PERSPECTIVE | FOCUS
https://doi.org/10.1038/s41591-018-0316-z
nature medicine

A guide to deep learning in healthcare

Andre Esteva[1,3]*, Alexandre Robicquet[1,3], Bharath Ramsundar[1], Volodymyr Kuleshov[1], Mark DePristo[2], Katherine Chou[2], Claire Cui[2], Greg Corrado[2], Sebastian Thrun[1] and Jeff Dean[2]

Here we present deep-learning techniques for healthcare, centering our discussion on deep learning in computer vision, natural language processing, reinforcement learning, and generalized methods. We describe how these computational techniques can impact a few key areas of medicine and explore how to build end-to-end systems. Our discussion of computer vision focuses largely on medical imaging, and we describe the application of natural language processing to domains such as electronic health record data. Similarly, reinforcement learning is discussed in the context of robotic-assisted surgery, and generalized deep-learning methods for genomics are reviewed.

Deep learning[1], a subfield of machine learning (ML), has seen a dramatic resurgence in the past 6 years, largely driven by increases in computational power and the availability of massive new datasets. The field has witnessed striking advances in the ability of machines to understand and manipulate data, including images[2], language[3], and speech[4]. Healthcare and medicine stand to benefit immensely from deep learning because of the sheer volume of data being generated (150 exabytes or 10^{18} bytes in United States alone, growing 48% annually[5]) as well as the increasing proliferation of medical devices and digital record systems.

ML is distinct from other types of computer programming in that it transforms the inputs of an algorithm into outputs using statistical, data-driven rules that are automatically derived from a large set of examples, rather than being explicitly specified by humans. Historically, constructing a ML system required domain expertise and human engineering to design feature extractors that transformed raw data into suitable representations from which a learning algorithm could detect patterns. In contrast, deep learning is a form of representation learning—in which a machine is fed with raw data and develops its own representations needed for pattern recognition—that is composed of multiple layers of representations. These layers are typically arranged sequentially and composed of a large number of primitive, 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[1]. As data flows through the layers of the system, the input space becomes iteratively warped until data points become distinguishable (Fig. 1a). In this manner, highly complex functions can be learned.

Deep-learning models scale to large datasets—in part owing to their ability to run on specialized computing hardware—and continue to improve with more data, enabling them to outperform many classical ML approaches. Deep-learning systems can accept multiple data types as input—an aspect of particular relevance for heterogeneous healthcare data (Fig. 1b). The most common models are trained using supervised learning, in which datasets are composed of input data points (e.g., skin lesion images) and corresponding output data labels (e.g., 'benign' or 'malignant'). Reinforcement learning (RL), in which computational agents learn by trial and error or by expert demonstration, has progressed with the adoption of deep learning, achieving remarkable feats in areas such as game playing (e.g., Go[6]). RL can be useful in healthcare whenever learning requires physician demonstration, for instance in learning to suture wounds for robotic-assisted surgery[7].

缝合伤口

Computer vision

Some of the greatest successes of deep learning have been in the field of computer vision (CV)[2]. CV focuses on image and video understanding, and deals with tasks such as object classification, detection, and segmentation—which are useful in determining whether a patient's radiograph contains malignant tumors. Convolutional neural networks (CNNs)[1,2], a type of deep-learning algorithm designed to process data that exhibits natural spatial invariance (e.g., images, whose meanings do not change under translation), have grown to be central in this field.

Medical imaging, for instance, can greatly benefit from recent advances in image classification and object detection[2,8]. Many studies have demonstrated promising results in complex diagnostics spanning dermatology[9,10], radiology[11–14], ophthalmology[15–17], and pathology[18–21] (Fig. 2). Deep-learning systems could aid physicians by offering second opinions and flagging concerning areas in images.

Image-level diagnostics have been quite successful at employing CNN-based methods (Fig. 2). This is largely due to the fact that CNNs have achieved human-level performance in object-classification tasks[2], in which a CNN learns to classify the object contained in an image. These same networks have demonstrated strong performance in transfer learning[22], in which a CNN initially trained on a massive dataset that is unrelated to the task of interest (e.g., ImageNet[2], a dataset of millions of common everyday objects) is further fine-tuned on a much smaller dataset related to the task of interest (e.g., medical images). In the first step, the algorithm leverages large amounts of data to learn of the natural statistics in images—straight lines, curves, colorations, etc.—and in the second step, the higher-level layers of the algorithm are retrained to distinguish between diagnostic cases. Similarly, object detection and segmentation algorithms identify specific parts of an image that correspond to particular objects. CNN methods take image data as input and iteratively warp it through a series of convolutional and nonlinear operations until the original raw data matrix is transformed into a probability distribution over potential image classes (e.g., medical diagnostic cases) (Fig. 2).

Remarkably, deep-learning models have achieved physician-level accuracy at a broad variety of diagnostic tasks, including identifying moles from melanomas[9,10], diabetic retinopathy, cardiovascular risk, and referrals from fundus[15,16] and optical coherence tomography (OCT)[17] images of the eye, breast lesion detection in mammograms[13], and spinal analysis with magnetic resonance imaging[23]. A single deep-learning model has even been shown to be effective at diagnosis across medical modalities (e.g., radiology and ophthalmology)[24].

[1]Stanford University, Stanford, CA, USA. [2]Google Research, San Jose, CA, USA. [3]These authors contributed equally: Andre Esteva, Alexandre Robicquet.
*e-mail: andre.esteva@gmail.com

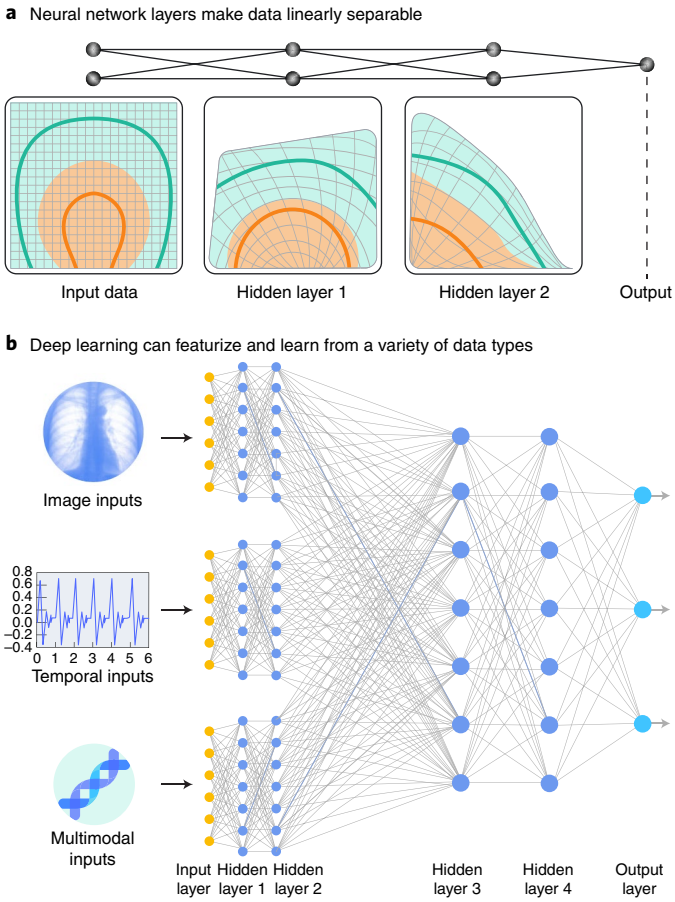

Fig. 1 | Deep learning. a, A simple, multilayer deep neural network takes two classes of data, denoted by the different colors, and makes them linearly separable by iteratively distorting the data as it flows from layer to layer. The final output layer serves as a classifier by outputting the probability of either one of the classes. This example illustrates the basic concept used by large scale networks. Conceptual illustration adapted with permission from http://colah.github.io/. **b**, Example large-scale network that accepts as input a variety of data types (images, time-series, etc.), and for each data type learns a useful featurization in its lower-level towers. The data from each tower is then merged and flows through higher levels, allowing the DNN to perform inference across data types—a capability that is increasingly important in healthcare.

However, a key limitation across studies that compare human to algorithmic performance has been a lack of clinical context—they constrain the diagnosis to be performed using just the images at hand. This often increases the difficulty of the diagnostic task for the human reader, who in real-world clinical settings has access to both the medical imagery and supplemental data, including the patient history and health record, additional tests, patient testimony, etc.

Clinics are beginning to employ object detection and segmentation in images for urgent and easily missed cases, such as flagging large-artery occlusion in the brain using radiological images[14], during which patients have a limited amount of time (a few minutes) before permanent brain damage occurs. Further, cancer histopathology reads, which require human experts to laboriously scan and diagnose gigapixel images (or equivalently large physical slides) can be supplemented with CNNs trained to detect mitotic cells[18] or tumor regions[19]. They can be trained to quantify the amount of PD-L1 present in a histopathology image[20]—a task important in determining which type of immuno-oncology drug a patient would be receptive to. Combined with pixel-level analyses, CNNs have even been used to discover biological features of tissue associated with survival probability[21].

The primary limitation to building a supervised deep-learning system for a new medical imaging task is access to a sufficiently large, labeled dataset. Small and labeled datasets for specific tasks are easier to collect, but result in algorithms that tend to perform poorly on new data. In these cases, techniques for heavy data augmentation have been shown to be effective at helping algorithms generalize[25]. Similarly, large but unlabeled datasets are also easier to collect, but will require a shift towards improved semisupervised and unsupervised techniques, such as generative adversarial networks[26].

Natural language processing

Natural language processing (NLP) focuses on analyzing text and speech to infer meaning from words. Recurrent neural networks (RNNs)—deep learning algorithms effective at processing sequential inputs such as language, speech, and time-series data[27]—play an important role in this field. Notable successes of NLP include machine translation[28], text generation[29], and image captioning[30]. In healthcare, sequential deep learning and language technologies power applications within domains such as electronic health records (EHRs).

EHRs are rapidly becoming ubiquitous[31]. The EHR of a large medical organization can capture the medical transactions of over 10 million patients throughout the course of a decade. A single hospitalization alone typically generates ~150,000 pieces of data. The potential benefits derived from this data are significant. In aggregate, an EHR of this scale represents 200,000 years of doctor wisdom and 100 million years of patient outcome data, covering a plethora of rare conditions and maladies. As such, applyication of deep-learning methods to EHR data is a rapidly expanding area[32,33].

Figure 3 outlines the technical steps in building deep-learning systems for EHRs. Raw data are first aggregated across across institutions in order to ensure that a generalizable system is built. The data are then standardized and parsed temporally and across patients, which makes them suitable for deep-learning training. From this, we can then infer answers to high-level medical questions, such as 'What past history is relevant to the patient's current diagnosis?', 'What is the patient's current problem list?', and 'What opportunities are there to intervene?'

When making predictions, most work to date uses supervised learning on limited sets of structured data, including lab results, vitals, diagnostic codes, and demographics. To account for the structured and unstructured data contained in EHRs, researchers are beginning to employ unsupervised learning approaches, such as auto-encoders—in which networks are first trained to learn useful representations by compressing and then reconstructing unlabeled data—to predict specific diagnoses[34]. Recent uses of deep learning model the temporal sequence of structured events that occurred in a patient's record with convolutional and recurrent neural networks in order to predict future medical incidents[35–38]. Much of this work focuses on the Medical Information Mart for Intensive Care (MIMIC) dataset[39] (e.g., for the prediction of sepsis[40]), which contains intensive care unit (ICU) patients from a single center. While ICU patients generate more EHR data than non-ICU patients, they are significantly outnumbered by non-ICU patients. As such, it is still uncertain how well techniques derived from this data will generalize to broader populations.

The next generation of automatic speech recognition[32] and information extraction models will likely develop clinical voice assistants to accurately transcribe patient visits. Doctors easily spend 6 hours in an 11-hour workday working on documentation in the EHR, which leads to burnout and reduces time with patients[31]. Automated transcription will alleviate this and facilitate more affordable scribing services. Consider RNN-based language translation[27],

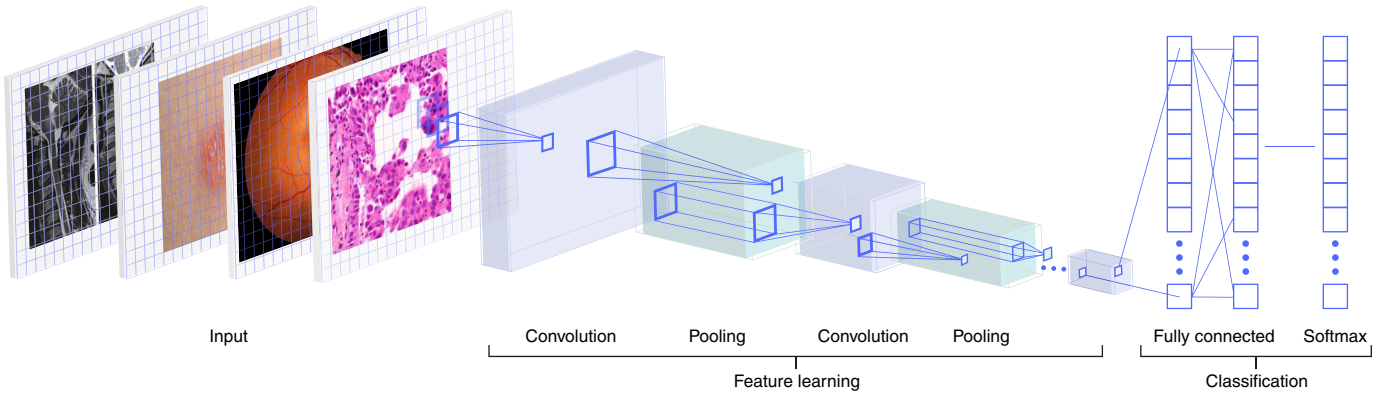

Fig. 2 | Medical Imaging. CNNs can be trained on a variety of medical imagery, including radiology, pathology, dermatology, and ophthalmology. Information flows left to right. CNNs take input images and sequentially transform them, using simple operations such as convolutional, pooling, and fully connected layers, into flattened vectors. The elements of the output vector (softmax layer) represent the probabilities of the presence of disease. During the training process, the internal parameters of the network layers are iteratively adjusted to improve accuracy. Typically, lower layers (left) learn simple image features—edges and basic shapes—which influence the high-level representations (right). Prediction tasks include both classification of the images (i.e., cancerous versus benign) as well as localization of medical features such as tumors.

which uses an end-to-end technique to translate directly from speech in one language to text in another. Adapted to EHRs, this technique could translate a patient–provider conversation directly into a transcribed text record. The key challenge lies in classifying the attributes and status of each medical entity from the conversation while accurately summarizing the dialogue. Though promising in early human–computer interaction experiments, these techniques have yet to be widely deployed in medical practice.

Future work will likely focus on developing algorithms to better leverage some of the information-rich yet unstructured data in EHRs. Clinical notes, for instance, are often omitted or redacted when developing predictive systems. Here, large-scale RNNs are beginning to demonstrate impressive predictive results by combining structured and unstructured data in a semisupervised way[33]. This data combination allows them to learn from broader populations across more diverse data types, outperforming other techniques across tasks including mortality, readmission, length of stay, and diagnosis predictions.

Reinforcement learning

Reinforcement learning (RL) refers to a class of techniques designed to train computational agents to successfully interact with their environment, typically to achieve specific goals. This learning can happen through trial and error, through demonstration, or through a hybrid approach. As an agent takes actions within its environment, an iterative feedback loop of reward and consequence trains the agent to better accomplish the goals at hand. Learning from expert demonstration is accomplished either by learning to predict the expert's actions directly via supervised learning (i.e., imitation learning) or by inferring the expert's objective (i.e., inverse RL). To successfully train an agent, it is critical to have a model function that can take as input sensory signals from the environment and output the next actions for the agent to take. Deep RL, in which a deep-learning model serves as the model function, shows promise.

One healthcare domain that can benefit from deep RL is robotic-assisted surgery (RAS). Currently, RAS largely depends on a surgeon guiding a robot's instruments in a teleoperated fashion. Deep learning can enhance the robustness and adaptability of RAS by using computer vision models (e.g., CNNs) to perceive surgical environments and RL methods to learn from a surgeons physical motions[41,42].

These techniques support the automation and speed of highly repetitive and time-sensitive surgical tasks, such as suturing and knot-tying[7]. For instance, computer vision techniques (e.g., CNNs for object detection/segmentation and stereovision) can reconstruct the landscape of an open wound from image data, and a suturing or knot-tying trajectory can be generated by solving a path optimization problem that attempts to find an optimal trajectory while accounting for external constraints, such as joint limits and obstacles[43]. Similarly, image-trained RNNs can learn to tie knots autonomously by learning sequences of events, in this case physical maneuvers, from surgeons[44].

These techniques are particularly advantageous for fully autonomous robotic surgery or minimally invasive surgery. Consider modern laparoscopic surgery (MLS)—in which several small incisions are used to insert a number of instruments into the body, including cameras and surgical tools, which surgeons then teleoperate. Deep imitation learning, RNNs, and trajectory transfer algorithms can fully automate certain teleoperated manipulation tasks of the surgical procedure[7]. In MLS, the automation of repetitive tasks is even more time-critical than in open surgery. For instance, it may take 3 minutes to tie a knot in MLS instead of a few seconds, as in open surgery.

One of the main challenges during semiautonomous teleoperation is correctly localizing an instrument's position and orientation in the vicinity of surgical scenes. Here, recent pixel-wise instrument segmentation techniques[45], developed using an improved U-Net architecture CNN[25,46], begin to show promise. Another challenge for the progression of deep learning in surgical robotics is data collection. Deep imitation learning requires large training datasets with many examples per surgical action. Given that many surgeries are nuanced and unique, it remains difficult to collect sufficient data for more general surgical tasks. Further, it remains difficult for autonomous systems to adapt to completely unknown and unobserved situations highly dissimilar from anything previously seen, such as an anomalous surgical accident.

Generalized deep learning

Beyond CV, NLP, and RL tasks, deep learning is adaptable to domains in which input data is nuanced and requires specialized treatment. For illustrative purposes, here we consider genomics, an example domain in which deep learning has been adapted beyond conventional (e.g., CNN- or RNN-based) approaches to work with unique (e.g., nonimage, nontemporal) data representations.

Modern genomic technologies collect a wide variety of measurements, from an individual's DNA sequence to the quantity of

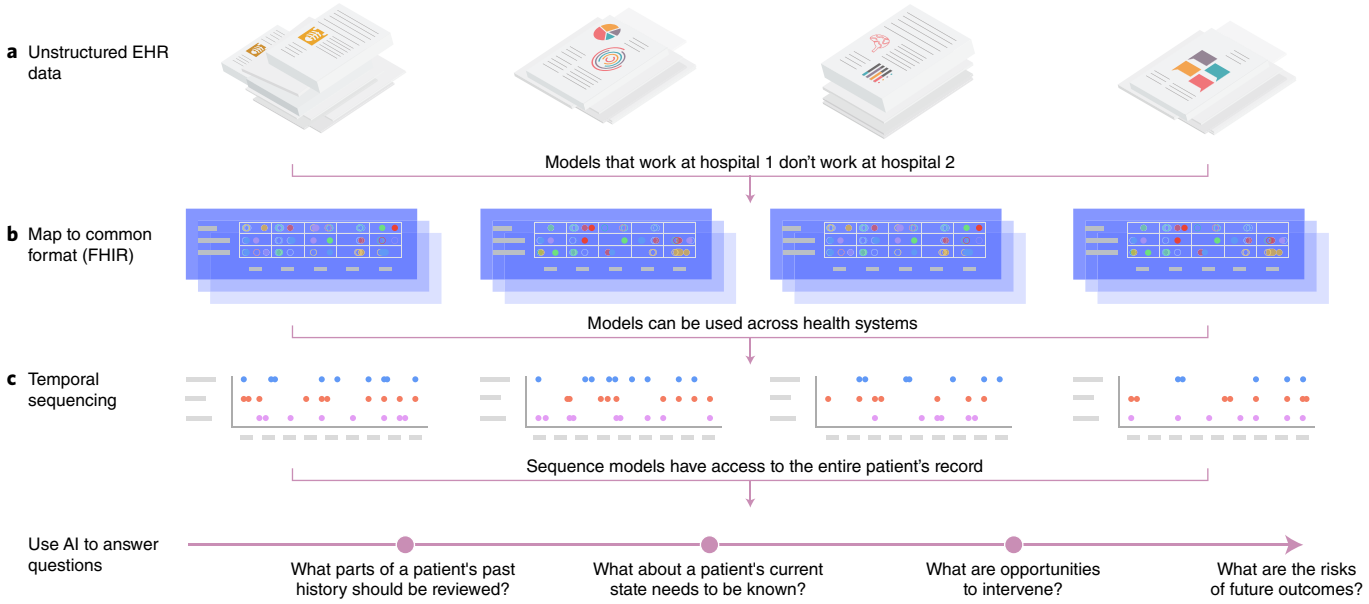

Fig. 3 | Making predictions using EHRs. a, Unstructured EHR data. Medical records are stored in idiosyncratic (异质) data structures and formats such that models built on a given hospital's record do not necessarily work with data from a different hospital. **b**, Data standardization. By mapping data from multiple sites to a single format based on FHIR, data are standardized into a homogeneous format. **c**, Sequencing. By temporally sequencing all data into a patient timeline, time-based deep-learning techniques can be applied on the entirety of EHR datasets for making predictions about single patients.

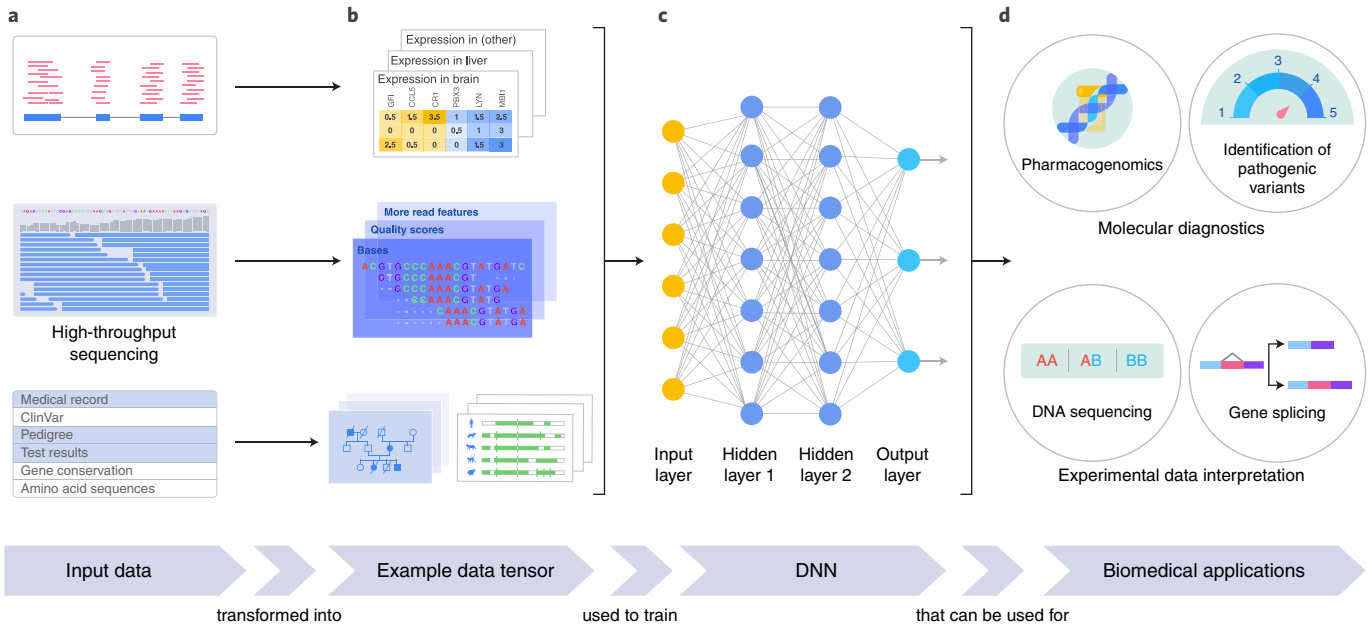

Fig. 4 | ML in genomics. a, Input data. Genomic data consists of experimental measurements from which certain properties or outcomes of interest may be predicted. This data is often diverse and may include sequencing, gene expression, and functional data as well as other forms of molecular data. **b** Example data tensors. Raw experimental measurements need to be transformed into a form that is suitable for consumption by deep-learning algorithms, which take as input multidimensional data tensors and associated target labels. **c**, DNN. Labeled tensors are used to train DNNs to predict the label from the input data tensor. **d**, Biomedical applications. Trained DNNs can be used in biomedical applications, such as in predicting labels for previously unseen data tensors or examining the relationship between input data and output labels. Example applications include interpreting experimental data (e.g., inferring DNA sequences from the output of a sequencing instrument or inferring the effects of DNA mutations on gene splicing) and molecular diagnostics (e.g., predicting the effects of genetic mutations on disease risk or drug response), among many others.

various proteins in their blood. There are many opportunities for deep learning to improve the methods used to analyze these measurements, which will ultimately help clinicians provide more accurate treatments and diagnoses. The typical pipeline for building a deep-learning system in genomics involves taking raw data (e.g., gene expression data), converting this raw data into input data tensors, and feeding these tensors through neural networks which then power specific biomedical applications (Fig. 4).

One set of opportunities centers on genome-wide association (GWA) studies—large case-control studies that seek to discover causal genetic mutations affecting specific traits. Analyzing GWA studies requires algorithms that scale to very large patient cohorts and that deal with latent confounders. These challenges can be addressed via optimization tools and techniques developed for deep learning—including stochastic optimization and other modern algorithms[47] combined with software frameworks for scaling computation in parallel[48]—as well as through modeling techniques that handle unseen confounders[49]. In the near future, models that integrate external modalities and additional sources of biological data into GWA studies—e.g., medical images or measurements of splicing and other intermediary molecular phenotypes—may also benefit from deep learning to more accurately identify disease-associated causal mutations[50].

Understanding the genetics of disease allows clinicians to recommend treatments and provide more accurate diagnoses. A key challenge for physicians is determining if novel variants in a patient's genome are medically relevant. In part, this decision relies on predicting the pathogenicity of mutations; a task which already uses features like protein structure and evolutionary conservation to train learning algorithms[51]. Given their greater power and ability to effectively integrate disparate data types, deep-learning techniques are likely to provide more accurate pathogenicity predictions than are possible today[52].

Machine learning also plays a role in phenotype prediction from genetic data, including complex traits such as height as well as disease risk. Deep learning can further enhance such models by integrating additional modalities such as medical images, clinical history, and wearable device data[53]. A particularly promising approach to phenotype prediction is to predict intermediate molecular phenotypes—e.g., gene expression or gene splicing—which then feed into downstream disease predictors[54]. Intermediate molecular states can be easier to predict than human traits because of larger, more proximal signals and more extensive training data. These two features make the problem a good fit for deep learning, which has shown success at predicting splicing[55] and transcription factor binding[56].

Genomic data can also directly serve as a biomarker for the onset and progression of disease. For example, blood contains small fragments of cell-free DNA released from cells present elsewhere in the body. These fragments are noninvasive indicators of organ rejection (i.e., the immune system attacking graft cells[57]), bacterial infection[58], and early-stage cancer[59]. Cell-free DNA is successfully used in prenatal diagnostics: fetal DNA present in the mother's blood indicates chromosomal aberrations and can reveal the whole genome of the fetus[60]. Biomarker data are often noisy and requires sophisticated analysis (e.g., to determine whether cell-free DNA is indicative of cancer); deep-learning systems can enhance the quality of biomarker assays targeting DNA sequences[61], methylation[62], gene expression[63], chromatin[64] profiles, and many other measurements.

Received: 17 October 2018; Accepted: 28 November 2018;
Published online: 7 January 2019

References

1. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
2. Russakovsky, O.et al. Imagenet large scale visual recognition challenge. *Int. J. Compute. Vis.* **115,** 211–252 (2015).
3. Hirschberg, J. & Manning, C. D. Advances in natural language processing. *Science* **349,** 261–266 2015).
4. Geoffrey Hinton, et al. Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. *IEEE Signal Process. Mag.* **29,** 82–97 (2012).
5. Stanford Health. Harnessing the power of data in health. *Stanford Medicine 2017 Health Trends Report* (2017).
6. Silver, D. et al. Mastering the game of go with deep neural networks and tree search. *Nature* **529**, 484–489 (2016).
7. Yohannes Kassahun, et al. Surgical robotics beyond enhanced dexterity instrumentation: a survey of machine learning techniques and their role in intelligent and autonomous surgical actions. *Int. J. Comput. Assist. Radio. Surg.* **11**, 553–568 (2016).
8. Litjens, G. et al. A survey on deep learning in medical image analysis. *Med. Image Anal.* **42**, 60–88 (2017).
9. Esteva, A. et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* **542**, 115–118 (2017).
10. Haenssle, H. A. et al. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Ann. Oncol.* **29**, 1836–1842 (2018).
11. Cheng, J.-Z. et al. Computer aided diagnosis with deep learning architecture: applications to breast lesions in us images and pulmonary nodules in CT scans. *Sci. Rep.* **6**, 24454 (2016).
12. Cicero, M. et al. Training and validating a deep convolutional neural network for computer-aided detection and classification of abnormalities on frontal chest radiographs. *Invest. Radiol.* **52**, 281–287 (2017).
13. Kooi, T. et al. Large scale deep learning for computer aided detection of mammographic lesions. *Med. Image Anal.* **35**, 303–312 (2017).
14. Barreira, C. M. et al. Abstract WP61: Automated large artery occlusion detection in st roke imaging-paladin study. *Stroke* **49,** AWP61 (2018).
15. Gulshan, V. et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* **316**, 2402–2410 (2016).
16. Poplin, R. et al. Webster. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nat. Biomed. Eng.* **2**, 158–164 (2018).
17. Fauw, J. et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat. Med.* **24**, 1342 (2018).
18. Cireşan, D. C., Giusti, A., Gambardella, L. M. & Schmidhuber, J. Mitosis detection in breast cancer histology images with deep neural networks. In *International Conference on Medical Image Computing and Computer-assisted Intervention* 411–418 (Springer, 2013).
19. Liu, Y. et al. Detecting cancer metastases on gigapixel pathology images. Preprint at https://arxiv.org/abs/1703.02442 (2017).
20. Charoentong, P. et al. Pan-cancer immunogenomic analyses reveal genotype–immunophenotype relationships and predictors of response to checkpoint blockade. *Cell Rep.* **18**, 248–262 (2017).
21. Beck, A. H. et al. Systematic analysis of breast cancer morphology uncovers stromal features associated with survival. *Sci. Transl. Med.* **3**, 108ra113 (2011).
22. Yosinski, J., Clune, J., Bengio, Y. and Lipson, L. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems* 3320–3328 (2014).
23. Jamaludin, A., Kadir, T. and Zisserman, A. Spinenet: automatically pinpointing classification evidence in spinal mris. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* 166–175 (Springer, 2016).
24. Kermany, D. S. et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell* **172**, 1122–1131 (2018).
25. Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2015).
26. Goodfellow, I. et al. Generative adversarial nets. In *Advances in Neural Information Processing Systems* 2672–2680 (2014).
27. Sutskever, I., Vinyals, O. & Le, Q. V. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems* 3104–3112 (2014).
28. Wu, Y. et al. Google's neural machine translation system: bridging the gap between human and machine translation. Preprint at https://arxiv.org/abs/1609.08144 (2016).
29. Kannan, A. et al. Smart reply: automated response suggestion for email. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, 2016).
30. Vinyals, O., Toshev, A., Bengio, S. & Erhan, D. Show and tell: a neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 3156–3164 (2015).
31. The Office of the National Coordinator for Health Information Technology. Quick stats: health IT dashboard. https://dashboard.healthit.gov/quickstats/quickstats.php (2017).
32. Shickel, B., Tighe, P. J., Bihorac, A. & Rashidi, P. Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. *IEEE J. Biomed. Health Inform.* **22**, 1589–1604 (2017).
33. Rajkomar, A. et al. Scalable and accurate deep learning with electronic health records. *NPJ Digit. Med.* **1**, 18 (2018).
34. Miotto, R. et al. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Sci. Rep.* **6**, 26094 (2016).

35. Liu, V., Kipnis, P., Gould, M. K. & Escobar, G. J. Length of stay predictions: improvements through the use of automated laboratory and comorbidity variables. *Med. Care* **48**, 739–744 (2010).
36. Choi, E. et al. Doctor AI: predicting clinical events via recurrent neural networks. In *Machine Learning for Healthcare* 301–318 (2016).
37. Che, Z. et al. Recurrent neural networks for multivariate time series with missing values. *Rep.* **8**, 1–12 (2018).
38. Suresh, H. et al. Clinical intervention prediction and understanding with deep neural networks. *PMLR* **68**, 322–377 (2017).
39. Johnson, A. E. W.et al. Mimic-iii, a freely accessible critical care database. *Sci. Data* **3,** 160035 (2016).
40. Mao, Q.et al. Multicentre validation of a sepsis prediction algorithm using only vital sign data in the emergency department, general ward and icu. *BMJ Open* **8,** e017833 (2018).
41. Abbeel, P. & Ng, A. Y. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the Twenty-First International Conference on Machine Learning* 1 (ACM, 2004).
42. Ratliff, N. D., Silver, D. & Bagnell, J. A. Learning to search: functional gradient techniques for imitation learning. *Autonomous Robots* **27,** 25–53 (2009).
43. Schulman, J. et al. A case study of trajectory transfer through non-rigid registration for a simplified suturing scenario. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* 4111–4117 (IEEE, 2013).
44. Mayer, H.et al. A system for robotic heart surgery that learns to tie knots using recurrent neural networks. *Adv. Robot* **22**, 1521–1537 (2008).
45. Shvets, A., Rakhlin, A., Kalinin, A. A. and Iglovikov, V. Automatic instrument segmentation in robot-assisted surgery using deep learning. Preprint at https://arxiv.org/abs/1803.01207 (2018).
46. Jin, A. et al. Tool detection and operative skill assessment in surgical videos using region-based convolutional neural networks. Preprint at https://arxiv.org/abs/1802.08774 (2018).
47. Loh, P.-R. et al. Efficient bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* **47**, 284 (2015).
48. Abadi, M. et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. http://download.tensorflow.org/paper/whitepaper2015.pdf (2015).
49. Tran, D. and Blei, D. M. Implicit causal models for genome-wide association studies. In *International Conference on Learning Representations* (2018).
50. Lee, S.-I. et al. Learning a prior on regulatory potential from eqtl data. *PLoS Genet.* **5**, e1000358 (2009).
51. Kircher, M. et al. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
52. Quang, D., Chen, Y. & Xie, X. Dann: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics* **31**, 761–3 (2015).
53. Dudley, J. T. et al. Personalized medicine: from genotypes, molecular phenotypes and the quantified self, towards improved medicine. In *Pacific Symposium on Biocomputing* 342–346 (2014).
54. Leung, M. K. K., Delong, A., Alipanahi, B. & Frey, B. J. Machine learning in genomic medicine: a review of computational problems and data sets. *Proc. IEEE* **104**, 176–197 (2016).
55. Xiong, H. Y. et al. The human splicing code reveals new insights into the genetic determinants of disease. *Science* **347**, 1254806 (2015).
56. Alipanahi, B. et al. Predicting the sequence specificities of dna-and rna-binding proteins by deep learning. *Nature Biotechnol.* **33**, 831–838 (2015).
57. Snyder, T. M., Khush, K. K., Valantine, H. A. & Quake, S. R. Universal noninvasive detection of solid organ transplant rejection. *Proc. Natl Acad. Sci. USA* **108**, 6229–6234 (2011).
58. Abril, M. K. et al. Diagnosis of capnocytophaga canimorsus sepsis by whole-genome next-generation sequencing. In *Open Forum Infectious Diseases* Vol. 3, ofw144 (Oxford University Press, 2016).
59. Forshew, T. et al. Noninvasive identification and monitoring of cancer mutations by targeted deep sequencing of plasma dna. *Sci. Transl. Med.* **4**, 136ra68–136ra68 (2012).
60. Fan, H. C. et al. Non-invasive prenatal measurement of the fetal genome. *Nature* **487**, 320–324 (2012).
61. Ching, T. et al. Opportunities and obstacles for deep learning in biology and medicine. Preprint at https://doi.org/10.1101/142760 (2017).
62. Angermueller, C., Lee, H. J., Reik, W. & Stegle, O. Deepcpg: accurate prediction of single-cell dna methylation states using deep learning. *Genome Biol.* **18**, 67 (2017).
63. Chen, Y. et al. Gene expression inference with deep learning. *Bioinformatics* **32**, 1832–1839 (2016).
64. Koh, P. W., Pierson, E. & Kundaje, A. Denoising genome-wide histone chip-seq with convolutional neural networks. *Bioinformatics* **33**, i225–i233 (2017).

Acknowledgements

The authors would like to thank D. Wang, E. Dorfman, and A. Rajkomar for the visual design of the figures in this paper and P. Nejad for insightful conversation and ideas.

Author contributions

B.R., V.K., M.D., and K.C. share second authorship. C.C., G.C., and S.T. share third authorship. J.D. is the principal investigator. A.E. and A.R. conceptualized the structure of the review and contributed to the computer vision and reinforcement learning sections. V.K., B.R., and M.D. contributed to the generalized deep learning section. K.C. and J.D. contributed to the natural language processing section. C.C., G.C., S.T., and J.D. oversaw the work. All authors contributed to multiple parts of the review, as well as the style and overall contents.

Competing interests

M.D., C.C., K.C., G.C. and J.D. are employees of Google Inc. This work was internally funded by Google Inc. G.C. is a board member at the Partnership on AI to Benefit People and Society. S.T. is an employee of Udacity, Inc. and the Kitty Hawk Corporation. He is on the faculty of Stanford University and Georgia Institute of Technology. B.R. is a partner of Computable LLC.

Additional information

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence should be addressed to A.E.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.