




Deep Learning in Medicine

Samuel P. Heilbroner and Riccardo Miotto 

CJASN 18: 397–399, 2023. doi: <https://doi.org/10.2215/CJN.0000000000000080>

Medicine is in a new era where clinical decisions are influenced by analysis on large quantities of data. However, making full use of biomedical data such as electronic health records (EHRs), -omics biobanks, clinical images, and wearable measurements is challenging, owing to their high dimensionality, heterogeneity, temporal dependency, sparsity, redundancy, bias, and irregularity.¹

A common approach is to have a domain expert specify the phenotype of interest in an *ad hoc* manner. However, this approach scales poorly and misses opportunities to discover novel patterns. Advances in machine learning (ML) can be exploited to overcome these limitations. ML is a general-purpose paradigm that learns relationships from the data without the need to define them *a priori*.²

For decades, constructing an ML system required careful engineering and domain expertise to transform the raw data into a suitable internal representation that facilitates learning such relationships. By contrast, a deep learning system uses neural networks to automatically develop its own representations from the raw data. Neural network layers are typically arranged sequentially and composed of a large number of basic, nonlinear operations, such that the representation of one layer (beginning with the raw data input) is fed into the next layer and transformed into a more abstract representation. As data flow through the layers of the system, the input space is iteratively reshaped until attributes of interest become distinguishable.³

The use of deep learning in medicine has been increasing since 2012, with several models successfully going from research to clinical deployment.⁴ These models can scale to very large datasets, mostly because of their ability to run on specialized computing hardware, and continue to improve with more data, enabling them to outperform many classical ML approaches. Another advantage of deep learning models is their ability to naturally ingest multimodal data (e.g., EHRs, medical images, and genomic data), leading to frameworks that can holistically represent a patient's clinical status.

Several types of neural networks are available as building blocks of deep learning architectures. In the bag-of-data-points scenario, a fully connected network can capture correlations among different features. If data have a natural and invariant adjacency structure, such as images, a convolutional neural network (CNN)

can take advantage of that structure by emphasizing local relationships, especially in early layers of the model. When data have a strong temporal component (e.g., time series), recurrent neural networks (RNNs) can model the time sequences of the events. Both RNNs and CNNs, however, struggle to perceive dependencies between data points that are temporally or spatially far from each other. Attention-based architectures such as transformers more efficiently account for these distant interactions and are quickly becoming the state of the art in a number of applications.

Deep architectures are trained in different ways using variations of the back-propagation algorithm. A typical scenario is the supervised learning paradigm, where models learn to map an input (e.g., a chest x-ray scan or a group of laboratory test results) to a label (e.g., a diagnosis of pneumonia or the prediction of the onset of a metabolic condition). There are several potential applications of supervised learning in medicine for adverse event detection, medical image classification and segmentation, and patient risk stratification.¹ For example, deep learning was used to predict AKI from EHRs⁵ and classify kidney biopsy images of patients with diabetic nephropathy.⁶

Training using supervised learning requires datasets in which each input is annotated with its corresponding label. These are commonly derived manually and must be of high quality (gold label) to obtain generalizable models that can effectively assign one of the labels to new data points. Given the time and expense associated with this process, it is often advantageous to use weak supervision, which defines silver labels that are not perfect but are strongly correlated with the gold values. In medicine, billing codes are an abundant source of silver labels, with the advantage of also being regularly updated, leading to annotations that account for changes in the population. In cases where silver labeling is not possible, a significant logistical and financial investment is often necessary to acquire reliable labels. Another option is unsupervised learning.

Unsupervised learning is used to draw inferences from datasets consisting of input data without any labeled annotation. In medicine, the goal is often to cluster patients according to their clinical characteristics to identify novel disease subtypes and phenotypes (namely, patient stratification), which can inform more personalized clinical care or provide avenues for future research. As an example, unsupervised learning was

Tempus Labs,
Chicago, Illinois

Correspondence:
Dr. Riccardo Miotto,
Tempus Labs,
Chicago, IL 60654.
Email: riccardo.miotto@tempus.com

used to subphenotype patients with sepsis-induced AKI in ways that informed the underlying physiology and patient mortality.⁷

Self-supervised learning is another type of unsupervised learning. In this paradigm, a first phase pretrains models in a preparation task using large-scale unlabeled datasets. Preparation tasks could be defined by occluding portions of the data and expecting the model to predict what was hidden or by providing two samples from the same distribution and training the model to associate them strongly (contrastive learning). After such preliminary training, these general architectures can be fine-tuned on a much smaller set of labeled examples for various supervised learning tasks. This pretraining phase is valuable because the model can learn how to find relevant attributes in large-scale data, even before seeing any labeled data, and has led to significant improvements in scalability and performance. Self-supervised learning has been used in medicine with pathology slides, electrocardiograms, clinical notes, EHRs, and x-ray scans.⁸

Different neural network architectures and learning strategies are used in unsupervised learning. Denoising autoencoders derive compressed representations of data, which focus only on areas of real value, and are trained to recreate the original data from a corrupted input. Variational autoencoders also recreate the original data while enforcing a probability distribution on the model's internal representation. Generative adversarial networks (GANs) are composed of two networks: a generator and a discriminator. The generator is trained to create realistic synthetic examples that can trick the discriminator, whereas the discriminator is trained to correctly separate real examples from synthetic ones. All of these models can augment downstream ML efforts. Consider the task of classifying magnetic resonance imaging slices: A GAN could be used to generate additional training examples, while autoencoders could generate useful feature representations for supervised modeling.

In the medical domain, ethical, legal, and logistical hurdles make it challenging to aggregate data from across institutions into something large enough to train a deep learning model. Deidentification can smooth the path to data aggregation and access. Another strategy is federated learning. ML models are trained across multiple institutions, such that data from each institution never leave their own servers—alleviating some of the ethical and legal challenges related to aggregating datasets.

It is important to consider how deep architectures can be obtained from the literature into the real world,⁹ a challenge highlighted by how few of these models are currently used to improve patient care. The level of evidence required to bring an ML model into clinical practice depends on the specific risks and benefits of that model and its use case. In an ideal world, all models would be validated using a randomized controlled trial (RCT), where clinical practice is aided by a model for a randomly selected group of patients and their outcomes are compared with the controls. However, many algorithms used in clinical practice today did not reach this bar. For example, the ubiquitous CHADS-Vasc score has never

been validated with an RCT. There is no one-size-fits-all approach, and this decision should be made with a multidisciplinary team of ML scientists, engineers, practitioners (*e.g.*, nurses, clinicians), hospital administration, and the patients themselves. It is also necessary that these changes be embedded in the system without disrupting it, following the regulatory frameworks provided by the Food and Drug Administration.

At minimum, all ML models should undergo rigorous validation, which estimates how a model would actually perform in the real world. This is typically done by measuring model performance against a test set (data that the model was not trained on). Performance on the testing set is not affected by overfitting, a process in which the model memorizes individual training examples without abstracting the general concepts necessary to make future inferences. The simplest way to generate a test set is by randomly dividing the data into separate training and testing subsets. This does not work when the model is applied to data from a different population, such as patients from a different hospital. In this case, scientists often use out-of-sample validation, in which the external data are used for additional testing.

Because unsupervised models are trained without an objective, they are harder to overfit. However, the process of extracting insights from results of an unsupervised analysis is subject to the very human bias to see patterns, which is itself a form of overfitting. Because there are no labels, it is also more difficult to design a validation experiment, resulting in laxer validation standards. Although these models are often created without an initial hypothesis, the insights generated can actually be thoughtfully validated, even if the process is less formulaic than for supervised learning. Taking patient stratification as an example, an unsupervised model can be used to classify patients into disease subtypes identified by clustering. Ideally, these subtypes have statistically and clinically significant differences in their attributes, and clinicians could use this information to inform treatment decisions. An RCT, where patients who have treatment decisions informed by their subtype membership are compared with the standard of care, can then also be used to formally test this hypothesis.

Several barriers such as interpretability and robustness to all populations also stand between deep learning and wide adoption in medicine.¹⁰ While the steps outlined above to move deep learning into the clinic represent significant challenges, we envision that there will be an increasing number of success stories in the foreseeable future, leading to new sets of practices and tools that will significantly affect patient health and clinical practice.

Disclosures

S.P. Heilbroner reports employment with Tempus Labs; consultancy agreements with Bristol Myers Squibb; ownership interest in ConcertAI and Tempus Labs; and patents or royalties from ConcertAI. R. Miotto reports employment with and ownership interest in Tempus Labs.

Funding

None.

Acknowledgments

This article is part of the Artificial Intelligence and Machine Learning in Nephrology series, led by series editor Girish N. Nadkarni.

The content of this article reflects the personal experience and views of the author(s) and should not be considered medical advice or recommendation. The content does not reflect the views or opinions of the American Society of Nephrology (ASN) or CJASN. Responsibility for the information and views expressed herein lies entirely with the author(s).

Author Contributions

S.P. Heilbroner and R. Miotto wrote the original draft.

References

1. Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: review, opportunities and challenges. *Brief Bioinform*. 2018;19(6):1236–1246. doi:10.1093/bib/bbx044
2. Jordan ML, Mitchell TM. Machine learning: trends, perspectives, and prospects. *Science*. 2015;349(6245):255–260. doi:10.1126/science.aaa8415
3. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015; 521(7553):436–444. doi:10.1038/nature14539
4. Yu KH, Beam AL, Kohane IS. Artificial intelligence in healthcare. *Nat Biomed Eng*. 2018;2(10):719–731. doi:10.1038/s41551-018-0305-z
5. Tomašev N, Glorot X, Rae JW, et al. A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature*. 2019;572(7767):116–119. doi:10.1038/s41586-019-1390-1
6. Ginley B, Lutnick B, Jen KY, et al. Computational segmentation and classification of diabetic glomerulosclerosis. *J Am Soc Nephrol*. 2019;30(10):1953–1967. doi:10.1681/ASN.2018121259
7. Chaudhary K, Vaid A, Duffy Á, et al. Utilization of deep learning for subphenotype identification in sepsis-associated acute kidney injury. *Clin J Am Soc Nephrol*. 2020;15(11):1557–1565. doi:10.2215/CJN.09330819
8. Krishnan R, Rajpurkar P, Topol EJ. Self-supervised learning in medicine and healthcare. *Nat Biomed Eng*. 2022;6(12):1346–1352. doi:10.1038/s41551-022-00914-1
9. Chen D, Liu S, Kingsbury P, et al. Deep learning and alternative learning strategies for retrospective real-world clinical data. *NPJ Digit Med*. 2019;2(1):43–45. doi:10.1038/s41746-019-0122-0
10. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in healthcare. *Lancet Digital Health*. 2021;3(11):e745–e750. doi:10.1016/S2589-7500(21)00208-9

Published online ahead of print. Publication date available at www.cjasn.org.