loss) function that calculates how closely the model predicts the output for a particular training case. The ultimate goal of training is to minimize the cost function. A common cost function is to compute the error (E) between the trained (f_{train}) and desired output (f), possibly the absolute difference between them (the square of the difference, and classification accuracy, are also common cost functions):

$$E = |f - f_{train}|. \quad \text{Eq. 1}$$

The values of w_1–w_{64} that give the best performance are obtained by iterating through the training cases: the weights are initialized,

TABLE 1
Terms Commonly Encountered When Discussing Neural Networks

Term	Explanation	Comment
Fully connected layer	Each input to layer is used to compute each output from layer	Figure 2C illustrates fully connected layer with 64 inputs and 1 output; although number of output data points could be smaller than number of input data points, this is not required
Kernel	Matrix of numbers in CNN where numbers are typically learned through exposure to training dataset	3 × 3 kernels or 3 × 3 × 3 kernels are common
Stride	Number that represents how many pixels a kernel skips each time it processes image in CNN	Figure 6 illustrates stride; output image has fewer pixels than input image, resulting in output image represented by matrix of lower dimension
Pooling	Operation in CNN that reduces image resolution by averaging or taking maximum of local region	Pooling layer could have, as input, image represented by 128 × 128 matrix and produce, as output, image represented by 64 × 64 matrix by dividing input matrix into 2 × 2 blocks and then reducing each block of 4 numbers to 1 number representing maximum value
FLOP	FLOP stands for floating-point operation and represents measure of computing power	FLOPs associated with network typically refer to computing power needed for network to run after it has been trained; in Figure 2, there are 64 multiplications and 63 pairwise additions, representing 127 FLOPs (omitting sigmoid function); CNN might require billions of FLOPs, whereas simple ML algorithm such as random forest or support vector machine might require thousands

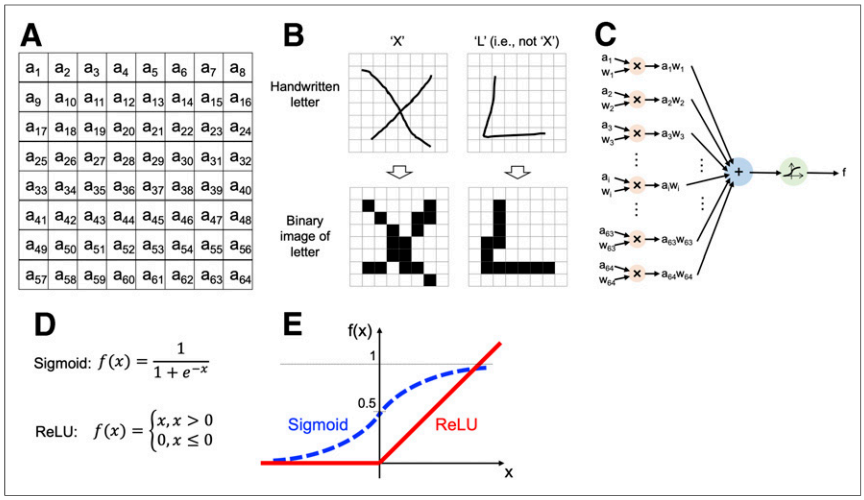

FIGURE 2. (A) Input: 8 × 8 matrix (a_1–a_{64}) of pixels, where each pixel is −1 (white) or +1 (black). (B) Examples from training dataset of handwritten letters mapped to binary input. (C) Single-layer neural network with input shown in A, 64 weights (w_1–w_{64}), sigmoid activation function, and 1 output (f) that is analog number between 0 and 1 reflecting probability that input is X. (D) Mathematic expressions for activation functions. (E) Graphs showing sigmoid (blue) and ReLU (red); for sigmoid, output is constrained to be between 0 and 1; for ReLU, negative inputs are zeroed.

a training case is input to the ANN, the error function is calculated, the weights are adjusted to nudge the ANN toward a lower cost, and a new training case is presented to the ANN. The process is done iteratively until the learned weights give a satisfactory cost.

Weight adjustment is often done using a gradient-descent algorithm, such as stochastic gradient descent. The gradient of the cost function is calculated, essentially the partial derivative (i.e., slope) of the cost with respect to each weight. Once the partial derivatives are known, the weights are adjusted in the direction of steepest descent. However, there is no guidance as to how much weight adjustment is needed. Too little adjustment and little progress is made toward the end-goal; too much adjustment and the output might degrade. Consider a function $E(x)$, where we are trying to identify a minimal point (Fig. 3). If we start from point A, we should move right. However, a large step (point D) moves us too far.

For our ANN to detect the letter X, we derive the partial derivative of E with respect to each weight. First, we express the output f based on the input pixels a_i and weights w_i:

$$f = \frac{1}{1 + e^{-\sum_{i=1}^{64} a_i u w_i}}. \quad \text{Eq. 2}$$

Taking the partial derivative of Eq. 1 with respect to w_i gives

$$\frac{\partial E}{\partial w_i} = \frac{(f_{train} - f)}{E} \frac{\left(e^{-\sum_{j=1}^{64} a_j w_j}\right)}{\left(1 + e^{-\sum_{j=1}^{64} a_j w_j}\right)^2} a_i. \quad \text{Eq. 3}$$

Each partial derivative shows the amount by which its corresponding weight should be adjusted per learning iteration.

We now train the ANN using 24 cases (handwritten samples): 12 of the letter X and 12 of other letters. After training, the weights look like the 8 × 8 matrix shown in Figure 4. The grayscale representation of weights in our ANN resembles an X (Fig. 4A). This makes intuitive sense: since the weights reflect a probability map, an image of the weights resembles what the ANN is trained to detect.

Complex ANNs: Number of Layers and Architecture Design

ANNs capable of deep learning typically have many layers (*8*). Consider the processing involved with your brain reading this text as an example of this multistep processing. It might go as follows: (1) you input an image through your eyes to your brain, (2) your brain identifies strokes and puts strokes together determining how they form a pattern, (3) you recognize the pattern (or character), (4) you assemble neighboring characters and identify words, (5) words come together into sentences, (6) meaning is extracted from sentences, and (7) you process information and perform an analysis. Although a programmer interested in deep learning might create a complex ANN, the task done at each layer is often not predefined by the programmer. Rather, the ANN operates for all intents and purposes as a black box. An ANN with more layers might be able to learn more but would also likely necessitate higher computational power, possibly using graphics processing units (GPUs) or a remote server over the cloud. Some ANNs have over 100 layers and millions of weights to optimize. The challenge is to build an ANN to solve a problem with a small number of operations, through optimizing architectural design.

Convolutional Neural Networks

Convolutional neural networks (CNNs) are a type of ANN where layers are structured in such a way that a convolutional kernel can be applied, which is important for image processing. A convolution is a common mathematic function, and a kernel refers to a matrix of weights that can be either preset or, more commonly, learned in the case of a CNN with access to training data. CNNs take a series of medical images, often single or multimodality, as input; perform

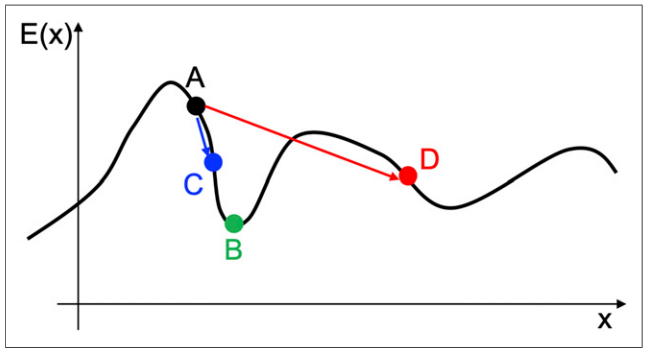

FIGURE 3. Starting from point A, we wish to find lowest point in function E(x), labeled B. Suppose we know slope of E(x) at point A; gradient-based search suggests we move right to lower point. Ideally, we prefer small step, to C, rather than large step, to D.

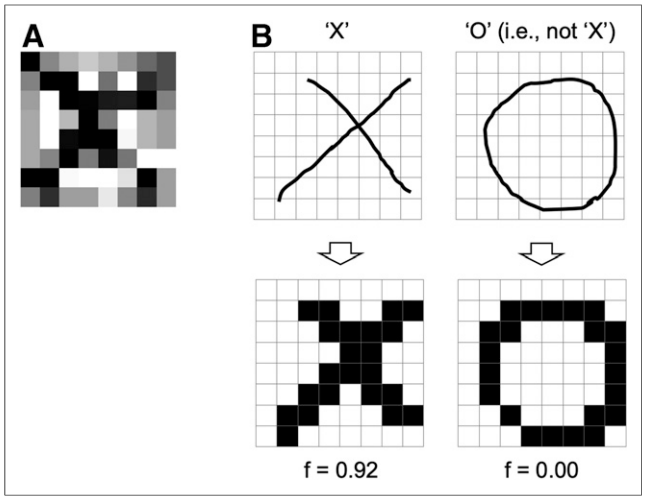

FIGURE 4. (A) Grayscale image of matrix of weights (kernel) after 24 training cases. Darker coloring represents pixels with higher weights. Notice that image resembles X. (B) New examples of images of handwritten letters input to trained ANN show that letter X is identified with 92% likelihood (ANN output f = 0.92) and O is not interpreted as X, with likelihood of 0% (f = 0.00).

operations; calculate weights and biases for different layers; and optimize parameters to minimize cost based on the desired output. Typically, input regions of interest or features are not required. Although simple ML algorithms can process images at the pixel level, CNNs have greater capacity for complex decision making and often outperform them—for example, in terms of classification accuracy.

Two common CNN architectural designs are illustrated in Figure 5. For applications such as image reconstruction or segmentation, where the desired output is an image, variations on an encoder–decoder architecture are commonly used (Fig. 5A). An encoder reduces input data in a stepwise process to identify components or features. This can be accomplished through the use of the concept of stride, defined in Table 1 and illustrated in Figure 6, or pooling, defined in Table 1. A decoder then builds the output image from the features using a stepwise process, possibly including interpolation or up-sampling to increase resolution. Some architectures that follow this style include U-Net (2-dimensional data) and V-Net (3-dimensional data). For applications such as disease detection where the output is a classification (e.g., disease present or absent), an ANN might only have an encoder phase where input data is reduced in a stepwise process that leads to the output classification (Fig. 5B). Res-Net is one such architecture.

Hardware Aspects

ANNs are typically programmed using software languages such as Matlab or Python. However, the hardware on which these programs run can significantly affect their speed. Simple ANNs can easily run on a standard laptop. However, more complex ANNs often need powerful hardware. GPUs have emerged as an effective hardware solution for ANNs since they are capable of performing many simple computations simultaneously, which improves speed. Sometimes, GPUs can be so powerful that they can perform all of the computations required for a convolution operation simultaneously. When ANNs get to be large enough that even a single GPU is insufficient, compute clusters (supercomputers often consisting of banks of GPUs) in data centers accessed over the cloud may be needed.

NEURAL NETWORKS IN NUCLEAR MEDICINE: A SPECTRUM OF APPLICATIONS

Complex ANNs are used across a spectrum of nuclear medicine applications. A search of "machine learning" on PubMed returned 595 papers in 2009, 2,402 in 2014, and 11,297 in 2019, several including the use of ANNs. ANNs can help with image reconstruction or to create standard-dose from low-dose images, as well as to improve scatter and attenuation correction (*9–18*). ANNs can also assist with disease detection and segmentation (*19–26*), disease diagnosis, and outcome predictions (*27–32*). In this paper we have chosen to focus on a few applications with specific examples.

Neural Networks Used for Image Reconstruction and Low-Dose PET

Signal noise is inherent in nuclear imaging and may be aggravated by using low-dose techniques or reducing image acquisition time. CNNs can be used during image reconstruction to generate higher-quality images than is possible with conventional techniques and to improve the perceived quality of noisy images. An array of architectural designs may be used (and details in the literature are often limited).

One approach, focusing on image reconstruction, is illustrated by Häggström et al. (*11*). The authors programmed a CNN using an encoder–decoder architecture, similar to that presented in Figure 5A, to reconstruct PET images from data synthesized almost entirely using a combination of phantom, simulation, and augmentation techniques. The input to the CNN was PET sinogram data represented by a $288 \times 289 \times 1$ matrix; the output was image data represented by a $128 \times 128 \times 1$ matrix. The encoder reduced the input data through sequential layers applying convolution kernels with decreasing kernel size and stride 2, as well as activation functions including batch normalization and ReLUs. The decoder upsampled the data using sequential layers to apply convolution kernels, increase matrix size, and apply activation functions including batch normalization and ReLUs to produce the final PET images. Several design modifications were studied, including differing numbers of layers and kernel size, among others. The CNN was able to generate PET images with higher quality than could techniques such as ordered-subset expectation maximization or filtered backprojection.

Often the CNN includes several layers with parallel paths (also referred to as parallel channels) to apply a host of specific kernels to dissect out certain image features and then combine feature information through a series of layers to generate the noiseless output image. Sometimes the input is a noisy image and the CNN is designed to reduce these input data to a series of low-resolution images that identify abstract features such as edges and texture and then progressively reconstruct a noiseless output image at the same resolution as the input image. These CNNs typically undergo supervised training using pairs of noisy input and noiseless output images. The key is to ascertain that no significant information is lost or false information added.

As an illustration Chen et al. (*12*) used a CNN with a U-Net architecture, similar to that presented in Figure 5A, to synthesize full-dose ^{18}F-florbetaben PET/MR images from low-dose images obtained using 1% of the raw list-mode PET data. The quality of the synthesized images was subjectively evaluated on a 5-point scale by 2 readers, whereas Bland–Altman plots were used to compare SUV ratios. The authors found that the synthesized images showed improved quality metrics compared with low-dose images, with

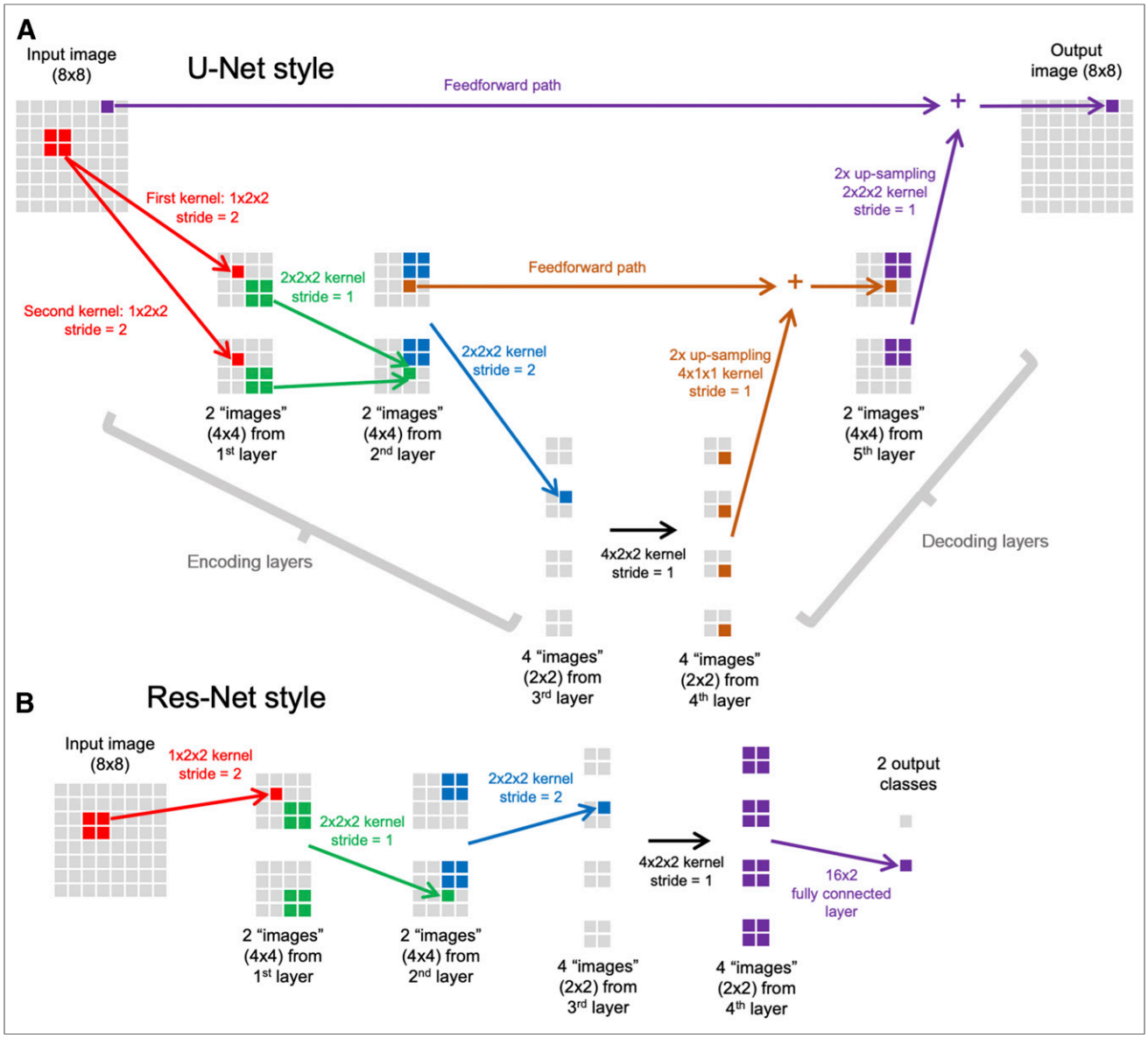

FIGURE 5. Illustration of ANN architectures. (A) Encoder–decoder design is helpful for image segmentation. In encoder, input image resolution is reduced while number of images increases. First layer produces 2 images, first by applying 2×2 kernel using convolutional operation with stride of 2 and then by applying a different kernel. Since 1 image is input, we denote these as $1 \times 2 \times 2$ kernels. Second layer produces 2 output images from 2 input images that are treated as volume, again using 2 different kernels, denoted as $2 \times 2 \times 2$, with stride of 1. Third layer applies 4 different $2 \times 2 \times 2$ kernels with stride of 2, to generate 4 images that are input to fourth layer. In decoder, up-sampling creates higher-resolution images so CNN input and output resolutions are similar. A feed-forward path adds data from earlier layers. U shape gives rise to name *U-Net.* (B) Encoder design is helpful for disease detection. Over consecutive layers, image resolution is decreased, to identify features that are encoded into feature maps. Final layer is often fully connected; the 2 outputs shown each use weighted sum of every pixel from preceding layer. Res-Nets are example of this.

high accuracy for amyloid status and intrareader reproducibility similar to that for full-dose images. A review of CNN approaches for handling low-dose PET was previously published (*33*).

Neural Networks Used for Disease Detection and Segmentation

A common application of neural networks is disease detection and segmentation, such as to quantify disease burden. A time-consuming task in practice, essentially, this is a pixelwise classification problem: each pixel must be tagged as normal or abnormal and joined to the region where it belongs (e.g., liver or spleen). Typically, the output is an image at the same resolution as the input image, with feature information extracted by the neural network used to create overlying segmentation images. Similar to denoising, input and output images are at the same resolution and training is usually supervised, using combinations of raw and segmented images. Several papers have been written on lesion detection and segmentation using neural networks (*20–26*) with differing architectural designs, although often a U-Net.

As an illustration, consider a paper by Zhao et al. (*26*). The authors created CNNs with the aim of automatically segmenting sites of disease on ^{68}Ga-PSMA-11 PET/CT images, to provide a yes–no answer as to whether a voxel reflected a lesion. The overall framework consisted of 2 components operating in series: 3 parallel

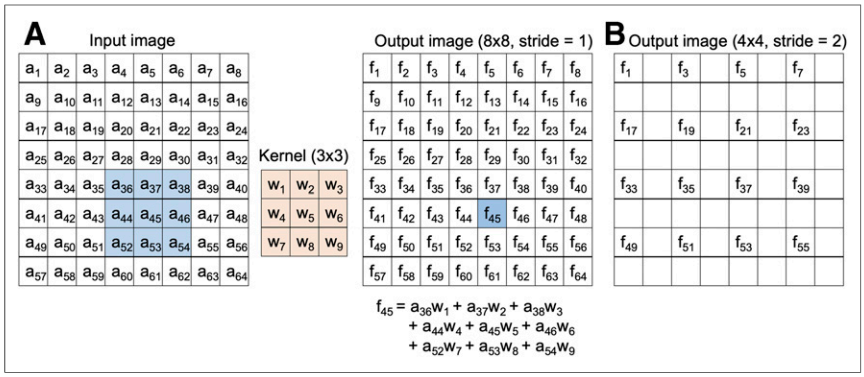

FIGURE 6. Illustration of stride. (A) Input 8 × 8 matrix is processed in convolutional layer with 3 × 3 kernel (weights w_1–w_9). Each pixel in output 8 × 8 matrix is calculated by multiplying the 9 neighbors nearest to corresponding input pixel by respective kernel weights. As illustration, calculation for output pixel f_{45} is shown. (B) Using stride of 2, every second output pixel is calculated in both dimensions, resulting in 4 × 4 output image.

CNN paths, each designed to detect lesions in 1 of 3 different planes, and per-voxel final majority voting based on intermediate decisions from each plane's CNN. The CNNs had a U-Net structure consisting of an encoding stack followed by a decoding stack that fused feature maps with original images, similar in structure to Figure 5A. The encoding stack included 3 × 3 convolutions, 2 × 2 maximum pooling with stride 2 for down-sampling, ReLU, and batch normalization. The decoding stack synthesized the information using a transposed convolution with kernel size 2 × 2 and stride 2, a concatenation operation, and 3 × 3 convolutions with ReLU and batch normalization. At the last layer of the CNN, the sigmoid function helped map features to a segmentation probability map. The Dice similarity coefficient was used to evaluate the accuracy of anatomic segmentation. ^{68}Ga-PSMA-11 PET/CT scans from 193 men with metastatic castration-resistant prostate cancer were randomly divided into 130 training scans and 63 testing scans. All lesions in the pelvis were manually delineated (i.e., 1,003 bone lesions and 626 lymph node lesions, among others). A 5-fold cross-validation was used for optimization. Using the manually annotated images as ground truth, a lesion was considered to be correctly detected when the overlap ratio exceeded a threshold of 10%. The detection accuracy, sensitivity, and F1 score (harmonic mean of accuracy and sensitivity) were 0.99, 0.99, and 0.99, respectively, for bone lesions and 0.94, 0.90, and 0.92, respectively, for lymph nodes. The image segmentation accuracy was lower than the lesion detection accuracy. The overall model achieved average Dice similarity coefficients of 65% and 54%, PPVs of 80% and 67%, and specificities of 61% and 55% for bone and lymph node lesions, respectively.

Although the possibility of using ANNs for lesion detection and image segmentation has enormous impact for clinical practice, manual assessment is still often used.

Neural Networks Used for Disease Diagnosis and Outcome Prediction

ANNs can assist with disease diagnosis and outcome prediction (*27–31*). Often, only a small set of input images or data is needed, and models that input full-resolution images or several data sources gradually reduce this to distill a diagnosis or outcome by the final layer. Typically, these are classification problems, training is supervised, and often a Res-Net architectural design is used, similar to that presented in Figure 5B. Although early results are promising, rigorous evidence supporting ML models is lacking. A systematic review by Nagendran at al. published last year found 1 randomized clinical trial related to breast ultrasound and 2 nonrandomized prospective studies investigating intracranial hemorrhage (*34*). The field is young, and it is important to remember to temper our claims of imminent clinical impact.

Mayerhoefer et al. provide an illustration of a neural network use for a predictive application (*32*). Specifically, the authors proposed to determine whether radiomic features on ^{18}F-FDG PET/CT alone or in combination with clinical, laboratory, and biologic parameters were predictive of 2-y progression-free survival in subjects with mantle cell lymphoma. A multilayer feed-forward neural network was used, which relied on a back-propagation learning algorithm (*8*) in combination with logistic regression analysis for feature selection. Few specific details are given, although we are told there was a minimum of 1 hidden layer with a minimum of 3 neurons per hidden layer. The input included a guess of weights for individual radiomic features, and the classification step was repeated 5 times. The data consisted of 107 ^{18}F-FDG PET/CT scans in treatment-naïve mantle cell lymphoma patients with baseline and follow-up data to the date of progression, death, or a minimum of 2 y. Cases were randomly split into 75 training and 32 validation cases for each classification step repetition. A semiautomatic process was used for lesion delineation, and several parameters were included for analysis: SUV_{max}, SUV_{mean}, SUV_{peak}, total lesion glycolysis, and 16 textural features derived from the gray-level cooccurrence matrix calculated in 3 dimensions. Outcome measures included the area under the receiver-operating-characteristic curve and classification accuracy. Although radiomic features were not significantly correlated with absolute progression-free survival (in months), 2-y progression-free

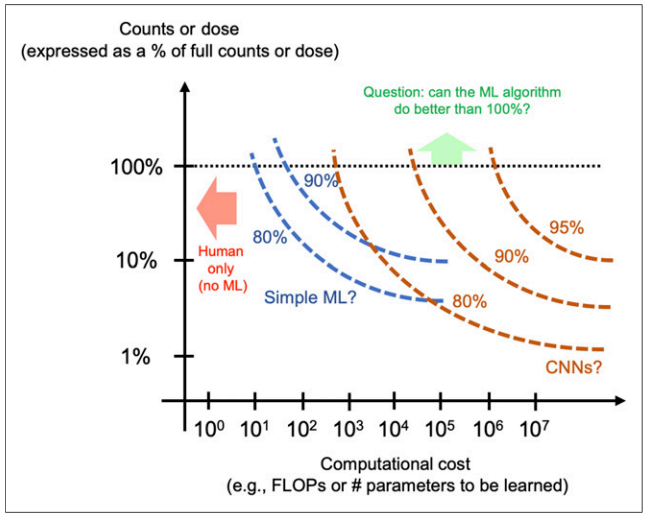

FIGURE 7. Conceptual graph showing how classification accuracy (dotted curves) and counts might be impacted by ML algorithm computational cost (and ability to learn complex tasks). Such graphs require that researchers provide specific details about their ML implementations. FLOP = floating-point operation.

TABLE 2
Suggested Checklist to Include for ML-Related Algorithm Reporting

Question	Possible metric	Comment
ML algorithm?	Family of ML algorithms	That is, CNN, random forest, support vector machine…
Architecture details?	Dependent on algorithm	That is, for CNN, report number of layers, kernel size, and strides and show complete block diagram with sufficient detail that model could be independently reconstructed
Computational cost?	Number of parameters, floating-point operations	Although a consulting computing expert, similarly to consulting statistician for clinical trials, is suggested, authors may generate this themselves
Data?	Training, validation, testing	That is, data type, number of validation/testing cases, use of cross-validation, data source (algorithms trained with data from single institution might not perform well using data from another institution)
Figure of merit?	Classification accuracy, dose reduction…	That is, key numeric performance results should be given, such as classification accuracy; ultimately, this should be standardized for a given application

survival status correlated with SUV_{mean} ($P = 0.022$) and entropy ($P = 0.034$) in a multivariate analysis. When SUV_{mean} and entropy values were input to the neural network, areas under the receiver-operating-characteristic curve for 2-y progression-free survival prediction were 0.70–0.73 (median, 0.72), with classification accuracies of 71.0%–76.7% (median, 74.4%) in training cases and 70.6%–86.8% (median, 74.3%) in validation cases, improving when combined with additional clinical, laboratory, or biologic data.

Common Themes and a Call to Arms

Ultimately, we arrive at a few conclusions regarding ANNs in nuclear medicine. The first is that good performance is often achieved with fewer than 10 layers. Many papers use data from small patient cohorts (~20–200) supplemented with data augmentation techniques to generate larger training or cross-validation datasets or generate data using simulation software. The second is that the computational cost of an ML algorithm is rarely reported yet should not be ignored, as it directly impacts reproducibility and clinical practicality. Those papers that do describe the algorithm structure often omit key information, making it nearly impossible for a reader to recreate the model. Floating-point operations, the cost metric commonly used by computer scientists and engineers, are rarely reported. The third is that there is a lack of well-conducted, systematic studies, with few to no randomized clinical trials evaluating applications in routine clinical practice.

We are in the early days of the application of ML to nuclear medicine, and it is becoming evident there is a need for the community to come together and design standard elements of reporting needed for the field's evolution. This would make it easier to assess algorithm effectiveness, cost, and appropriate use. If we had the details, we could graph metrics of input, algorithm complexity, and output to establish algorithms that are most effective for a specific task. As a starting point for discussion, Figure 7 portrays a conceptual graph that could be plotted if standardized details of algorithms were reported, and which would provide insight into trends. The graph uses the example of low-dose PET and plots percentage dose versus ML algorithm computational cost. Any paper that reports dose, computational cost, and algorithm family could be included as a point on the graph. As more data become available, we would see trend lines emerge, such as the dashed lines shown representing constant classification accuracy for algorithm family. Bounds on algorithm family capability might be inferred. For example, Minarik et al. (*10*) report performance results for a CNN at various image noise levels (analogous to percentage dose). Although the CNN performs well, the computational cost is not reported, and it difficult to exactly replicate what was done. With additional information, we could have plotted several points on our graph, on which to base future work.

To gather insight into algorithms best suited for a given task, and the computational cost needed to achieve a desired output, we advocate that our community use a checklist for reporting ML algorithms. Table 2 provides our top 5 points to include. We hope this represents a start for further discussion.

CONCLUSION

We are witnessing a potentially phenomenal development in clinical nuclear medicine. Although ANNs are becoming ever more common in nuclear medicine, new families of algorithms are being developed. Further, as databases of shared images continue to be created, there will be expanding datasets useful for training, validation, and testing purposes. Several issues remain, notably those surrounding ethics and privacy of data collection, deidentification, and ownership. In some situations, it may prove easier to download an algorithm to multiple sites instead of uploading multisite data to a communal database. Regardless, to understand where we are, a standardized practice for reporting ML algorithm metrics would be helpful. We present a list of our top 5 items to include (Table 2) and suggest how data could be compiled to generate graphs showing which family of ML algorithms might be best suited for a given application. We hope this paper has provided insight into how ANNs work, the spectrum of clinical tasks they can help with, and where we might go from here.

REFERENCES

1. Uribe CF, Mathotaarachchi S, Gaudet V, et al. Machine learning in nuclear medicine: part 1—introduction. *J Nucl Med.* 2019;60:451–458.
2. Zukotynski K, Gaudet V, Kuo PH, et al. The use of random forests to classify amyloid brain PET. *Clin Nucl Med.* 2019;44:784–788.
3. Nuvoli S, Spanu A, Fravolini ML, et al. ^{123}metaiodobenzylguanidine (MIBG) cardiac scintigraphy and automated classification techniques in Parkinsonian disorders. *Mol Imaging Biol.* 2020;22:703–710.
4. Perk T, Bradshaw T, Chen S, et al. Automated classification of benign and malignant lesions in ^{18}F-NaF PET/CT images using machine learning. *Phys Med Biol.* 2018;63:225019.
5. Nicastro N, Wegrzyk J, Preti MG, et al. Classification of degenerative parkinsonism subtypes by support-vector-machine analysis and striatal ^{123}I-FP-CIT indices. *J Neurol.* 2019;266:1771–1781.
6. Kim JP, Kim J, Kim Y, et al. Staging and quantification of florbetaben PET images using machine learning: impact of predicted regional cortical tracer uptake and amyloid stage on clinical outcomes. *Eur J Nucl Med Mol Imaging.* 2020;47:1971–1983.
7. Mayerhoefer ME, Materka A, Langs G, et al. Introduction to radiomics. *J Nucl Med.* 2020;61:488–495.
8. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature.* 2015;521:436–444.
9. Xiang L, Qiao Y, Nie D, An L, Wang Q, Shen D. Deep auto-context convolutional neural networks for standard-dose PET image estimation from low-dose PET/MRI. *Neurocomputing.* 2017;267:406–416.
10. Minarik D, Enqvist O, Trägårdh E. Denoising of scintillation camera images using a deep convolutional neural network: a Monte Carlo simulation approach. *J Nucl Med.* 2020;61:298–303.
11. Häggström I, Schmidtlein CR, Campanella G, Fuch TJ. DeepPET: a deep encoder-decoder network for directly solving the PET image reconstruction inverse problem. *Med Image Anal.* 2019;54:253–262.
12. Chen KT, Gong E, de Carvalho Macruz FB, et al. Ultra-low-dose ^{18}F-florbetaben amyloid PET imaging using deep learning with multi-contrast MRI inputs. *Radiology.* 2019;290:649–656.
13. Gao F, Shah V, Sibille L, Zuehlsdorff S. An AI system to determine reconstruction parameters and improve PET image quality [abstract]. *J Nucl Med.* 2018;59(suppl 1):31.
14. Hwang D, Kim KY, Kang SK, et al. Improving the accuracy of simultaneously reconstructed activity and attenuation maps using deep learning. *J Nucl Med.* 2018;59:1624–1629.
15. Leynes AP, Yang J, Wiesinger F, et al. Zero-echo-time and Dixon deep pseudo-CT (ZeDD CT): direct generation of pseudo-CT images for pelvic PET/MRI attenuation correction using deep convolutional neural networks with multiparametric MRI. *J Nucl Med.* 2018;59:852–858.
16. Spuhler KD, Gardus J III, Gao Y, DeLorenzo C, Parsey R, Huang C. Synthesis of patient-specific transmission data for PET attenuation correction for PET/MR neuroimaging using a convolutional neural network. *J Nucl Med.* 2019;60:555–560.
17. Torrado-Carvajal A, Vera-Olmos J, Izquierdo-Garcia D, et al. Dixon-VIBE deep learning (DIVIDE) pseudo-CT synthesis for pelvis PET/MR attenuation correction. *J Nucl Med.* 2019;60:429–435.
18. Gong K, Guan J, Kim K, et al. Iterative PET image reconstruction using convolutional neural network representation. *IEEE Trans Med Imaging.* 2019;38:675–685.
19. Lindgren Belal S, Sadik M, Kaboteh R, et al. Deep learning for segmentation of 49 selected bones in CT scans: first step in automated PET/CT-based 3D quantification of skeletal masses. *Eur J Radiol.* 2019;113:89–95.
20. Gsaxner C, Roth PM, Wallner J, Egger J. Exploit fully automatic low-level segmented PET data for training high-level deep learning algorithms for the corresponding CT data. *PLoS One.* 2019;14:e0212550.
21. Huang B, Chen Z, Wu PM, et al. Fully automated delineation of gross tumor volume for head and neck cancer on PET-CT using deep learning: a dual-center study. *Contrast Media Mol Imaging.* 2018;2018:8923028.
22. Zhao X, Li L, Lu W, Tan S. Tumor co-segmentation in PET/CT using multimodality fully convolutional neural network. *Phys Med Biol.* 2018;64:015011.
23. Hatt M, Laurent B, Ouahabi A, et al. The first MICCAI challenge on PET tumor segmentation. *Med Image Anal.* 2018;44:177–195.
24. Bi L, Kim J, Kumar A, Wen L, Feng D, Fulham M. Automatic detection and classification of regions of FDG uptake in whole-body PET-CT lymphoma studies. *Comput Med Imaging Graph.* 2017;60:3–10.
25. Xu L, Tetteh G, Lipkova J, et al. Automated whole-body bone lesion detection for multiple myeloma on ^{68}Ga-pentixafor PET/CT imaging using deep learning methods. *Contrast Media Mol Imaging.* 2018;2018:2391925.
26. Zhao Y, Gafita A, Vollnberg B, et al. Deep neural network for automatic characterization of lesions on ^{68}Ga-PSMA-11 PET/CT. *Eur J Nucl Med Mol Imaging.* 2020;47:603–613.
27. Commandeur F, Goeller M, Razpour A, et al. Fully automated CT quantification of epicardial adipose tissue by deep learning: a multicenter study. *Radiol Artif Intell.* 2019;1:e190045.
28. Eisenberg E, Commandeur F, Chen X, et al. Deep learning-based quantification of epicardial adipose tissue volume and attenuation predicts major adverse cardiovascular events in asymptomatic subjects. *Circ Cardiovasc Imaging.* 2020;13:e009829.
29. Hartenstein A, Lubbe F, Baur ADJ, et al. Prostate cancer nodal staging: using deep learning to predict ^{68}Ga-PSMA-positivity from CT imaging alone. *Sci Rep.* 2020;10:3398.
30. van Velzen SGM, Lessmann N, Velthuis BK, et al. Deep learning for automatic calcium scoring in CT: validation using multiple cardiac CT and chest CT protocols. *Radiology.* 2020;295:66–79.
31. Huang Y, Xu J, Zhou Y, et al. Diagnosis of Alzheimer's disease via multimodality 3D convolutional neural network. *Front Neurosci.* 2019;13:509–520.
32. Mayerhoefer ME, Riedl CC, Kumar A, et al. Radiomic features of glucose metabolism enable prediction of outcome in mantle cell lymphoma. *Eur J Nucl Med Mol Imaging.* 2019;46:2760–2769.
33. Zaharchuk G. Next generation research applications for hybrid PET/MR and PET/CT imaging using deep learning. *Eur J Nucl Med Mol Imaging.* 2019;46:2700–2707.
34. Nagendran M, Chen Y, Lovejoy CA, et al. Artificial intelligence versus clinicians: a systematic review of design, reporting standards, and claims of deep learning studies. *BMJ.* 2020;368:m689.