Deep Learning and Multimodal Artificial Intelligence in Orthopaedic Surgery

Anthony Bozzo, MD, MSc, FRCSC
James M. G. Tsui, MD, PhD, FRCPC
Sahir Bhatnagar, PhD
Jonathan Forsberg, MD, PhD

ABSTRACT

This review article focuses on the applications of deep learning with neural networks and multimodal neural networks in the orthopaedic domain. By providing practical examples of how artificial intelligence (AI) is being applied successfully in orthopaedic surgery, particularly in the realm of imaging data sets and the integration of clinical data, this study aims to provide orthopaedic surgeons with the necessary tools to not only evaluate existing literature but also to consider AI's potential in their own clinical or research pursuits. We first review standard deep neural networks which can analyze numerical clinical variables, then describe convolutional neural networks which can analyze image data, and then introduce multimodal AI models which analyze various types of different data. Then, we contrast these deep learning techniques with related but more limited techniques such as radiomics, describe how to interpret deep learning studies, and how to initiate such studies at your institution. Ultimately, by empowering orthopaedic surgeons with the knowledge and know-how of deep learning, this review aspires to facilitate the translation of research into clinical practice, thereby enhancing the efficacy and precision of real-world orthopaedic care for patients.

From the Division of Orthopaedic Surgery, McGill University, Canada (Bozzo), the Division of Radiation Oncology, McGill University, Canada (Tsui), the Department of Epidemiology and Biostatistics, Department of Diagnostic Radiology, McGill University, Canada (Bhatnagar), and the Memorial Sloan Kettering Cancer Center (Forsberg).

None of the following authors or any immediate family member has received anything of value from or has stock or stock options held in a commercial company or institution related directly or indirectly to the subject of this article: Bozzo, Tsui, Bhatnagar, and Forsberg.

J Am Acad Orthop Surg 2024;32:e523-e532

DOI: 10.5435/JAAOS-D-23-00831

Artificial intelligence (AI) and machine learning (ML) methods applied to orthopaedic research have produced a substantial body of literature with hundreds of papers published in 2022 alone. While these studies highlight the prospects of this exciting new technology, delving properly into the methods demands a certain level of knowledge that may be outside of a clinician's expertise. As such, it is pertinent for orthopaedic surgeons to familiarize themselves with AI methodology and the associated terminology, enabling them to not only be able to critically appraise emerging research but also potentially conduct such studies themselves.

A comprehensive review by Shah et al[1] summarized applications of classic ML techniques and terminology in orthopaedics. However, there has been a notable change in the AI landscape in the past decade, with deep learning using deep neural networks (DNNs) surpassing classic ML by demonstrating a greater capacity for complex pattern recognition.[2,3] Deep learning on

medical image data such as radiographs, CTs, or MRIs enables the automatic extraction and selection of the most salient features, obviating the need to generate handcrafted or human-engineered features such as those used in radiomics. Deep learning provides a richer representation of the imaging data that can then be combined with data from other modalities (image data, genomic data, clinical variables). The use of multimodal AI (MMAI) is among the most promising use cases for AI.[4]

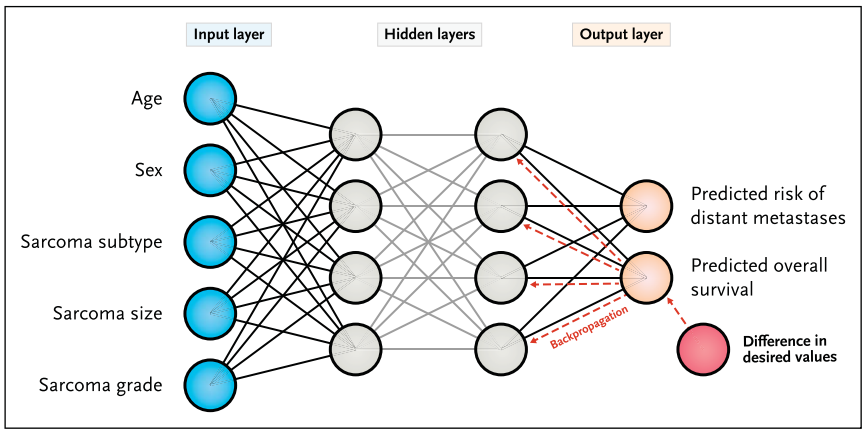
Therefore, this review article focuses on the applications of deep learning with neural networks and multimodal neural networks in the orthopaedic domain. By providing practical examples of how AI is being applied successfully in orthopaedic surgery, particularly in the realm of imaging data sets and the integration of clinical data, this paper aims to provide orthopaedic surgeons with the necessary tools to not only evaluate existing literature but also to consider AI's potential in their own clinical or research pursuits.

In this review, we will examine common AI terms, real-world case studies, and explore the diverse applications of deep learning in orthopaedic practice. We begin with standard DNNs which can analyze traditional numerical clinical variables, then describe convolutional neural networks (CNNs) which can analyze image data, and then introduce MMAI models which analyze various types of different data. Then, we contrast these deep learning techniques with related but more limited techniques such as radiomics, describe how to interpret deep learning studies, and how to initiate such studies at your institution. Ultimately, by empowering orthopaedic surgeons with the knowledge and know-how of deep learning, this review aspires to facilitate the translation of research into clinical practice, thereby enhancing the efficacy and precision of real-world orthopaedic care for patients.

What Is a Deep Neural Network, and How Do They Learn?

Deep learning involves a network of layers of connected artificial neurons (Figure 1). The first layer accepts input variables which are termed "features." If a model has 10 input variables, there would be 10 neurons in the first layer. Similar to natural neurons, each artificial neuron emits an output to neurons in the next layer, and the strength of the connection between two neurons (the weights) can be adjusted. DNNs are neural networks with multiple "hidden" layers between the first input layer and the output layer. The final layer in a neural network converges on output neurons. The number of output neurons will depend on the specific task the model aims to accomplish. For instance, in a binary classification task where the model predicts the presence or absence of an event (will an infection occur?), only one output neuron is needed. In multiclass classification task, such as classifying a femoral neck fracture according to the Garden classification system, the number of output neurons would correspond to the total number of classes (4), with each output neuron representing one class. The neuron with the highest value among the outputs corresponds to the chosen class. When the neural network's predictions are wrong, an error function, much like a "feedback loop," is back

Figure 1



A DNN which accepts 5 input variables, has two hidden layers, and predicts two outcomes is depicted. When the predictions are wrong, back propagation is used to adjust the weights between neurons in each layer. This process repeats iteratively until errors are minimized. DNNs = deep neural networks

propagated through the network which adjusts the weights between neurons and thus the network's outputs (predictions) are adjusted. The network determines the appropriate direction for the weight adjustments by following the "gradient descent," a process using calculus aimed at reducing errors. The network will keep nudging the weights in the correct direction until errors are minimized. This process repeats iteratively over the training set of data until the model's predictions are as accurate as possible. This is how DNNs learn. Some but not all classic ML algorithms also use gradient descent, but much fewer weights or parameters are tuned (dozens in classic ML versus up to billions in deep learning). As deep learning is a small subset of ML algorithms, these terms should not be used interchangeably.

Anastasio et al[5] used a DNN to identify the best combination of biologic agents to aid in bone repair. They extracted data from 225 published studies. Their DNN has 16 input neurons, for the 16 different combinations of agents such as bone-morphogenic-protein-2 (BMP2), BMP7 among others, and different doses, reported in their literature search. Their DNN has 26 output neurons with each output neuron predicting a different aspect of bone healing. In their case, outputs ranged from "bone healing at 3 months" to "mean time to radiographic union," "need for repeat bone grafting," and others. They averaged the predicted performance on each of these outcomes. Their DNN model predicts that using a combination of three factors (BMP2, BMP7, and osteogenin) was more effective than using any single factor, given this combination of inputs had the highest average predicted performance across all 26 outputs.[5]

This is an example of "supervised learning," where the desired outputs or classifications, known as "labels," are known. When large data sets are available but no appropriate groupings or labels are present, "unsupervised learning" can be used, often to reduce the dimension of input or to cluster the inputs together, for example, to group patients with similar features.

What Is "Learned" by Deep Neural Networks?

Underlying how neural networks "learn" is the fundamental concept that any causal relationship between inputs, whether traditional clinical variables, image data, or genomic data, and outputs such as predictions or classifications, can be expressed through a mathematical function.[6] The objective of deep learning algorithms is to discover the mathematical function that best describes the relationships between inputs and outputs. In other words, deep learning uncovers the best mapping or pattern recognition between the inputs available and the predictions or classifications desired in a given study.

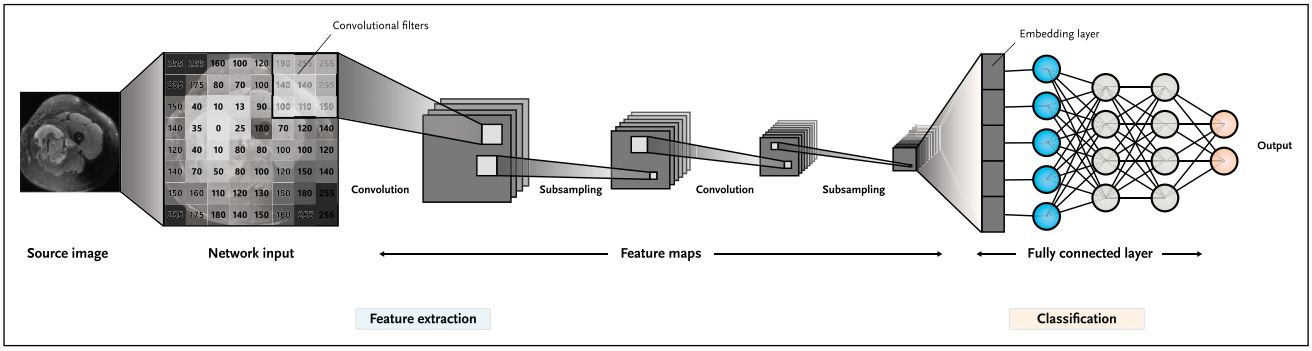
To accomplish this, deep learning algorithms must be trained on high-quality data. The available variables or "features" should accurately reflect and represent the clinical problem. For example, if it is unknown whether a patient is diabetic or not, an algorithm will be missing a key piece of information when predicting whether a prosthetic joint infection (PJI) will occur. Second, the outcomes of interest or "ground truth labels" must be accurately defined. If patients who never developed a PJI are mistakenly labelled in the data set as having a PJI, the algorithm will learn the wrong patterns in the data and will make more errors when tested on new patients. For instance, ImageNet is an open-source data set of images used to train image models.[7] Even state-of-the-art models cannot demonstrate 100% performance on this data set because many of the images are incorrectly labelled.[8] However, if proper input features are available to train on and the labels are accurate, with sufficient data, deep learning algorithms will use gradient descent to find the optimal combination of weights between neurons that best ties input features to the correct outputs. Note, however, that the final model will also reflect any biases in the data used for training.[9]

How Are Images Analyzed by Deep Neural Networks?

When humans interpret the world around us or images such as a radiograph or MRI, the initial processing of visual information within the visual cortex begins with V1. This area is responsible for processing low-level features such as lines and edges. The outputs of V1 are then sent to subsequent areas that would assemble these low-level features into more complex features such as blobs and color (V2), shapes (V3, V4), movements (V5), and even objects (inferotemporal cortex) and persons (fusiform face area).[10]

For a neural network to interpret an image, it must first be represented by a grid of numbers, where each number represents the pixel intensity in that location (Figure 2). For CT scans, each pixel is denoted by a numeric value known as the Hounsfield unit. As opposed to DNNs which accept a set of clinical variables as their input, CNNs are a type of DNN designed to accept a grid of numbers representing images as input.

Figure 2



A convolutional neural network (CNN) is a neural network designed to accept images as input. The image is represented as a grid of numbers representing the information at each pixel. Initial layers of the CNN extract low-level features such as lines and edges, while higher order features are extracted in deeper layers. The core information of an image is represented in the embedding layer. This information can then be used for the specific research question. In this case, a DNN is used to relate the embedding layer to the outcomes of interest (outputs). CNNs = convolutional neural networks, DNNs = deep neural networks

They can accept 2D images or 3D volumes of varying sizes. Their name, CNN, refers to the convolution operation. This mathematical operation allows low-level features such as lines or edges to be extracted from images. Subsequent layers of a CNN will combine these low-level features into higher-order features, similarly to the mammalian visual system. AlexNet is the first CNN that won the ImageNet competition in 2012 and demonstrated the superiority of deep learning to radiomics and other methods for image analysis.[11] While AlexNet contained 12 layers, the CNNs that followed have advanced markedly, with many of the more modern algorithms such as DenseNet-121[12] having more than 100 layers and other novel elements such as combining multiple levels of extracted features at different layers.

These CNNs extract low-level features (lines, edges) and combine them spatially to extract higher-order features. For example, a 256 × 256 radiograph or 256 × 256 × 30 (30 slices) MRI or CT is converted to a smaller set of numbers that represent what the algorithm considers to be the essential information within the image. This is known as the "embedding" layer. This downsampling reduces the spatial dimensions while retaining the most salient information, similar to how MP3 files zipped (compressed) to a smaller file size still retain their core information. Just as zipped files can be unzipped, once an image is downsampled to an "embedding" layer, the original image can then be recreated. Moreover, the core features encoded in this "embedding" layer can be used to convert, for example, a CT scan into an MRI.[13]

For our purposes, this embedding layer contains the most relevant features from the input data tailored to the task at hand. These features are identified using the back propagation method of learning described above. This process of downsampling and back propagation allows the CNN to learn intricate and abstract patterns from raw medical image data and excel in handling complex tasks with remarkable performance, whether the task is image classification, object detection, segmentation, or others.

In fact, most of the published orthopaedic literature using ML uses CNNs to interpret radiographs. The number of orthopaedic studies using CNNs are too numerous to individually cite, but published uses include identifying hip fractures[14] and distal radius fractures,[15] classifying knee osteoarthritis based on MRI,[16] and predicting bone age based on radiographs.[17] Several groups have trained CNNs to identify hip arthroplasty implants based on radiographs.[18,19]

Borjali demonstrate 100% accuracy in recognizing whether an implant is an Accolade II (Stryker), Coral (DePuy Synthes), or S-ROM (DePuy Synthes) with a sample size of only 252 patients.[19] They used an 80/10/10 train/validation/test partition with their data, meaning that a model was trained on 80% of the cohort. Ten percent of the cohort, the "validation set," was used to guide training and to tune the "hyperparameters"—the modifiable aspects of a neural network aside from the weights, such as learning rate, number of neurons in each layer, and the activation and loss functions used. Once the model's weights and hyperparameters were optimized, the performance was then evaluated on the remaining 10% of patients in the "test" set—which are patients who were never encountered during training. Both 80/10/10 and 70/15/15 splits are commonly used in training image models. Ideally, the test data would be taken from a different institution to better assess if the

Table 1. Common Metrics in Deep Learning Research

Metric	Explanation
F1 score	The harmonic mean of sensitivity and positive predictive value. Better than raw accuracy for imbalanced classes
Area under the receiver operator curve (AUC/AUROC)	Graphical depiction of model performance that plots true positive against false-positive rates at different classification thresholds. It quantifies the overall performance of a classification task and represent the ability of a model to distinguish between classes at various threshold values
C-Index	Extension of AUC to survival curves. Represents the proportion of patients that can be ordered such that patients with higher predicted survival actually survive longer
Brier score	Represents the mean squared difference between the model's predictions and the actual outcomes. Lower Brier scores are favorable
Calibration plot	Depicts differences between predicted and actual outcomes at different ranges of the predicted values. Helps to determine whether the model performs equally well across the entire range of possible outcomes or eg, better in the middle range of usual outcomes (where there may be more training data) and poorer at higher and lower ends of the outcome

model generalizes well; this is known as external validation.

To help their model learn and achieve great performance on their classification task, Borjali et al used an important process known as "data augmentation" to increase the amount of training data. Some examples of data augmentation are rotating, zooming in or out, or flipping the images along an axis. This step mimics the variety of what can be encountered in real life and allows the CNN to be invariant to relative positioning or orientation of the object within the image. An "epoch" is when the model has completed one training iteration over each radiograph in the training set, and usually several hundred epochs are required in deep learning. Borjali et al trained their network for 350 epochs, and with transformations applied at each epoch to the radiographs, they state that more than 69,000 radiographs were used to train the model.

The number of epochs required is often guided by when performance ceases to improve on the validation data set. It is important to note that more epochs do not always translate to better performance. It is possible for a model to overtrain for too many epochs and "overfit" to the training data: This is demonstrated by better performance on the validation set and worsening performance on the final test set. This means the model is memorizing its training data, without learning associations that generalize to unseen data. Therefore, once performance on the validation set plateaus, training can be stopped.

How Are Models Evaluated?

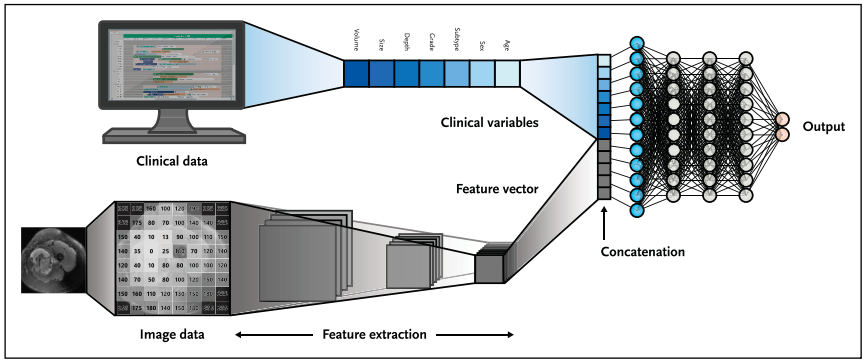
Borjali et al report overall accuracy as their outcome metric. Given their 100% accuracy, this would be reflected in other performance metrics as well. However, is raw accuracy a good measure of a model's performance?

When prediction models are tested on imbalanced classes, that is, outcomes are not split nearly 50/50, then raw accuracy is not a good metric. For example, if a PJI occurs in only ~2% of an elective total knee arthroplasty cohort, an algorithm can be "98% accurate" by predicting no PJIs at all, without having learned any insight into the problem. For unbalanced classes, metrics which combine sensitivity and specificity, such as the F1 score, are preferred. Models predicting time-based outcomes such as implant survival or patient survival have their own metrics. Please see Table 1 for list of common metrics used in deep learning research.

What Is Multimodal Artificial Intelligence?

Thus far, the deep learning models described have been trained on and evaluate only clinical variables (DNNs) or only image data (CNNs). However, relying solely on information derived from one data type may not provide a comprehensive understanding of a patient's condition.[20] In addition to clinical variables and image data, genomic information or data from histopathology can also play a notable role in specific tasks. Integrating

Figure 3



To achieve multimodal AI, it is essential to represent information from different domains in consistent numerical formats. CNNs extract features from images. Initially as a 2D or 3D grid of numbers, the image data is then represented by a 1D vector of numbers (the embedding layer in Figure 2). Similarly, clinical variables information can be transformed into a numerical vector. These numerical representations from various sources can then be concatenated or serially combined. The resulting sequence of numbers representing the core feature of multiple modalities can then be evaluated by a final DNN and related to the outcome of interest. CNNs = convolutional neural networks, DNNs = deep neural networks

diverse information from multiple modalities into a unified task is the essence of MMAI.

To achieve MMAI, it is essential to represent all information from different domains as numerical formats. As discussed earlier, CNNs can extract features from diagnostic imaging, essentially generating a vector of numbers (the embedding layer in Figure 2). Similarly, clinical and genetic information can be transformed into numerical vectors, and histopathology slides can also be encoded into numerical representations using CNNs. These numerical representations from various sources can then be concatenated or serially combined. The resulting sequence of numbers representing the core feature of multiple modalities can then be evaluated by a final DNN (or a classic ML algorithm) and related to the outcome of interest (Figure 3).

Recently, a success story in the field of MMAI was published in Nature in 2022.[21] The paper focuses on prostate cancer. The MMAI model was trained and validated using clinical and histopathological data from five phase III randomized trials. In this study, a simple CNN (ResNet50) was employed to extract deep features from digital pathology slides. These extracted deep features were then combined with essential clinical variables, including age, prostate-specific antigen (PSA), and Gleason score. The resulting concatenated feature vector was then used to generate an AI score that predicts oncological outcomes better than previous prediction models in their field.[21]

After the publication, the National Comprehensive Cancer Network recommended the use of this AI algorithm as a prognostic tool, providing Level 1 evidence for its validation. However, it is worth noting that the prostate cancer study had access to prospectively collected trial data and histopathology data from 5,654 patients, which contributes to the robustness of the findings. Given that orthopaedic cancers, sarcomas, are a relatively rarer and heterogeneous disease entity, finding a sufficient number of cases for MMAI may remain challenging. However, recent work indicates that smaller data sets that contain multimodal data can outperform larger data sets with only one data modality.[4] Within the orthopaedic literature, a multimodal deep learning model has been used to evaluate surgical knot tying ability.[22] The model combined analysis of both images of the final knot and kinematic time series data from a sensor to score "overall performance," "respect for tissue," and "time and motion."[22]

Comparison With Radiomics

Radiomics is a method to extract features from images which is not based on deep learning. The key difference is that unlike CNNs, the radiomics features are "human engineered"—they follow predetermined rules for combining the pixel information. These features can include measures of contrast, brightness, heterogeneity, texture, and others. While CNNs can learn an embedding of an image that captures its core information, radiomics can only determine which of the preset features, and combinations of those preset features, are related the outcome. It cannot learn new features, and the data set may be better represented by features not included in the radiomics package.

While radiomics has demonstrated the ability to predict the risk of metastasis of patients with sarcoma based on positron emission tomography (PET) images[23] and to differentiate between more aggressive and less aggressive osteosarcomas based on MRI images,[24] critical concerns of radiomics include feature stability, reproducibility, and repeatability.[25] Different radiomics extraction packages may yield varied results, as demonstrated in a study by Carloni et al.[26] Furthermore, when radiomics is compared with CNNs on the same data set, radiomics features commonly prove less accurate and perform poorer on external validation.[27-29] Radiomics is certainly capable of extracting features from images and is useful because of its easier implementation and lower computational requirements compared with CNNs. However, given their higher accuracy and robustness to external validation, CNNs represent the benchmark for extracting features from images and analyzing them in unimodal or multimodal AI models.

How to Critically Appraise Papers Using These Methods

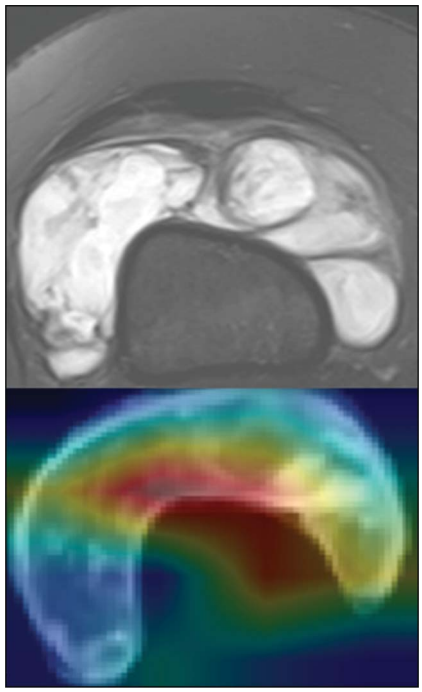
Several studies in orthopaedics and other specialties exist where ML is touted in the title, but linear regression demonstrates the best performance.[30] Readers must be able to separate hype from true advances. When evaluating a paper which uses ML algorithm, one should first identify what type of algorithms was used: classic ML or DNNs? Were one or multiple data modalities incorporated?

What is the sample size? Was data augmentation used? Did they make use of transfer learning? Transfer learning involves using knowledge gained from a model pretrained on a related task and then retraining it on a target task to enhance performance.

What performance metrics are reported? As mentioned above, reporting just overall accuracy of the model is not representative of an algorithm's ability to learn when classes are unbalanced. Was performance reported only on the training data set or validation data set? Was there an external testing data set?

What aspects aiding model interpretability have been reported by the authors? It is not enough to simply provide a prediction. In order to peel back the 'black box' of machine learning and trust the predictions, clinicians should know why the model is making a prediction. Some classic ML algorithms like random forests allow the relative importance of predictor variables to be visualized. To determine this, the algorithm removes a predictor variable and records the resulting decrease in performance. Variables are then ranked by their contribution to overall accuracy of the model.

Figure 4



A, An axial MRI slice depicting a soft-tissue sarcoma used to train a multimodal neural network. **B**, A corresponding heat map from the same patient is displayed. Pixels shaded in red are those that the network deemed most salient for its predictions.

For models analyzing images, attention maps or heat maps can be used to determine which pixels are most contributing to the model's predictions. For example, Navarro et al[31] used a unimodal CNN (DenseNet 121) to analyze individual slices of the MRI of a sarcoma and predict if it is high grade or low grade. They demonstrate heat maps for examples where the model was correct (heat map corresponds to areas of the sarcoma) and instances where the model was wrong (heat map is overlayed onto empty pixels). By knowing which pixels a CNN is paying attention to, physicians can understand what aspects of an image are deemed most salient or if spurious associations were made (Figure 4).

Reporting guidelines such as Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD-AI), Checklist for Artificial Intelligence in Medical Imaging, and other risk of bias tools have been developed for AI models and continue to be refined.[32,33] These checklists can be applied to the AI studies being reviewed and should be used when designing a study using deep learning methods. We provide a checklist to aid in critically appraising studies using ML (Table 2).

Table 2. Checklist of Discussion Items When Reviewing and Critically Appraising a Machine Learning Article

Area	Questions
Algorithms	Was classic machine learning or deep learning used? Was multimodal AI used? Was the network architecture explained?
Data	What is the sample size? Was data augmentation used? Did they make use of transfer learning? Do they specify how their data were preprocessed? If it is supervised learning, how were the labels obtained? Are the output labels (classes) well balanced, and if not, what measures were taken to correct this? Is the data single center or multicenter? Retrospective or prospective?
Model training	What split in the data was used to train and validate the model? Was external validation with a test set performed?
Results	What outcome measure is reported? Do the authors use performance metrics beyond simple overall accuracy? Are Brier scores or calibration plots provided? Does a benchmark for comparison exist, and if so, was it compared? Is the improvement clinically significant? Is an increase in net benefit to the patient demonstrated?
Model interpretability	For models based on clinical variables, was variable importance reported? For models based on image data, were heat maps provided?
Reporting guidelines	Were TRIPOD-AI, CLAIM, and/or other risk of bias tools used?

AI = artificial intelligence, CLAIM = Checklist for Artificial Intelligence in Medical Imaging

What Data and Expertise Are Needed for Your Center to Do Machine Learning Research?

Analysis of clinical variables alone does not require any special implant or infrastructure. Analysis can be performed in several coding languages such as R and Python. The training of CNNs on image data requires dedicated graphics processing units (GPUs) and may take days to complete training depending on the complexity of the model. However, several options for researchers are now available. For example, Google's Colab provides free access to a Python environment and GPUs through their cloud service. Access to GPUs is also available from Amazon Web Services and other providers.

Successful MMAI studies require collaboration between orthopaedic surgeons and data scientists to optimize the data collection and network architectures required for different clinical questions.

How should your image data be prepared for neural network analysis? "Preprocessing" is the term used to describe preparation of the image data set before analysis. After preprocessing, pixel data are standardized between images in the data set so that true differences between the images can be extracted. Common preprocessing techniques include N4 bias correction and Z-score normalization.[34]

To reduce computational costs and the time needed to train a CNN, sometimes only a region of interest or volume of interest within the larger MRI is analyzed. This is achieved with the use of segmentation masks which indicate which pixels should be included in the analysis. Segmentation can be performed manually or automatically by trained neural networks.[35]

What to Expect Next

Given the emerging popularity of ML in orthopaedic research, it is important that systematic reviews be performed. One of the first such reviews demonstrated that most papers fail to report the standardized checklists.[36] It is imperative that the orthopaedic community embrace and require checklists such as TRIPOD-AI, just as we have for reporting the results of other types of clinical research.

Furthermore, deep learning algorithms will continue to advance at a rapid pace, and we will likely witness the integration of deep learning solutions in patient care pathways. The rate limiting step for impactful deep learning studies is the availability of high-quality data. Current challenges to AI studies include the fact that data from different modalities can exist in disparate "silos" within an institution, that is, the patient's

laboratory values are not stored with their imaging data nor with their clinical notes. Ultimately, large multicenter databases of high-quality prospective data are needed to maximize deep learning to benefit our patients. This is especially true for rare conditions with low prevalence, such as sarcomas. Fortunately, federated learning, a privacy preserving method of data analysis, can be used to transcend historical and emerging barriers of data sharing between institutions. Similar to a traveling fellowship, an algorithm in the cloud can be sent to train on data from multiple centers, without the centers ever having to send data outside of their institution.[37]

Acknowledgments

The authors would like to thank the amazing Jiawen Deng, affiliated with the Temerty Faculty of Medicine at the University of Toronto, for the production of Figures 1–3. Anthony Bozzo was supported by the Cedars Cancer Foundation, the Montreal General Hospital Foundation, and the Richard H. Tomlinson Doctoral Fellowship.

References

1. Shah RM, Wong C, Arpey NC, Patel AA, Divi SN: A surgeon's guide to understanding artificial intelligence and machine learning studies in orthopaedic surgery. *Curr Rev Musculoskelet Med* 2022;15:121-132.

2. Huang S, Arpaci I, Al-Emran M, Klçarslan S, Al-Sharafi MA: A comparative analysis of classical machine learning and deep learning techniques for predicting lung cancer survivability. *Multimedia Tools Appl* 2023;82:34183-34198.

3. Schulz M-A, Yeo BT, Vogelstein JT, et al: Different scaling of linear models and deep learning in UKBiobank brain images versus machine-learning datasets. *Nat Commun* 2020;11:4238.

4. Acosta JN, Falcone GJ, Rajpurkar P, Topol EJ: Multimodal biomedical AI. *Nat Med* 2022;28:1773-1784.

5. Anastasio AT, Zinger BS, Anastasio TJ: A novel application of neural networks to identify potentially effective combinations of biologic factors for enhancement of bone fusion/repair. *PLoS One* 2022;17:e0276562.

6. Sieg W: Calculations by man and machine: Conceptual analysis, in *Reflections on the foundations of Mathmathematics (essays in honor of Solomon Feferman)*, 2001, vol. 15, pp 387-406.

7. Russakovsky O, Deng J, Su H, et al: Imagenet large scale visual recognition challenge. *Int J Comput Vis* 2015;115:211-252.

8. Luccioni AS, Rolnick D. Bugs in the data: How ImageNet misrepresents biodiversity. Paper presented at the Proceedings of the AAAI Conference on Artificial Intelligence, 2023;37:14382-14390.

9. Parikh RB, Teeple S, Navathe AS: Addressing bias in artificial intelligence in health care. *JAMA* 2019;322:2377-2378.

10. Grill-Spector K, Malach R: The human visual cortex. *Annu Rev Neurosci* 2004;27:649-677.

11. Verdhan V, Verdhan V: VGGNet and AlexNet networks, in *Computer Vision Using Deep Learning: Neural Network Architectures with Python and Keras*, 2021, pp 103-139.

12. Wang S-H, Zhang Y-D: DenseNet-201-based deep neural network with composite learning factor and precomputation for multiple sclerosis classification. *ACM Trans Multimedia Comput Commun Appl* 2020;16:1-19.

13. Jin C-B, Kim H, Liu M, et al: Deep CT to MR synthesis using paired and unpaired data. *Sensors* 2019;19:2361.

14. Yamada Y, Maki S, Kishida S, et al: Automated classification of hip fractures using deep convolutional neural networks with orthopedic surgeon-level accuracy: Ensemble decision-making with antero-posterior and lateral radiographs. *Acta Orthopaedica* 2020;91:699-704.

15. Suzuki T, Maki S, Yamazaki T, et al: Detecting distal radial fractures from wrist radiographs using a deep convolutional neural network with an accuracy comparable to hand orthopedic surgeons. *J Digital Imaging* 2022;35:39-46.

16. Guida C, Zhang M, Shan J: Knee osteoarthritis classification using 3d CNN and MRI. *Appl Sci* 2021;11:5196.

17. Liu Z-Q, Hu Z-J, Wu T-Q, et al: Bone age recognition based on mask R-CNN using xception regression model. *Front Physiol* 2023;14:1062034.

18. Kang Y-J, Yoo J-I, Cha Y-H, Park CH, Kim J-T: Machine learning–based identification of hip arthroplasty designs. *J Orthop Translat* 2020;21:13-17.

19. Borjali A, Chen AF, Muratoglu OK, Morid MA, Varadarajan KM: Detecting total hip replacement prosthesis design on plain radiographs using deep convolutional neural network. *J Orthop Res* 2020;38:1465-1471.

20. Callegaro D, Miceli R, Gronchi A: Sarcoma nomograms: A light over the darkness. *Oncoscience* 2017;4:15-16.

21. Esteva A, Feng J, van der Wal D, et al: Prostate cancer therapy personalization via multi-modal deep learning on randomized phase III clinical trials. *NPJ Digital Med* 2022;5:71.

22. Kasa K, Burns D, Goldenberg MG, Selim O, Whyne C, Hardisty M: Multi-Modal deep learning for assessing surgeon technical skill. *Sensors* 2022;22:7328.

23. Vallières M, Freeman CR, Skamene SR, El Naqa I: A radiomics model from joint FDG-PET and MRI texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities. *Phys Med Biol* 2015;60:5471-5496.

24. White LM, Atinga A, Naraghi AM, et al: T2-weighted MRI radiomics in high-grade intramedullary osteosarcoma: Predictive accuracy in assessing histologic response to chemotherapy, overall survival, and disease-free survival. *Skeletal Radiol* 2023;52:553-564.

25. Welch ML, McIntosh C, Haibe-Kains B, et al: Vulnerabilities of radiomic signature development: The need for safeguards. *Radiother Oncol* 2019;130:2-9.

26. Carloni G, Garibaldi C, Marvaso G, et al: Brain metastases from NSCLC treated with stereotactic radiotherapy: Prediction mismatch between two different radiomic platforms. *Radiother Oncol* 2023;178:109424.

27. Ziegelmayer S, Reischl S, Harder F, Makowski M, Braren R, Gawlitza J: Feature robustness and diagnostic capabilities of convolutional neural networks against radiomics features in computed tomography imaging. *Invest Radiol* 2022;57:171-177.

28. Whitney HM, Li H, Ji Y, Liu P, Giger ML: Comparison of breast MRI tumor classification using human-engineered radiomics, transfer learning from deep convolutional neural networks, and fusion methods. *Proc IEEE* 2020;108:163-177.

29. Nyflot MJ, Thammasorn P, Wootton LS, Ford EC, Chaovalitwongse WA: Deep learning for patient-specific quality assurance: Identifying errors in radiotherapy delivery by radiomic analysis of gamma images with convolutional neural networks. *Med Phys* 2019;46:456-464.

30. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B: A systematic review shows no performance benefit of machine

learning over logistic regression for clinical prediction models. *J Clin Epidemiol* 2019;110:12-22.

31. Navarro F, Dapper H, Asadpour R, et al: Development and external validation of deep-learning-based tumor grading models in soft-tissue sarcoma patients using MR imaging. *Cancers* 2021;13:2866.

32. Collins GS, Dhiman P, Andaur Navarro CL, et al: Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. *BMJ Open* 2021;11:e048008.

33. Ibrahim H, Liu X, Denniston AK: Reporting guidelines for artificial intelligence in healthcare research. *Clin Exp Ophthalmol* 2021;49:470-476.

34. Carré A, Klausner G, Edjlali M, et al: Standardization of brain MR images across machines and protocols: Bridging the gap for MRI-based radiomics. *Scientific Rep* 2020;10:12340.

35. Harrison K, Pullen H, Welsh C, Oktay O, Alvarez-Valle J, Jena R: Machine learning for auto-segmentation in radiotherapy planning. *Clin Oncol* 2022;34:74-88.

36. Kunze KN, Krivicich LM, Clapp IM, et al: Machine learning algorithms predict achievement of clinically significant outcomes after orthopaedic surgery: A systematic review. *Arthrosc J Arthroscopic Relat Surg* 2022;38:2090-2105.

37. Rieke N, Hancox J, Li W, et al: The future of digital health with federated learning. *NPJ digital Med* 2020;3:119.