Deep learning for healthcare: review, opportunities and challenges

Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, Joel T Dudley
Author Notes
Briefings in Bioinformatics, Volume 19, Issue 6, November 2018, Pages 1236–1246, https://doi.org/10.1093/bib/bbx044
Published:

06 May 2017
Article history

- PDF
- Split View
- Cite
- Permissions Icon Permissions
- Share Icon Share

Abstract

Gaining knowledge and actionable insights from complex, high-dimensional and heterogeneous biomedical data remains a key challenge in transforming health care. Various types of data have been emerging in modern biomedical research, including electronic health records, imaging, -omics, sensor data and text, which are complex, heterogeneous, poorly annotated and generally unstructured. Traditional data mining and statistical learning approaches typically need to first perform feature engineering to obtain effective and more robust features from those data, and then build prediction or clustering models on top of them. There are lots of challenges on both steps in a scenario of complicated data and lacking of sufficient domain knowledge. The latest advances in deep learning technologies provide new effective paradigms to obtain end-to-end learning models from complex data. In this article, we review the recent literature on applying deep learning technologies to advance the health care domain. Based on the analyzed work, we suggest that deep learning approaches could be the vehicle for translating big biomedical data into improved human health. However, we also note limitations and needs for improved methods development and applications, especially in terms of ease-of-understanding for domain experts and citizen scientists. We discuss such challenges and

suggest developing holistic and meaningful interpretable architectures to bridge deep learning models and human interpretability.
deep learning, health care, biomedical informatics, translational bioinformatics, genomics, electronic health records

Introduction

Health care is coming to a new era where the abundant biomedical data are playing more and more important roles. In this context, for example, precision medicine attempts to 'ensure that the right treatment is delivered to the right patient at the right time' by taking into account several aspects of patient's data, including variability in molecular traits, environment, electronic health records (EHRs) and lifestyle [1–3].

The large availability of biomedical data brings tremendous opportunities and challenges to health care research. In particular, exploring the associations among all the different pieces of information in these data sets is a fundamental problem to develop reliable medical tools based on data-driven approaches and machine learning. To this aim, previous works tried to link multiple data sources to build joint knowledge bases that could be used for predictive analysis and discovery [4–6]. Although existing models demonstrate great promises (e.g. [7–11]), predictive tools based on machine learning techniques have not been widely applied in medicine [12]. In fact, there remain many challenges in making full use of the biomedical data, owing to their high-dimensionality, heterogeneity, temporal dependency, sparsity and irregularity [13–15]. These challenges are further complicated by various medical ontologies used to generalize the data (e.g. Systematized Nomenclature of Medicine-Clinical Terms (SNOMED-CT) [16], Unified Medical Language System (UMLS) [17], International Classification of Disease-9th version (ICD-9) [18]), which often contain conflicts and inconsistency [19]. Sometimes, the same clinical phenotype is also expressed in different ways across the data. As an example, in the EHRs, a patient diagnosed with 'type 2 diabetes mellitus' can be identified by laboratory values of hemoglobin A1C >7.0, presence of 250.00 ICD-9 code, 'type 2 diabetes mellitus' mentioned in the free-text clinical notes and so on. Consequently, it is nontrivial to

harmonize all these medical concepts to build a higher-level semantic structure and understand their correlations [6, 20].

A common approach in biomedical research is to have a domain expert to specify the phenotypes to use in an ad hoc manner. However, supervised definition of the feature space scales poorly and misses the opportunities to discover novel patterns. Alternatively, representation learning methods allow to automatically discover the representations needed for prediction from the raw data [21, 22]. Deep learning methods are representation-learning algorithms with multiple levels of representation, obtained by composing simple but nonlinear modules that each transform the representation at one level (starting with the raw input) into a representation at a higher, slightly more abstract level [23]. Deep learning models demonstrated great performance and potential in computer vision, speech recognition and natural language processing tasks [24–27].

Given its demonstrated performance in different domains and the rapid progresses of methodological improvements, deep learning paradigms introduce exciting new opportunities for biomedical informatics. Efforts to apply deep learning methods to health care are already planned or underway. For example, Google DeepMind has announced plans to apply its expertise to health care [28] and Enlitic is using deep learning intelligence to spot health problems on X-rays and Computed Tomography (CT) scans [29].

However, deep learning approaches have not been extensively evaluated for a broad range of medical problems that could benefit from its capabilities. There are many aspects of deep learning that could be helpful in health care, such as its superior performance, end-to-end learning scheme with integrated feature learning, capability of handling complex and multi-modality data and so on. To accelerate these efforts, the deep learning research field as a whole must address several challenges relating to the characteristics of health care data (i.e. sparse, noisy, heterogeneous, time-dependent) as need for improved methods and tools that enable deep learning to interface with health care information workflows and clinical decision support.

In this article, we discuss recent and forthcoming applications of deep learning in medicine, highlighting the key aspects to significantly impact health care. We do not aim to provide a comprehensive background on technical details (see e.g. [21, 30–32]) or general application of deep learning (see e.g. [23]). Instead, we focus on biomedical data only, in particular those originated from clinical imaging, EHRs, genomes and wearable devices. While additional sources of information, such as metabolome, antibodyome and other omics information are expected to be valuable for health monitoring, at this point deep learning has not been significantly used in these domains. Thus, in the following, we briefly introduce the general deep learning framework, we review some of its applications in the medical domain and we discuss the opportunities, challenges and applications related to these methods when used in the context of precision medicine and next-generation health care.

Deep learning framework

Machine learning is a general-purpose method of artificial intelligence that can learn relationships from the data without the need to define them a priori [33]. The major appeal is the ability to derive predictive models without a need for strong assumptions about the underlying mechanisms, which are usually unknown or insufficiently defined [34]. The typical machine learning workflow involves four steps: data harmonization, representation learning, model fitting and evaluation [35]. For decades, constructing a machine learning system required careful engineering and domain expertise to transform the raw data into a suitable internal representation from which the learning subsystem, often a classifier, could detect patterns in the data set. Conventional techniques are composed of a single, often linear, transformation of the input space and are limited in their ability to process natural data in their raw form [21].

Deep learning is different from traditional machine learning in how representations are learned from the raw data. In fact, deep learning allows computational models that are composed of multiple processing layers based

on neural networks to learn representations of data with multiple levels of abstraction [23]. The major differences between deep learning and traditional artificial neural networks (ANNs) are the number of hidden layers, their connections and the capability to learn meaningful abstractions of the inputs. In fact, traditional ANNs are usually limited to three layers and are trained to obtain supervised representations that are optimized only for the specific task and are usually not generalizable [36]. Differently, every layer of a deep learning system produces a representation of the observed patterns based on the data it receives as inputs from the layer below, by optimizing a local unsupervised criterion [37]. The key aspect of deep learning is that these layers of features are not designed by human engineers, but they are learned from data using a general-purpose learning procedure. Figure 1 illustrates such differences at a high level: deep neural networks process the inputs in a layer-wise nonlinear manner to pre-train (initialize) the nodes in subsequent hidden layers to learn 'deep structures' and representations that are generalizable. These representations are then fed into a supervised layer to fine-tune the whole network using the backpropagation algorithm toward representations that are optimized for the specific end-to-end task.

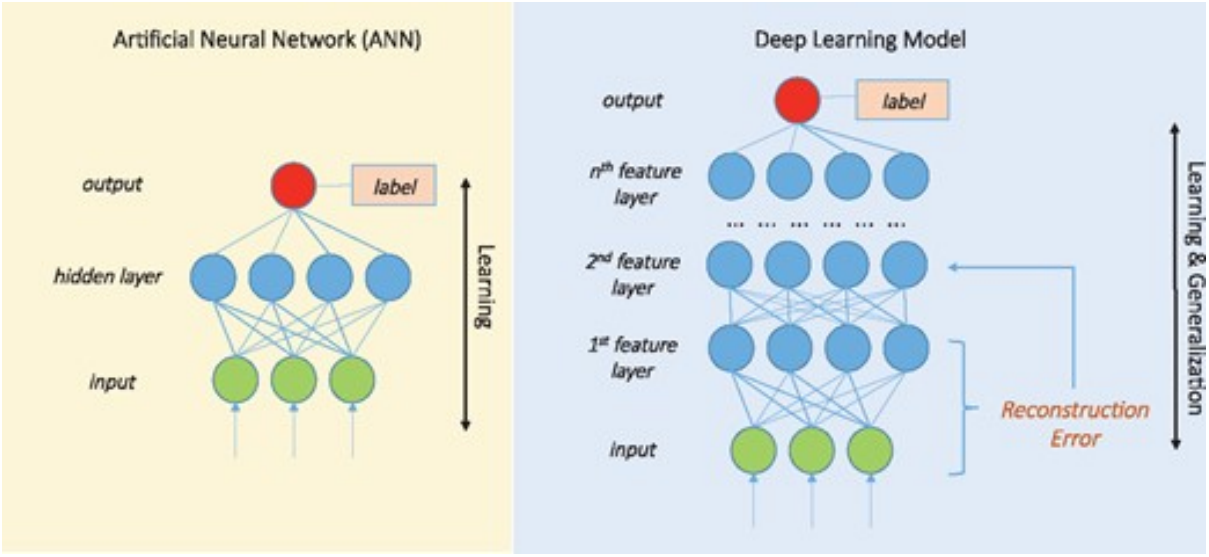

Figure 1

Comparison between ANNs and deep architectures. While ANNs are usually composed by three layers and one transformation toward the final outputs, deep learning architectures are constituted by several layers of neural networks. Layer-wise unsupervised pre-training allows deep networks to be

tuned efficiently and to extract deep structure from inputs to serve as higher-level features that are used to obtain better predictions.
Open in new tabDownload slide

The unsupervised pre-training breakthrough [23, 38], new methods to prevent overfitting [39], the use of general-purpose graphic processing units to speedup computations and the development of high-level modules to easily build neural networks (e.g. Theano [40], Caffe [41], TensorFlow [42]) allowed deep models to establish as state-of-the-art solutions for several tasks. In fact, deep learning turned out to be good at discovering intricate structures in high-dimensional data and obtained remarkable performances for object detection in images [43, 44], speech recognition [45] and natural language understanding [46] and translation [47]. Relevant clinical-ready successes have been obtained in health care as well (e.g. detection of diabetic retinopathy in retinal fundus photographs [48], classification of skin cancer [49], predicting of the sequence specificities of DNA- and RNA-binding proteins [50]), initiating the way toward a potential new generation of intelligent tools-based deep learning for real-world medical care.

Literature review

The use of deep learning for medicine is recent and not thoroughly explored. In the next sections, we will review some of the main recent literature (i.e. 32 papers) related to applications of deep models to clinical imaging, EHRs, genomics and wearable device data.

Table 1 summarizes all the papers mentioned in this literature review, in particular highlighting the type of networks and the medical data considered. To the best of our knowledge, there are no studies using deep learning to combine neither all these data sources, nor a part of them (e.g. only EHRs and clinical images, only EHRs and genomics) in a joint representation for medical analysis and prediction. A few preliminary studies evaluated the combined use of EHRs and genomics (e.g. see [9, 80]), without applying deep learning though; for this reason, they were not considered relevant to this review. The deep architectures applied to the health care domain have been mostly based

on convolutional neural networks (CNNs) [81], recurrent neural networks (RNNs) [82], Restricted Boltzmann Machines (RBMs) [83] and Autoencoders (AEs) [84]. Table 2 briefly reviews these models and provides the main ideas behind their structures.

Table 1

Summary of the articles described in the literature review with highlighted the deep learning architecture applied and the medical domain considered

Data	Author	Application	Model	Reference
Clinical imaging	Liu *et al.* (2014)	Early diagnosis of Alzheimer disease from brain MRIs	Stacked Sparse AE	[51]
	Brosch *et al.* (2013)	Manifold of brain MRIs to detect modes of variations in Alzheimer disease	RBM	[52]
	Prasoon *et al.* (2013)	Automatic segmentation of knee cartilage MRIs to predict the risk of osteoarthritis	CNN	[53]
	Yoo *et al.* (2014)	Segmentation of multiple sclerosis lesions in multi-channel 3D MRIs	RBM	[54]
	Cheng *et al.* (2016)	Diagnosis of breast nodules and lesions from ultrasound images	Stacked Denoising AE	[55]
	Gulshan *et al.* (2016)	Detection of diabetic retinopathy in retinal fundus photographs	CNN	[48]
	Esteva *et al.* (2017)	Dermatologist-level classification of skin cancer	CNN	[49]
Electronic health records	Liu *et al.* (2015)	Prediction of congestive heart failure and chronic obstructive pulmonary disease from longitudinal EHRs	CNN	[56]

Data	Author	Application	Model	Reference
	Lipton *et al.* (2015)	Diagnosis classification from clinical measurements of patients in pediatric intensive unit care	LSTM RNN	[57]
	Pham *et al.* (2016)	DeepCare: a dynamic memory model for predictive medicine based on patient history	LSTM RNN	[58]
	Miotto *et al.* (2016)	Deep Patient: an unsupervised representation of patients that can be used to predict future clinical events	Stacked Denoising AE	[59]
	Miotto *et al.* (2016)	Prediction of future diseases from the patient clinical status	Stacked Denoising AE	[60]
	Liang *et al.* (2014)	Automatically assign diagnosis to patients from their clinical status	RBM	[61]
	Tran *et al.* (2015)	Predict suicide risk of mental health patients by low-dimensional representations of the medical concepts embedded in the EHRs	RBM	[62]
	Che *et al.* (2015)	Discovering and detection of characteristic patterns of physiology in clinical time series	Stacked AE	[63]

Data	Author	Application	Model	Reference
	Lasko *et al.* (2013)	Model longitudinal sequences of serum uric acid measurements to suggest multiple population subtypes and to distinguish the uric-acid signatures of gout and acute leukemia	Stacked AE	[64]
	Choi *et al.* (2016)	Doctor AI: use the history of patients to predict diagnoses and medications for a subsequent visit	GRU RNN	[65]
	Nguyen *et al.* (2017)	Deepr: end-to-end system to predict unplanned readmission after discharge	CNN	[66]
	Razavian *et al.* (2016)	Prediction of Disease Onsets from Longitudinal Lab Tests	LSTM RNN	[67]
	Dernoncourt *et al.* (2016)	De-identification of patient clinical notes	LSTM RNN	[68]
Genomics	Zhou *et al.* (2015)	Predict chromatin marks from DNA sequences	CNN	[69]
	Kelley *et al.* (2016)	Basset: open-source platform to predict DNase I hypersensitivity across multiple cell types and to quantify the effect of	CNN	[70]

Data	Author	Application	Model	Reference
		SNVs on chromatin accessibility		
	Alipanahi *et al.* (2015)	DeepBind: predict the specificities of DNA- and RNA-binding proteins	CNN	[50]
	Angermueller *et al.* (2016)	Predict methylation states in single-cell bisulfite sequencing studies	CNN	[71]
	Koh *et al.* (2016)	Prevalence estimate for different chromatin marks	CNN	[72]
	Fakoor *et al.* (2013)	Classification of cancer from gene expression profiles	Stacked Sparse AE	[73]
	Lyons *et al.* (2014)	Prediction of protein backbones from protein sequences	Stacked Sparse AE	[74]
Mobile	Hammerla *et al.* (2016)	HAR to detect freezing of gait in PD patients	CNN/RNN	[75]
	Zhu *et al.* (2015)	Estimation of EE using wearable sensors	CNN	[76]
	Jindal *et al.* (2016)	Identification of Photoplethysmography signals for health monitoring	RBM	[77]
	Nurse *et al.* (2016)	Analysis of electroencephalogram and local field potentials signals	CNN	[78]

Data	Author	Application	Model	Reference
	Sathyanarayana *et al.* (2016)	Predict the quality of sleep from physical activity wearable data during awake time	CNN	[79]

We report 32 different papers using deep learning on clinical images, EHRs, genomics and mobile data. As it can be seen, most of the papers apply CNNs and AEs, regardless the medical domain. To the best of our knowledge, no works in the literature jointly process these different types of data (e.g. all of them, only EHRs and clinical images, only EHRs and mobile data) using deep learning for medical intelligence and prediction.

RNN = recurrent neural network; CNN = convolutional neural network; RBM = restricted Boltzmann machine; AE = autoencoder; LSTM = long short-term memory; GRU = gated recurrent unit.

Table 2

Review of the neural networks shaping the deep learning architectures applied to the health care domain in the literature

Architecture	Description
CNN	CNNs are inspired by the organization of cat's visual cortex [85]. CNNs rely on local connections and tied weights across the units followed by feature pooling (subsampling) to obtain translation invariant descriptors [52]. The basic CNN architecture consists of one convolutional and pooling layer, optionally followed by a fully connected layer for supervised prediction. In practice, CNNs are composed by > 10 convolutional and pooling layers to better model the input space. The most successful applications of CNNs were obtained in computer vision [43, 44]. CNNs usually require a large data set of labeled documents to be properly trained.
RNN	RNNs are useful to process streams of data [53]. They are composed by one network performing the same task for every element of a

Architecture	Description
	sequence, with each output value dependent on the previous computations. In the original formulation, RNNs were limited to look back only a few steps owing to vanishing and exploding gradient problems. LSTM [86] and GRU [87] networks addressed this problem by modeling the hidden state with cells that decide what to keep in (and what to erase from) memory given the previous state, the current memory and the input value. These variants are efficient at capturing long-term dependencies and led to excellent results in Natural Language Processing applications [46, 47].
RBM	A RBM is a generative stochastic model that learns a probability distribution over the input space [54]. RBMs are a variant of Boltzmann machines, with the restriction that their neurons must form a bipartite graph. Pairs of nodes from each of the two groups (i.e. visible and hidden units) can have a symmetric connection between them, but there are no connections between nodes within a group. This restriction allows for more efficient training algorithms than the general class of Boltzmann machines, which allows connections between hidden units. RBMs had success in dimensionality reduction [55] and collaborative filtering [88]. Deep learning systems obtained by stacking RBMs are called Deep Belief Networks [89].
AE	An AE is an unsupervised learning model where the target value is equal to the input [55]. AEs are composed by a decoder, which transforms the input to a latent representation, and by a decoder, which reconstructs the input from this representation. AEs are trained to minimize the reconstruction error. By constraining the dimension of the latent representation to be different from input (and consequently from the output), it is possible to discover relevant patterns in the data. AEs are mostly used for representation learning [21] and are often regularized by adding noise to the original data (i.e. denoising AEs [90]).

Clinical imaging

Following the success in computer vision, the first applications of deep learning to clinical data were on image processing, especially on the analysis of brain Magnetic Resonance Imaging (MRI) scans to predict Alzheimer

disease and its variations [51, 52]. In other medical domains, CNNs were used to infer a hierarchical representation of low-field knee MRI scans to automatically segment cartilage and predict the risk of osteoarthritis [53]. Despite using 2D images, this approach obtained better results than a state-of-the-art method using manually selected 3D multi-scale features. Deep learning was also applied to segment multiple sclerosis lesions in multi-channel 3D MRI [54] and for the differential diagnosis of benign and malignant breast nodules from ultrasound images [55]. More recently, Gulshan *et al*. [48] used CNNs to identify diabetic retinopathy in retinal fundus photographs, obtaining high sensitivity and specificity over about 10 000 test images with respect to certified ophthalmologist annotations. CNNs also obtained performances on par with 21 board-certified dermatologists on classifying biopsy-proven clinical images of different types of skin cancer (keratinocyte carcinomas versus benign seborrheic keratoses and malignant melanomas versus benign nevi) over a large data set of 130 000 images (1942 biopsy-labeled test images) [49].

Electronic health records

More recently deep learning has been applied to process aggregated EHRs, including both structured (e.g. diagnosis, medications, laboratory tests) and unstructured (e.g. free-text clinical notes) data. The greatest part of this literature processed the EHRs of a health care system with a deep architecture for a specific, usually supervised, predictive clinical task. In particular, a common approach is to show that deep learning obtains better results than conventional machine learning models with respect to certain metrics, such as Area Under the Receiver Operating Characteristic Curve, accuracy and F-score [91]. In this scenario, while most papers present end-to-end supervised networks, some works also propose unsupervised models to derive latent patient representations, which are then evaluated using shallow classifiers (e.g. random forests, logistic regression).

Several works applied deep learning to predict diseases from the patient clinical status. Liu *et al*. [56] used a four-layer CNN to predict congestive heart

failure and chronic obstructive pulmonary disease and showed significant advantages over the baselines. RNNs with long short-term memory (LSTM) hidden units, pooling and word embedding were used in DeepCare [58], an end-to-end deep dynamic network that infers current illness states and predicts future medical outcomes. The authors also proposed to moderate the LSTM unit with a decay effect to handle irregular timed events (which are typical in longitudinal EHRs). Moreover, they incorporated medical interventions in the model to dynamically shape the predictions. DeepCare was evaluated for disease progression modeling, intervention recommendation and future risk prediction on diabetes and mental health patient cohorts. RNNs with gated recurrent unit (GRU) were used by Choi *et al*. [65] to develop Doctor AI, an end-to-end model that uses patient history to predict diagnoses and medications for subsequent encounters. The evaluation showed significantly higher recall than shallow baselines and good generalizability by adapting the resulting model from one institution to another without losing substantial accuracy. Differently, Miotto *et al*. [59] proposed to learn deep patient representations from the EHRs using a three-layer Stacked Denoising Autoencoder (SDA). They applied this novel representation on disease risk prediction using random forest as classifiers. The evaluation was performed on 76 214 patients comprising 78 diseases from diverse clinical domains and temporal windows (up to a 1 year). The results showed that the deep representation leads to significantly better predictions than using raw EHRs or conventional representation learning algorithms (e.g. Principal Component Analysis (PCA), k-means). Moreover, they also showed that results significantly improve when adding a logistic regression layer on top of the last AE to fine-tune the entire supervised network [60]. Similarly, Liang *et al*. [61] used RBMs to learn representations from EHRs that revealed novel concepts and demonstrated better prediction accuracy on a number of diseases.

Deep learning was also applied to model continuous time signals, such as laboratory results, toward the automatic identification of specific phenotypes. For example, Lipton *et al*. [57] used RNNs with LSTM to recognize patterns in multivariate time series of clinical measurements. Specifically, they trained a

model to classify 128 diagnoses from 13 frequently but irregularly sampled clinical measurements from patients in pediatric intensive unit care. The results showed significant improvements with respect to several strong baselines, including multilayer perceptron trained on hand-engineered features. Che *et al*. [63] used SDAs regularized with a prior knowledge based on ICD-9s for detecting characteristic patterns of physiology in clinical time series. Lasko *et al*. [64] used a two-layer stacked AE (without regularization) to model longitudinal sequences of serum uric acid measurements to distinguish the uric-acid signatures of gout and acute leukemia. Razavian *et al*. [67] evaluated CNNs and RNNs with LSTM units to predict disease onset from laboratory test measures alone, showing better performances than logistic regression with hand-engineered, clinically relevant features.

Neural language deep models were also applied to EHRs, in particular to learn embedded representations of medical concepts, such as diseases, medications and laboratory tests, that could be used for analysis and prediction [92]. As an example, Tran *et al*. [62] used RBMs to learn abstractions of ICD-10 codes on a cohort of 7578 mental health patients to predict suicide risk. A deep architecture based on RNNs also obtained promising results in removing protected health information from clinical notes to leverage the automatic de-identification of free-text patient summaries [68].

The prediction of unplanned patient readmissions after discharge recently received attention as well. In this domain, Nguyen *et al*. [66] proposed Deepr, an end-to-end architecture based on CNNs, which detects and combines clinical motifs in the longitudinal patient EHRs to stratify medical risks. Deepr performed well in predicting readmission within 6 months and was able to detect meaningful and interpretable clinical patterns.

Genomics

Deep learning in high-throughput biology is used to capture the internal structure of increasingly larger and high-dimensional data sets (e.g. DNA sequencing, RNA measurements). Deep models enable the discovery of high-level features, improving performances over traditional models, increasing

interpretability and providing additional understanding about the structure of the biological data. Different works have been proposed in the literature. Here we review the general ideas and refer the reader to [93–96] for more comprehensive reviews.

The first applications of neural networks in genomics replaced conventional machine learning with deep architectures, without changing the input features. For example, Xiong *et al*. [97] used a fully connected feed-forward neural network to predict the splicing activity of individual exons. The model was trained using >1000 predefined features extracted from the candidate exon and adjacent introns. This method obtained higher prediction accuracy of splicing activity compared with simpler approaches, and was able to identify rare mutations implicated in splicing misregulation.

More recent works apply CNNs directly on the raw DNA sequence, without the need to define features a priori (e.g. [50, 69, 70]). CNNs use less parameters than a fully connected network by computing convolution on small regions of the input space and by sharing parameters between regions. This allowed training the models on larger sequence windows of DNAs, improving the detection of the relevant patterns. For example, Alipanahi *et al*. [50] proposed DeepBind, a deep architecture based on CNNs that predicts specificities of DNA- and RNA-binding proteins. In the reported experiment, DeepBind was able to recover known and novel sequence of motifs, quantify the effect of sequence alterations and identify functional single nucleotide variations (SNVs). Zhou and Troyanskaya [69] used CNNs to predict chromatin marks from DNA sequence. Similarly, Kelley *et al*. [70] developed Basset, an open-source framework to predict DNase I hypersensitivity across multiple cell types and to quantify the effect of SNVs on chromatin accessibility. CNNs were also used by Angermueller *et al*. [71] to predict DNA methylation states in single-cell bisulfite sequencing studies and, more recently, by Koh *et al*. [72] to denoise genome-wide chromatin immunoprecipitation followed by sequencing data to obtain a more accurate prevalence estimate for different chromatin marks.

While CNNs are the most widely used architectures to extract features from fixed-size DNA sequence windows, other deep architectures have been proposed as well. For example, sparse AEs were applied to classify cancer cases from gene expression profiles or to predict protein backbones [74]. Deep neural networks also enabled researchers to significantly improve the state-of-the-art drug discovery pipeline for genomic medicine [98].

Mobile

Sensor-equipped smartphones and wearables are transforming a variety of mobile apps, including health monitoring [99]. As the difference between consumer health wearables and medical devices begins to soften, it is now possible for a single wearable device to monitor a range of medical risk factors. Potentially, these devices could give patients direct access to personal analytics that can contribute to their health, facilitate preventive care and aid in the management of ongoing illness [100]. Deep learning is considered to be a key element in analyzing this new type of data. However, only a few recent works used deep models within the health care sensing domain, mostly owing to hardware limitations. In fact, running an efficient and reliable deep architecture on a mobile device to process noisy and complex sensor data is still a challenging task that is likely to drain the device resources [101]. Several studies investigated solutions to overcome such hardware limitations. As an example, Lane and Georgiev [102] proposed a low-power deep neural network inference engine that exploited both Central Processing Unit (CPU) and Digital Signal Processor (DSP) of the mobile device, without leading to any major overburden of the hardware. They also proposed DeepX, a software accelerator capable of lowering the device resources required by deep learning that currently act as a severe bottleneck to mobile adoption. This architecture enabled large-scale deep learning to execute efficiently on mobile devices and significantly outperformed cloud-based off-loading solutions [103].

We did not find any relevant study applying deep learning on commercial wearable devices for health monitoring. However, a few works processed data

from phones and medical monitor devices. In particular, relevant studies based on deep learning were done on Human Activity Recognition (HAR). While not directly exploring a medical application, many studies argue that the accurate predictions obtained by deep models on HAR can leverage clinical applications as well. In the health care domain, Hammerla *et al*. [75] evaluated CNNs and RNNs with LSTM to predict the freezing of gait in Parkinson disease (PD) patients. Freezing is a common motor complication in PD, where affected individuals struggle to initiate movements such as walking. Results based on accelerometer data from above the ankle, above the knee and on the trunk of 10 patients showed that RNNs obtained the best results, with a significantly large improvement over the other models, including CNNs. While the size of this data set was small, this study highlights the potential of deep learning in processing activity recognition measures for clinical use. Zhu *et al*. [76] obtained promising results in predicting Energy Expenditure (EE) from triaxial accelerometer and heart rate sensor data during ambulatory activities. EE is considered important in tracking personal activity and preventing chronic diseases such as obesity, diabetes and cardiovascular diseases. They used CNNs and significantly improved performances over regression and a shallow neural network.

In other clinical domains, deep learning, in particular CNNs and RBMs, improved over conventional machine learning in analyzing portable neurophysiological signals such as Electroencephalogram, Local Field Potentials and Photoplethysmography [77, 78]. Differently, Sathyanarayana *et al*. [79] applied deep learning to predict poor or good sleep using actigraphy measurements of the physical activity of patients during awake time. In particular, by using a data set of 92 adolescents and one full week of monitored data, they showed that CNNs were able to obtain the highest specificity and sensitivity, with results 46% better than logistic regression.

Challenges and opportunities

Despite the promising results obtained using deep architectures, there remain several unsolved challenges facing the clinical application of deep learning to health care. In particular, we highlight the following key issues:

- ***Datavolume*:** Deep learning refers to a set of highly intensive computational models. One typical example is fully connected multi-layer neural networks, where tons of network parameters need to be estimated properly. The basis to achieve this goal is the availability of huge amount of data. In fact, while there are no hard guidelines about the minimum number of training documents, a general rule of thumb is to have at least about 10× the number of samples as parameters in the network. This is also one of the reasons why deep learning is so successful in domains where huge amount of data can be easily collected (e.g. computer vision, speech, natural language). However, health care is a different domain; in fact, we only have approximately 7.5 billion people all over the world (as per September 2016), with a great part not having access to primary health care. Consequently, we cannot get as many patients as we want to train a comprehensive deep learning model. Moreover, understanding diseases and their variability is much more complicated than other tasks, such as image or speech recognition. Consequently, from a big data perspective, the amount of medical data that is needed to train an effective and robust deep learning model would be much more comparing with other media.
- ***Dataquality*:** Unlike other domains where the data are clean and well-structured, health care data are highly heterogeneous, ambiguous, noisy and incomplete. Training a good deep learning model with such massive and variegate data sets is challenging and needs to consider several issues, such as data sparsity, redundancy and missing values.
- ***Temporality*:** The diseases are always progressing and changing over time in a nondeterministic way. However, many existing deep learning models, including those already proposed in the medical domain, assume static vector-based inputs, which cannot handle the time factor in a natural way. Designing deep learning approaches that can handle

temporal health care data is an important aspect that will require the development of novel solutions.

- ***Domaincomplexity*:** Different from other application domains (e.g. image and speech analysis), the problems in biomedicine and health care are more complicated. The diseases are highly heterogeneous and for most of the diseases there is still no complete knowledge on their causes and how they progress. Moreover, the number of patients is usually limited in a practical clinical scenario and we cannot ask for as many patients as we want.
- ***Interpretability*:** Although deep learning models have been successful in quite a few application domains, they are often treated as black boxes. While this might not be a problem in other more deterministic domains such as image annotation (because the end user can objectively validate the tags assigned to the images), in health care, not only the quantitative algorithmic performance is important, but also the reason why the algorithms works is relevant. In fact, such model interpretability (i.e. providing which phenotypes are driving the predictions) is crucial for convincing the medical professionals about the actions recommended from the predictive system (e.g. prescription of a specific medication, potential high risk of developing a certain disease).

All these challenges introduce several opportunities and future research possibilities to improve the field. Therefore, with all of them in mind, we point out the following directions, which we believe would be promising for the future of deep learning in health care.

- ***Featureenrichment*:** Because of the limited amount of patients in the world, we should capture as many features as possible to characterize each patient and find novel methods to jointly process them. The data sources for generating those features need to include, but not to be limited to, EHRs, social media (e.g. there are prior research leveraging patient-reported information on social media for pharmacovigilance [104, 105]), wearable devices, environments, surveys, online communities, genome profiles, omics data such as proteome and so on. The effective integration of such highly heterogeneous data and how to

use them in a deep learning model would be an important and challenging research topic. In fact, to the best of our knowledge, the literature does not provide any study that attempts to combine different types of medical data sources using deep learning. A potential solution in this domain could exploit the hierarchical nature of deep learning and process separately every data source with the appropriate deep model, and stack the resulting representations in a joint model toward a holistic abstraction of the patient data (e.g. using layers of AEs or deep Bayesian networks).

- ***Federatedinference*:** Each clinical institution possesses its own patient population. Building a deep learning model by leveraging the patients from different sites without leaking their sensitive information becomes a crucial problem in this setting. Consequently learning deep model in this federated setting in a secure way will be another important research topic, which will interface with other mathematical domains, such as cryptography (e.g. homomorphic encryption [106] and secure multiparty computation [107]).
- ***Modelprivacy***: Privacy is an important concern in scaling up deep learning (e.g. through cloud computing services). In fact, a recent work by Tramèr *et al*. [108] has demonstrated the vulnerability of Machine Learning (ML)-as-a-service (i.e. 'predictive analytics') on a set of common models including deep neural networks. The attack abides all authentication or access-control mechanisms but infers parameters or training data through exposed Application Program Interface (APIs), which breaks the model and personal privacy. This issue is well known to the privacy community, and researchers have developed a principled framework called 'differential privacy' [109, 110] to ensure the indistinguishability of individual samples in training data through their functional outputs [111]. However, naive approaches might render outputs useless or cannot provide sufficient protection [22], which makes the development of practically useful differential privacy solutions nontrivial. For example, Chaudhuri *et al*. [112] developed differential private methods to protect the parameters trained for the logistic regression model. Preserving the privacy of deep learning

models is even more challenging, as there are more parameters to be safeguarded and several recent works have pushed the fronts in this area [113–115]. Yet, considering all the personal information likely to be processed by deep models in clinical applications, the deployment of intelligent tools for next-generation health care needs to consider these risks and attempt to implement a differential privacy standard.

- ***Incorporatingexpertknowledge*:** The existing expert knowledge for medical problems is invaluable for health care problems. Because of the limited amount of medical data and their various quality problems, incorporating the expert knowledge into the deep learning process to guide it toward the right direction is an important research topic. For example, online medical encyclopedia and PubMed abstracts should be mined to extract reliable content that can be included in the deep architecture to leverage the overall performances of the systems. Also semi-supervised learning, an effective scheme to learn from large amount of unlabeled samples with only a few labeled samples, would be of great potential because of its capability of leveraging both labeled (which encodes the knowledge) and unlabeled samples [105].
- ***Temporalmodeling*:** Considering that the time factor is important in all kinds of health care-related problems, in particular in those involving EHRs and monitoring devices, training a time-sensitive deep learning model is critical for a better understanding of the patient condition and for providing timely clinical decision support. Thus, temporal deep learning is crucial for solving health care problems (as already shown in some of the early studies reported in the literature review). To this aim, we expect that RNNs as well as architectures coupled with memory (e.g. [86]) and attention mechanisms (e.g. [116]) will play a more significant role toward better clinical deep architectures.
- ***Interpretablemodeling*:** Model performance and interpretability are equally important for health care problems. Clinicians are unlikely to adopt a system they cannot understand. Deep learning models are popular because of their superior performance. Yet, how to explain the results obtained by these models and how to make them more understandable is of key importance toward the development of

trustable and reliable systems. In our opinion, research directions will include both algorithms to explain the deep models (i.e. what drives the hidden units of the networks to turn on/off along the process—see e.g. [117]) as well as methods to support the networks with existing tools that explain the predictions of data-driven systems (e.g. see [118]).

Applications

Deep learning methods are powerful tools that complement traditional machine learning and allow computers to learn from the data, so that they can come up with ways to create smarter applications. These approaches have already been used in a number of applications, especially for computer vision and natural language processing. All the results available in the literature illustrate the capabilities of deep learning for health care data analysis as well. In fact, processing medical data with multi-layer neural networks increased the predictive power for several specific applications in different clinical domains. Additionally, because of their hierarchical learning structure, deep architectures have the potential to integrate diverse data sets across heterogeneous data types and provide greater generalization given the focus on representation learning and not simply on classification accuracy.

Consequently, we believe that deep learning can open the way toward the next generation of predictive health care systems that can (i) scale to include many millions to billions of patient records and (ii) use a single, distributed patient representation to effectively support clinicians in their daily activities—rather than multiple systems working with different patient representations and data. Ideally, this representation would join all the different data sources, including EHRs, genomics, environment, wearables, social activities and so on, toward a holistic and comprehensive description of an individual status. In this scenario, the deep learning framework would be deployed into a health care platform (e.g. a hospital EHR system) and the models would be constantly updated to follow the changes in the patient population.

Such deep representations can then be used to leverage clinician activities in different domains and applications, such as disease risk prediction, personalized prescriptions, treatment recommendations, clinical trial recruitment as well as research and data analysis. As an example, Wang *et al*. recently won the Parkinson's Progression Marker's Initiative data challenge on subtyping Parkinson's disease using a temporal deep learning approach [119]. In fact, because Parkinson's disease is highly progressive, the traditional vector or matrix-based approach may not be optimal, as it is unable to accurately capture the disease progression patterns, as the entries in those vectors/matrices are typically aggregated over time. Consequently, the authors used the LSTM RNN model and identified three interesting subtypes for Parkinson's disease, wherein each subtype demonstrates common disease progression trends. We believe that this work shows the great potential of deep learning models in real-world health care problems and how it could lead to more reliable and robust automatic systems in the near future.

Last, more broadly, deep learning can serve as a guiding principle to organize both hypothesis-driven research and exploratory investigation in clinical domains (e.g. clustering, visualization of patient cohorts, stratification of disease populations). For this potential to be realized, statistical and medical tasks must be integrated at all levels, including study design, experiment planning, model building and refinement and data interpretation.