■ EDITORIAL

Deep learning in orthopaedic research

WEIGHING IDEALISM AGAINST REALISM

J. M. Vigdorchik,
S. J. Jang,
M. J. Taunton,
F. S. Haddad

From facial recognition to self-driving cars, artificial intelligence (AI) technology is already deeply embedded in our daily lives. There has been a recent surge in the orthopaedic deep learning literature due to the increasing accessibility of AI tools for researchers and health systems.[1,2] The value of deep learning literature in orthopaedic surgery is likely to continue to increase. Yet, for deep learning to reach its full potential and translate into orthopaedic clinical practice, high-quality research with transparent data, sound methodology, and external validation is necessary.

AI entails the use of algorithms to perform human cognitive tasks. It has various branches. Machine learning is a subset of AI involving the use of algorithms and statistical analysis to identify patterns in a dataset based on previous examples (i.e. observed or trained data).[3]

Computer vision is an interdisciplinary field described as the machine understanding of visual data, and is also a growing field of research in medicine.[4,5] This has been especially true in trauma and orthopaedics due to the vast numbers of radiological images routinely acquired for the care of our patients.

Computer vision tasks are conducted through forms of deep learning using multi-layered neural networks.[6] As the name suggests, a neural network is a computational structure originally inspired by the biological nature of neurones (input = dendrite, output = axons, weight = synapse, function = neurone body, and network = multiple linked neurones).[7] In theory, a deep learning model can 'learn' like a connection of neurones would in the brain.

Recently, we have seen an increase in literature harnessing the power of deep learning in aspects of orthopaedic surgery, such as the classification of implants by manufacturers (i.e. image classification),[8,9] the identification of fractures (i.e. object detection),[10] and the annotation of bony landmarks for the automation of measurements (i.e. object segmentation).[11,12] We are just beginning to see the potential of the applications of deep learning in our specialty.

However, we must be aware of the limitations of the quality of data in this promising area of research.

Models of deep learning are created by training the algorithm on a specified dataset. Intrinsically, the model is only as robust as the data it learns from, and thus high-quality data are necessary. Using a homogenous dataset will only lead the model to extrapolate patterns specific to those data.[13,14] Concerningly, a dataset may inherently reflect clinical patterns driven by disparities in healthcare. A model can and will learn to replicate such flaws. For example, a study showed that a model of deep learning predicting pain scores from radiographs of the knee had greater racial disparity in its predictions when using a non-diverse training set compared with a diverse training set.[13]

Likewise, a dataset can be mistakenly mislabelled with serious implications as these errors will be perpetuated during the clinical application of these tools. From the initial design of a study, investigators need to carefully consider the characteristics of the data. They must identify its weaknesses and potential biases and explicitly acknowledge the difficulties. These recommendations reflect established guidelines such as those of the Transparent Reporting of a multivariate prediction model for Individual Prognosis or Diagnosis (TRIPOD),[15] which have already been used to evaluate machine learning literature in orthopedics,[16] and the same standards for research should apply when using deep learning.

'Overfitting' can also occur when an algorithm or model learns specific features of a training data set and has good results on its own predetermined test set. However, the model may have not learned 'general' patterns for the task being investigated. When data from an external source are used, the performance of the algorithm can be poor.[17] Methods such as cross-validation or independent testing cohorts mitigate this problem.[18,19] Preferably, however, testing the model on data from an external, multi-institutional setting with large, heterogeneous groups of patients would truly examine a model's generalizability and applicability.[10] Obtaining this type of data is indeed a challenge in all medical research, but it is a goal which we should particularly strive to achieve in the digital age of big data.

Despite these potential limitations of the deep learning literature in orthopaedic surgery, there is an eagerness for it to grow and succeed. Some authors have reported the accurate prediction of clinical outcomes,[20] the expert assessment of disease,[21,22] and the automation of measurements from images with an accuracy which is comparable

Correspondence should be sent to S. J. Jang; email: sej4001@med.cornell.edu

doi:10.1302/0301-620X.104B8.BJJ-2022-0416 $2.00

Bone Joint J
2022;104-B(8):909–910.