Deep Learning in Omics Data Analysis and Precision Medicine

Jordi Martorell-Marugán[1] • Siham Tabik[2] • Yassir Benhammou[2] • Coral del Val[2] • Igor Zwir[2] • Francisco Herrera[2] • Pedro Carmona-Sáez[1]

[1]GENYO, Centre for Genomics and Oncological Research: Pfizer, University of Granada, Andalusian Regional Government, Granada, Spain; [2]Department of Computer Science and Artificial Intelligence, University of Granada, Granada, Spain

Authors for correspondence: Francisco Herrera, Department of Computer Science and Artificial Intelligence, University of Granada, Granada, Spain. Email: herrera@decsai.ugr.es; Pedro Carmona-Sáez, GENYO, Centre for Genomics and Oncological Research: Pfizer, University of Granada, Andalusian Regional Government, PTS Granada, Avenida de la Ilustración 114 – 18016, Granada, Spain. Email: pedro.carmona@genyo.es

Doi: http://dx.doi.org/10.15586/computationalbiology.2019.ch3

Abstract: The rise of omics techniques has resulted in an explosion of molecular data in modern biomedical research. Together with information from medical images and clinical data, the field of omics has driven the implementation of personalized medicine. Biomedical and omics datasets are complex and heterogeneous, and extracting meaningful knowledge from this vast amount of information is by far the most important challenge for bioinformatics and machine learning researchers. In this context, there is an increasing interest in the potential of deep learning (DL) methods to create predictive models and to identify complex patterns from these large datasets. This chapter provides an overview of the main applications of DL methods in biomedical research, with focus on omics data analysis and precision medicine applications. DL algorithms and the most popular architectures are introduced first. This is followed by a review of some of the main applications and problems approached by DL in omics data and medical image analysis. Finally, implementations for improving the diagnosis, treatment, and classification of complex diseases are discussed.

Keywords: artificial neural networks; biomedical informatics; deep learning; omics data analysis; precision medicine

INTRODUCTION

The amount of available biological data has increased exponentially since the emergence of high-throughput technologies such as microarrays and next-generation sequencing (1), introducing biology to the big data era. These methods initiated the so-called omics revolution, where large amounts of omics data providing global information about different properties of genes, proteins or biomolecules can be generated within a short period of time in a cost-effective way. These methods have revolutionized biomedical research by providing a more comprehensive understanding of the biological system under study and the molecular mechanisms underlying disease development. The generation of such a large amount of data in biomedicine requires the application of advanced informatics techniques in order to extract new insights and expand our knowledge about diseases, improve diagnosis, and design personalized treatments. In this context, DL algorithms have become one of the most promising methods in the area (2).

DL is a subset of machine learning (ML) algorithms characterized by the use of artificial neural networks (ANN). ANNs are inspired by biological neural networks in a sense that they are formed by interconnected artificial neurons, which receive an input, apply a transformation to the data, and return an output (which can be an input for another neuron). DL is gaining popularity as a powerful approach that can encode and learn from heterogeneous and complex data, in both supervised and unsupervised settings. DL methods have achieved considerable improvements in classical artificial intelligence challenges like language processing, speech recognition, and image recognition (3). In the context of biomedical research, DL methods have drawn the attention of many researchers, and there is an increasing number of applications in omics data analysis. Omics data analysis is frequently impeded by low signal to noise ratios, datasets with large number of variables and relatively small number of samples or large analytical variance. In this context, DL techniques have already over-performed previous methods in terms of sensitivity, specificity and efficiency (4). In addition, DL algorithms not only have the challenge of analyzing each kind of data separately but also have the challenge of integrating different omics layers or even other sources of information such as medical images or clinical health records. This big data analysis and integration is fueling the implementation of personalized medicine approaches allowing early detection and classification of diseases or personalized therapies for each patient depending on their biochemical background. This chapter reviews the main applications of DL methods to omics data analysis with a focus on the types of analysis, challenges, and opportunities in precision medicine.

DEEP LEARNING METHODS

DL networks are a class of ML algorithms whose aim is to determine a mathematical function f that maps a number of inputs, x, to their corresponding outputs, y, such as $y = f(x)$. A simple feedforward network $y = f(x;w) = LN(LN\text{-}1(….L1(x)))$ is

defined as a composition of N nonlinear transformations $Li(1<=i<=N)$ where each function Li corresponds to a hidden layer activation, and w is the learnable weight contained in all filter bank layers that are updated during the training.

Under the supervised learning approach, the training of these networks is often done iteratively in which a set of training data, also called batch, with their ground truth labels are provided to the network as input. After a feed-forward of this batch through the network's layers, the output layer computes the loss function as the difference between the calculated prediction and the correct response. After computing the loss function, all layers' weights are updated so that the loss error of the next iteration is minimized. This weight-tuning operation is performed using a back-propagation algorithm (5) where the error function gradient is propagated in the opposite direction through the network after a batch of feedforwards to adjust filter banks, thereby learning the value of the parameter w that results in the best function approximation.

Deep feedforward neural network (DFF)

DFFs, also called multilayer perceptrons, constitute the simplest DL architecture. In these models, the input information x flows to its corresponding output y through an intermediate function f being evaluated and learned inside the neural network layers. These models are called feedforward since there are no feedback connections in which outputs of the model are fed back into themselves.

Convolutional neural network (CNN)

CNNs are the most adequate DNNs to deal with high multi-dimensional data like medical images. In medical imaging applications, CNNs act like a long dimensionality reduction process, binding input images to their classification scores outputs (e.g., disease or healthy patient). The building block layers of a CNN are convolutional layer, pooling layer, and fully connected layer. Generally, DL CNNs are applied with a transfer learning strategy to enhance their performance in dealing with relatively small datasets. Transfer learning consists of transferring prior learned knowledge from a source domain into a target domain. This approach is carried out by using one of the well-known CNNs pre-trained on a large dataset such as ImageNet (6), either for further training on the new data or to reuse it as a features extractor (7). Rawat and Wang (8) wrote a more comprehensive review on CNNs history and their architectures. Some of the most influential CNNs are summarized in Table 1.

Recurrent neural network (RNN)

RNNs are neural networks used especially for sequential data in a way that the reached output decision at time step $t - 1$ affects the decision which will be reached one moment later at time step t. These networks have two input sources, the present and the recent past, which are combined to determine how they respond to new data.

TABLE 1		Summary of some of the most influential CNNs	
CNN	Layers	Parameters	Comments
LeNet	5	60 000	First CNN to be trained on a large dataset (5, 87)
AlexNet	7	60 million	Variation of LeNet. First CNN model to win the prestigious ILSVRC[a] in 2012 (88).
GoogLeNet	22	4 million	Winner of ILSVRC[a] in 2014 (89). The main contribution is the inception module which is composed of different parallel small convolutions.
VGGNet	16	-	Initially the runner-up in ILSVRC 2014 behind GoogleNet (90)
ResNet	18, 34, 50, 101 or 152	11.7 million – 60.2 million	To overcome the gradient vanishing issue, ResNet authors (91) proposed using a residual function $F(x) = H(x) - x$, where $H(x)$ is the standard mapping function that we want to learn with an input x through few stacked non-linear layers. By reformulating it as $H(x) = F(x) + x$, where $F(x)$ and x represent the stacked non-linear layers and the identity function, respectively. Based on their hypothesis, it is better to optimize the reformulated residual mapping function $F(x)$ than optimizing the original mapping $H(x)$.
DenseNet	121, 161, 169 or 201	8 million – 20 million	Presented in (92) to take advantage from previous findings regarding CNN's depth increasing and identity shortcut connections. The specificity of this new network architecture is that each layer is connected to all its previous and next layers.

[a]Large Scale Visual Recognition Challenge.

Long-/short-term memory (LSTM)

The main drawback of RNNs is the vanishing gradient problem. To address this issue, a variant of RNN called LSTM was proposed. LSTMs aim to preserve the error that can be back-propagated through time and layers. In fact, they allow recurrent nets to continue to learn over many steps by maintaining a more constant error. LSTMs contain information outside the normal flow of the recurrent network in a gated cell. Information can be stored in, written to or read from a cell, much like data in a computer's memory.

Deep belief network (DBN)

To learn deep features representation, a DBN (9) is built with a concatenation of several restricted Boltzmann machine (RBM) stacked on each other. RBM is the core component of DBN models (10), being a generative stochastic model that can

be used either for unsupervised or supervised learning. It is composed of two layers, an input visible layer and an adjacent hidden layer trained with the aim to learn a probability distribution in the input set. Nevertheless, unlike original Boltzmann machine (11), intra-connections between hidden–hidden or visible–visible layers in an RBM are disjointed forming a bipartite graph.

Autoencoder (AE)

Generally, AEs act in an unsupervised manner trying to learn a distribution of a given dataset (12) and are often used as a dimensionality reduction network (13). AEs try to learn a mapping function $M\ w,\ b\ (x) = x' \approx x$ throughout stacked hidden layers mapping an input data x to its similar identity x' Generally, an AE is composed of an encoder and a decoder. The first one is trying to learn a set of low-dimensional representation features z, while the second is trying to reconstruct a similar copy of x using only learned features z. A special case of AEs is sparse autoencoder (SAE) (14), where sparsity is introduced into the hidden units by making the number of nodes in the hidden layer z bigger than in the input layer x. When several SAEs with only their encoding parts are stacked on each other, we obtain a stacked sparse autoencoder (SSAE) which is often trained in a bottom–up greedy fashion to learn deep feature representation from the data (14).

DEEP LEARNING APPLICATIONS IN OMICS DATA ANALYSIS

DL algorithms are specially suitable to analyze complex, heterogeneous, and high-dimensional data such as omics datasets (15). This section reviews some cases of omics data analyses in which DL methods have provided significant insights, and the next section provides an overview of some of the main applications in the context of precision medicine, such as biomarker discovery for disease classification. A summary of the main applications is provided in Figure 1.

Genomics and sequence analysis

Genomics uses a set of techniques to analyze DNA sequences for studying the structure and function of genomes, gene regulation, and genetic alterations that can be associated with several diseases. During the last years, DL methods have been applied to genomics data to address several questions. For instance, Poplin et al. developed a method to detect single-nucleotide polymorphisms (SNP) and indels by applying CNNs, which outperformed previous tools (16). In this context, other approaches have applied ResNets (17), DFF (18) or CNN (19) to predict the pathogenic consequences of genetic variants. In addition, Xie et al. applied DFFs and SAEs to predict the effect of genetic variants on gene expression (20). In the field of functional genomics, DL algorithms have been applied to predict enhancers' sequences and regulatory motifs in the genome (21–25) from heterogeneous sources of data (histone modifications, chromatin accessibility and so on). Wang et al. applied CNNs to quantify transcription factor (TF)-DNA binding affinities (26). Oubounyt et al. combined a CNN and an
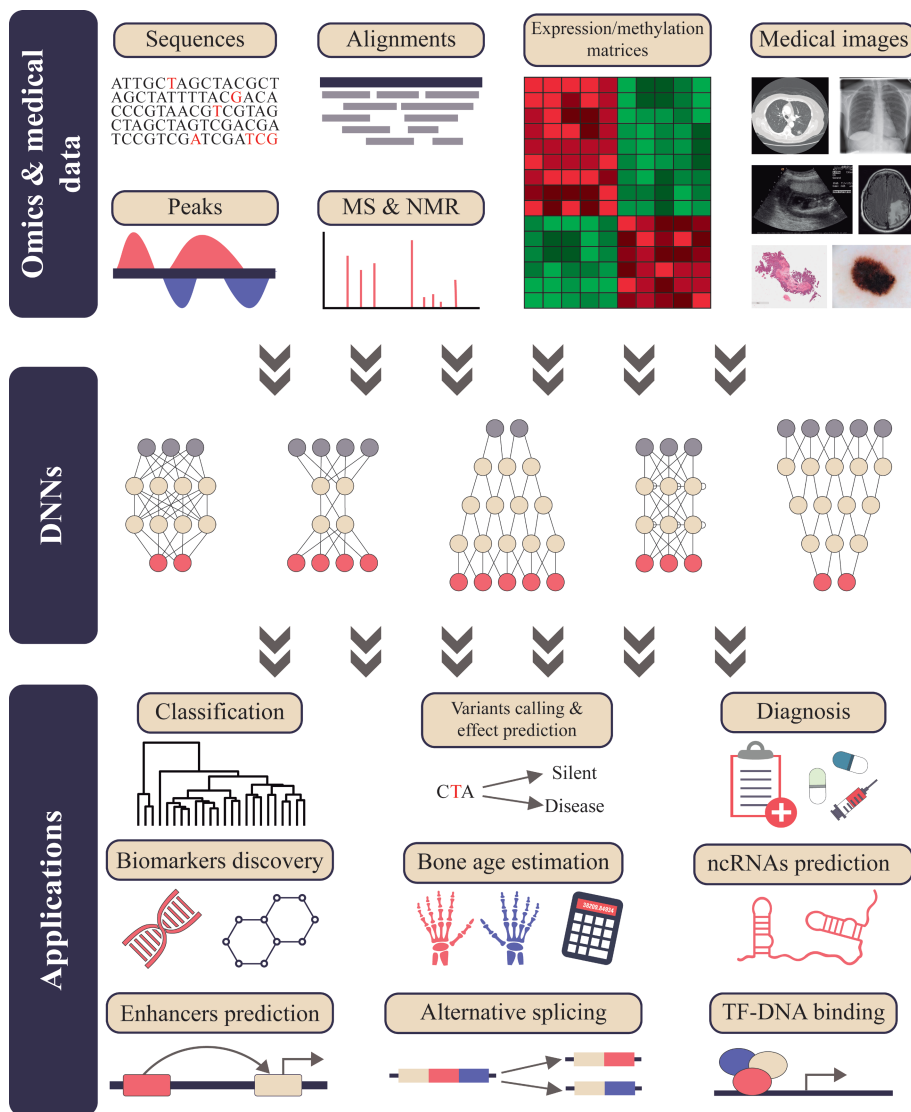
Figure 1 DNNs have been applied to several biological data types. At the top, there are the different types of data. At the middle, there are some examples of DNNs structures. At the bottom, there are some of the main applications achieved with these methodologies. Source of medical images: TCIA (93) for MRI and CT; Chest X-Ray database (94) for X-Ray; MedPix® (https://medpix.nlm.nih.gov) for US; TCGA (58) for the histopathological image and ISIC (https://www.isic-archive.com) for the skin lesion. Some graphical elements were downloaded from Stockio (https://www.stockio.com/) and Freepik (https://www.freepik.com/).

LSTM to predict promoter sequences in genes (27). DL algorithms have also helped to identify splice junctions through CNN (28).

Genome-wide association studies

Another important field of application of genomics techniques is the screening of genetic regions (loci) that associate with diseases/phenotypes, what is termed genome-wide association studies (GWAS). In this context, GWAS analyses identify SNPs in genomic locations that are incorporated into risk prediction models traditionally analyzed by polygenic risk scores (29). However, this method presents certain limitations such as the inability to reduce the missing heritability, dealing with epistasis, assumption of a global linear association model or the replication of results in different samples (30).

As an alternative, supervised learning algorithm, especially DL models, is gaining relevance in this field. Promising results have been shown by Montaez et al. (31) that developed a DL framework for the classification of obesity as a binary phenotype. However, the predictive capacity of these genetic markers is weak because it is based on single locus. More recently, Fergus et al. (32) modeled the epistatic effects of SNPs using SSAEs to classify term and preterm births observations in African-American women. Although it shows a good performance in classification and the capture of loci interactions, it suffers from the common black-box problem. The selected SNPs loose the GWAS context making it very difficult to evaluate their contribution to the phenotype. A different approach is the one proposed by PGMRA (33), a deep unsupervised and data-driven ML method designed for fusing genotypic–phenotypic analysis in a semi-supervised fashion including unsupervised non-negative matrix factorization (NMF) method as an AE (13), multiobjective optimization and pooling, interpretable association of types of knowledge, and labeling the associations. Each layer has its own learning process and constitutes the input of the next layer. The results from PGMRA are interpretable and have been able to decrease the missing heritability and identify the epistatic sets of markers that are composed of the genotypic–phenotypic architecture of a disease or trait (34).

Transcriptomics

Transcriptomics quantifies the expression level of all RNA transcripts that are produced in a cell. Transcriptomics raw data are usually processed to generate expression matrices containing an estimate of expression level of each gene or transcript across several samples and conditions, which are typically the input of DL methods. There is a broad range of transcriptomics applications in which DL has been successfully applied. For example, one of the main goals of gene expression data is the analysis of alternative splicing (i.e., the synthesis of different transcripts isoforms from the same gene). In this context, Zhang et al. notably achieved to analyze differential splicing between different samples using RNA-seq data and combining a DNN and a Bayesian statistical model (35). On the other hand, CNNs have been applied to identify actual splice junctions from false positives generated during RNA-seq reads alignment (36). In addition, Jha et al. proposed a model to integrate RNA-seq and CLIP-seq data in order to improve the study of alternative splicing (37).

Another major research focus in transcriptomics is the prediction of other types of RNAs, such as non-coding RNAs (ncRNAs), and the characterization of their expression. In this context, Hill et al. proposed an RNN to differentiate between coding and non-coding RNAs (38), demonstrating the capability of their algorithm to identify ncRNAs without providing their model with previous knowledge. Tripathi et al. developed a method to detect long ncRNAs (lncRNAs) (39). They reached a remarkable 99% accuracy rate applying a DFF to reference databases. Long intergenic ncRNAs (lincRNAs), a type of lncRNAs which are transcribed in intergenic regions, have been also successfully predicted feeding an AE with previous knowledge about lincRNAs (40).

Epigenomics

Epigenomic studies identify modifications in DNA that comprise markers that can potentially alter gene expression without modifying the DNA sequence itself. There are several epigenetic markers such as DNA methylation, histone modification, and specifically positioned nucleosomes. DNA methylation perhaps is the most studied epigenetic modification. DNA methylation studies generate methylation matrices that, like gene expression matrices, can be used for biomarker discovery or disease classification problems. In this context, DL methods have been used to accurately predict the sequences recognized by DNA- and RNA-binding proteins using CNNs (41). A key advantage of this method is the capability to integrate data from different technologies used in epigenomics studies, like chromatin immunoprecipitation (ChIP)-seq or cross-linking immunoprecipitation (CLIP)-seq. DNase I sequencing data have been also used for predicting the three-dimensional chromatin state in a cell using CNN (42). On the other hand, Wang et al. accurately predicted DNA methylation state feeding SAEs with sequence and Hi-C data (43). Histone modifications, similar to DNA methylation, do not affect DNA sequence but can modify its availability to the transcriptional machinery. Using CNNs, Yin et al. designed an algorithm to predict these histone modifications by integrating sequence and DNase data (44). In addition, Singh et al. used a CNN to infer gene expression from histone modifications data (45), while Sekhon et al. used a LSTM to predict differential gene expression, also from histone modifications data (46).

Proteomics and metabolomics

Proteomics comprises a set of techniques that can be used to quantify expression levels, post-translational modifications or localization of proteins in a cell or a biological sample. Metabolomics is the study of a complete metabolome, which are small molecules that participate in general metabolic reactions. The technologies used by these omics-streams are, among others, mass spectrometry (MS) or nuclear magnetic resonance (NMR), and the first challenge for researchers in this field is to assign raw instrumental signals to proteins or metabolites.

In proteomics, the most common experimental strategy is to split proteins into short amino acid chains (peptides) and to analyze these peptides in an MS. The MS output signals are compared to peptide profiles stored in public or proprietary databases to identify them. However, these databases are still incomplete

and inaccurate. In this context, Zhou et al. developed a software that uses a LSTM network to predict peptide MS/MS spectra (47). Knowing peptide spectra a priori facilitates the task of assigning MS/MS spectra to peptides comparing them to the theoretical spectra. Another proteomics application is de novo peptide sequencing, which is essential for proteins characterization. In this field, Tran et al. surpassed previous software combining CNN and LSTMs networks to effectively accomplish such a difficult task (48). Once the collection of peptides has been sequenced in a proteomics sample, the next challenge is to identify the proteins of origin of such peptides. Kim et al. addressed this problem through a CNN (49), getting better results than other dedicated libraries for this task. DL has been also applied to predict protein secondary structures from their amino acid sequences (50).

NMR technology is essential for both proteomics and metabolomics data generation. However, it has the technical limitation to return many noise signals that should be filtered in order to improve accuracy. Kobayashi et al. automated this necessary step by applying CNNs to remove noise peaks from NMR spectra (51), thereby improving the performance.

Applying DL methods to metabolomics data is especially challenging because they are unable to identify specific factors that contribute to individual samples, which is essential in these types of experiments (52). Despite this fact, some DL applications have been developed in this field providing interesting results. For instance, Date and Kikuchi combined DNN and mean decrease accuracy metric to analyze NMR-based metabolomics data (52). Asakura et al. also applied DNNs to metabolomics data, overperforming other ML applications (53).

CLINICAL APPLICATIONS AND PRECISION MEDICINE

Precision medicine basically aims to move away from general therapies for a broad population to individualized targeted therapies and treatment protocols depending on each patient's molecular background (54), or establish preventive medicine strategies based on disease susceptibility estimation (55). Omics data have a key role in this transition as they enable studying diseases from several simultaneous levels (e.g., DNA sequence, gene expression, and medical images) and identifying which parts of the complex biological functions are altered. In this new scenario, several ML-based approaches have been applied to medicine (56). However, although ML has been demonstrated to be useful in several precision medicine applications, it has some disadvantages that can be overcome by DL architectures. For instance, ML performance has a strong dependence on the data preprocessing to extract features, while DL models include this feature extraction (57).

Biomarker discovery and patient classification

One of the most common applications of omics technologies in biomedical research is the identification of new biomarkers for early diseases diagnosis, treatment response, and classification. The availability of large amounts of public omics data, especially in cancer, such as The Cancer Genome Atlas (TCGA) (58), has permitted the identification of new biomarkers with both DL and

non-DL strategies. A promising study applied an SDAE to classify breast cancer samples from the TCGA database into healthy or diseased using gene expression data (59). In addition, this method identified a set of highly interactive genes which could be good cancer biomarkers. Gene expression data from TCGA have also been exploited to accurately differentiate samples into different cancer types (60). On the other hand, Si et al. used an AE to classify healthy and breast cancer patients using methylation data (61), while Chatterjee et al. used CNN to classify different cancer types by their methylation patterns, achieving very promising results (62). Multiple omics (RNA-seq, miRNA-seq, and methylation data) have been combined by Chaudhary et al. to classify liver cancer patients into different survival groups (63). Authors used TCGA data to train their AE model, but they expect to improve their method using more clinical data in the future. In a similar work, Olivier et al. integrated the same kinds of omics data from TCGA to stratify bladder cancer patients by their survival chances (64). They used an AE approach to split patients into two survival groups. They also used these clusters to identify biomarkers linked to survival rates. Biomarkers for Alzheimer's disease have also been proposed using DFFs (65). Another precision oncology application is a tool developed by Yuan et al. to classify cancer types based on somatic mutations (66). The authors combined a DFF with other statistical techniques. They trained and tested their method with TCGA data for 12 cancer types.

Medical imaging

Medical imaging is one of the main tools for the transition from traditional medicine to precision medicine. This section reviews some DL-based imaging applications in the context of disease classification and diagnosis.

In skin cancer, the first step for diagnosing is based on visual inspections by dermatologists. Consequently, skin cancer diagnosis is a classical image recognition problem where researchers have applied ML methods and image recognition approaches. In a recent work, Estava et al. trained a CNN with thousands of clinical images to automatically identify whether a skin lesion is a skin cancer symptom (67). With their method, they obtained results as good as a panel of expert dermatologists. Some other studies addressed this problem with CNNs (68), all of them with promising results, and it is expected that this research will be translated in a few years into mobile applications able to accurately diagnose skin cancer lesions.

In the context of brain cancer, tumor segmentation is essential to define the shape and size of the tumor and apply diagnoses and therapies accordingly. This tumor segmentation is usually made manually by doctors using magnetic resonance imaging (MRI) images. However, this crucial task is very time-consuming and subjective. Therefore, there has been a lot of interest in automating tumor segmentation from MRI data. This task is very challenging because MRI data consist of 3D images where tumors are very different between patients, and in addition, they are very heterogeneous images depending on the device and experimental procedures employed (69). Several researchers addressed this challenge using CNNs (70–72) or SAEs (73).

Analysis of histopathological images is one of the most common tests for cancer diagnosis. As with brain tumor segmentation, the analysis of images is manually performed by pathologists, which is a time-consuming task. In this context,

several attempts have been made in order to automate this process. Litjens et al. reported a CNNs-based strategy for prostate and breast cancer diagnosis (74), although their results are very preliminary and much more research is necessary in this field. In addition, Xie et al. recently combined different DL algorithms to classify breast cancer subtypes from histopathological images (75). Colorectal polyps have been also classified applying a ResNet (76).

Computed tomography (CT) is used for the diagnosis of several diseases due to its capacity to generate three-dimensional anatomic images. Some DL approaches will likely enable the use of CT images in precision medicine. Roth et al., for instance, proposed the application of CNNs to automatically classify CT images into the different human anatomical parts (77). Such classification is the first step in many CT-based diagnostic strategies. There are also some specific applications in this field, for instance, for pancreas segmentation (78) or coronary artery calcium scoring (79).

Ultrasound (US) imaging is another imaging technique with many medical applications, for instance, in heart dysfunctions diagnosis. Carneiro and Nascimento innovated this field using DBNs to left ventricle endocardium tracking, allowing the automatic detection of different cardiopathies (80). On the other hand, Lekadir et al. applied a CNN to characterize carotid plaque composition (81). In addition, Biswas et al. developed a DL method to characterize liver US images, allowing the diagnosis and stratification of liver pathologies (82).

Some DL methods have been also applied to X-ray images. For instance, Nasr-Esfahani et al. used a CNN to detect vessel regions, a necessary step for coronary artery disease diagnosis (83). Bone age assessment is a common technique to detect growth abnormalities, and currently, it is done manually by comparing the X-ray images from databases. However, some authors applied DL algorithms to automate this process (84, 85).

Finally, facial images are being used with very promising results for automatic disease diagnosis. In a very recent work, Gurovich et al. have presented a facial analysis framework for genetic syndrome classification (86). They used patient facial images and CNNs to quantify similarities of facial features to hundreds of syndromes outperforming clinicians in diagnosis tasks.

CONCLUSION

Omics technologies are not only changing the way we study biomedicine but also introducing novel analytical challenges to bioinformatics analysts. DL is a promising approach to analyze these complex and heterogeneous datasets to drive precision medicine. This chapter reviewed some of the most common DL applications in omics data analysis and precision medicine. Although these methods have been used with very promising results, there are important considerations to take into account. The most successful application of DL in biomedical research to date has been in supervised learning; therefore, a crucial step is to avoid biases in training sets as quality of learning depends on the quality of the input data. No single method is universally applicable, and the choice of whether and how to use DL approaches will be problem-specific. Conventional analytical approaches will remain valid and have advantages when data are scarce or if the aim is to assess

statistical significance, which is currently difficult using DL methods. Another limitation of DL is the increased complexity, which applies both to model design and to the required computing environment. The application of DL methods to omics and precision medicine is a very new field. Although there are still some limitations, there is an increasing interest and research efforts that is resolving the major shortcomings and providing with very promising applications. The increasing availability of a larger number of omics datasets, medical images and clinical health records is fuelling the promising applications of DL technology that in the near future will play an increasingly important role in this field.

Conflict of Interest: The authors declare no potential conflicts of interest with respect to research, authorship and/or publication of this chapter.

REFERENCES

1. Cook CE, Bergman MT, Finn RD, Cochrane G, Birney E, Apweiler R. The European Bioinformatics Institute in 2016: Data growth and integration. Nucleic Acids Res. 2016 Jan 4;44(Database issue):D20–6. http://dx.doi.org/10.1093/nar/gkv1352
2. Grapov D, Fahrmann J, Wanichthanarak K, Khoomrung S. Rise of deep learning for genomic, proteomic, and metabolomic data integration in precision medicine. OMICS J Integr Biol. 2018 Oct 1;22(10):630–6. http://dx.doi.org/10.1089/omi.2018.0097
3. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015 May 28;521(7553):436–44. http://dx.doi.org/10.1038/nature14539
4. Min S, Lee B, Yoon S. Deep learning in bioinformatics. Brief Bioinform. 2017;18(5):851–69.
5. LeCun Y, Boser B, Denker JS, Henderson D, Howard RE, Hubbard W, et al. Backpropagation applied to handwritten zip code recognition. Neural Comput. 1989 Dec;1(4):541–51. http://dx.doi.org/10.1162/neco.1989.1.4.541
6. Deng J, Dong W, Socher R, Li L, Kai Li, Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, FL: IEEE (Institute of Electrical and Electronics Engineers); 2009. p. 248–55. http://dx.doi.org/10.1109/CVPR.2009.5206848
7. Orenstein EC, Beijbom O. Transfer learning and deep feature extraction for planktonic image data sets. In: 2017 IEEE Winter Conference on Applications of Computer Vision (WACV). Santa Rosa, CA: IEEE (Institute of Electrical and Electronics Engineers); 2017. p. 1082–8.
8. Rawat W, Wang Z. Deep convolutional neural networks for image classification: A comprehensive review. Neural Comput. 2017;29(9):2352–449. http://dx.doi.org/10.1162/neco_a_00990
9. Hinton GE. Deep belief networks. Scholarpedia. 2009 May 31;4(5):5947. http://dx.doi.org/10.4249/scholarpedia.5947
10. Fischer A, Igel C. An introduction to restricted Boltzmann machines. In: Alvarez L, Mejail M, Gomez L, Jacobo J, editors. Progress in pattern recognition, image analysis, computer vision, and applications. Berlin: Springer; 2012. p. 14–36. (Lecture Notes in Computer Science).

11. Ackley DH, Hinton GE, Sejnowski TJ. A learning algorithm for Boltzmann Machines*. Cogn Sci. 1985;9(1):147–69. http://dx.doi.org/10.1207/s15516709cog0901_7

12. Rumelhart DE, McClelland JL. Learning internal representations by error propagation. In: Parallel distributed processing: Explorations in the microstructure of cognition: Foundations [homepage on the Internet]. MITP; 1987 [cited 2019 May 16]. Available from: https://ieeexplore.ieee.org/document/6302929

13. Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. Science. 2006 Jul 28;313(5786):504–7. http://dx.doi.org/10.1126/science.1127647

14. Ng A, others. Sparse autoencoder. CS294A Lect Notes. 2011;72(2011):1–19.

15. Zhang Z, Zhao Y, Liao X, Shi W, Li K, Zou Q, et al. Deep learning in omics: A survey and guideline. Brief Funct Genomics. 2018 Sep 26; http://dx.doi.org/10.1093/bfgp/ely030

16. Poplin R, Chang P-C, Alexander D, Schwartz S, Colthurst T, Ku A, et al. A universal SNP and small-indel variant caller using deep neural networks. Nat Biotechnol. 2018;36(10):983–7. http://dx.doi.org/10.1038/nbt.4235

17. Qi H, Chen C, Zhang H, Long JJ, Chung WK, Guan Y, et al. MVP: Predicting pathogenicity of missense variants by deep neural networks. bioRxiv. 2018 Feb 2;259390. http://dx.doi.org/10.1101/259390

18. Quang D, Chen Y, Xie X. DANN: A deep learning approach for annotating the pathogenicity of genetic variants. Bioinforma Oxf Engl. 2015 Mar 1;31(5):761–3. http://dx.doi.org/10.1093/bioinformatics/btu703

19. Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning-based sequence model. Nat Methods. 2015 Oct;12(10):931–4. http://dx.doi.org/10.1038/nmeth.3547

20. Xie R, Wen J, Quitadamo A, Cheng J, Shi X. A deep auto-encoder model for gene expression prediction. BMC Genomics. 2017 Nov 18;18(Suppl 9):845. http://dx.doi.org/10.1038/nmeth.3547

21. Liu F, Li H, Ren C, Bo X, Shu W. PEDLA: Predicting enhancers with a deep learning-based algorithmic framework. Sci Rep. 2016 22;6:28517. http://dx.doi.org/10.1038/srep28517

22. Min X, Zeng W, Chen S, Chen N, Chen T, Jiang R. Predicting enhancers with deep convolutional neural networks. BMC Bioinformatics. 2017 Dec 1;18(Suppl 13):478. http://dx.doi.org/10.1186/s12859-017-1878-3

23. Kleftogiannis D, Kalnis P, Bajic VB. DEEP: A general computational framework for predicting enhancers. Nucleic Acids Res. 2015 Jan;43(1):e6. http://dx.doi.org/10.1093/nar/gku1058

24. Li Y, Shi W, Wasserman WW. Genome-wide prediction of cis-regulatory regions using supervised deep learning methods. BMC Bioinformatics. 2018 31;19(1):202. http://dx.doi.org/10.1186/s12859-018-2187-1

25. Eser U, Churchman LS. FIDDLE: An integrative deep learning framework for functional genomic data inference. bioRxiv. 2016 Oct 17;081380. http://dx.doi.org/10.1101/081380

26. Wang M, Tai C, E W, Wei L. DeFine: Deep convolutional neural networks accurately quantify intensities of transcription factor-DNA binding and facilitate evaluation of functional non-coding variants. Nucleic Acids Res. 2018 Jun 20;46(11):e69. http://dx.doi.org/10.1093/nar/gky215

27. Oubounyt M, Louadi Z, Tayara H, Chong KT. DeePromoter: Robust promoter predictor using deep learning. Front Genet. 2019;10:286. http://dx.doi.org/10.3389/fgene.2019.00286

28. Zuallaert J, Godin F, Kim M, Soete A, Saeys Y, De Neve W. SpliceRover: Interpretable convolutional neural networks for improved splice site prediction. Bioinforma Oxf Engl. 2018 Dec 15;34(24):4180–8. http://dx.doi.org/10.1093/bioinformatics/bty497

29. Wei Z, Wang K, Qu H-Q, Zhang H, Bradfield J, Kim C, et al. From disease association to risk assessment: An optimistic view from genome-wide association studies on type 1 diabetes. PLoS Genet. 2009 Oct;5(10):e1000678. http://dx.doi.org/10.1371/journal.pgen.1000678

30. The International Schizophrenia Consortium. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature. 2009 Aug;460(7256):748–52. http://dx.doi.org/10.1038/nature08185

31. Montaez CAC, Fergus P, Montaez AC, Hussain A, Al-Jumeily D, Chalmers C. Deep learning classification of polygenic obesity using genome wide association study SNPs. In: 2018 International Joint Conference on Neural Networks (IJCNN). Rio de Janeiro: IEEE, 2018. p. 1–8.

32. Fergus P, Montanez A, Abdulaimma B, Lisboa P, Chalmers C, Pineles B. Utilising deep learning and genome wide association studies for epistatic-driven preterm birth classification in African-American Women. IEEE/ACM Trans Comput Biol Bioinform. IEEE, 2018:1–1. http://dx.doi.org/10.1109/TCBB.2018.2868667

33. Arnedo J, del Val C, de Erausquin GA, Romero-Zaliz R, Svrakic D, Cloninger CR, et al. PGMRA: A web server for (phenotype x genotype) many-to-many relation analysis in GWAS. Nucleic Acids Res. 2013 Jul;41(Web Server issue):W142–149. http://dx.doi.org/10.1093/nar/gkt496

34. Arnedo J, Svrakic DM, Del Val C, Romero-Zaliz R, Hernández-Cuervo H, Molecular genetics of Schizophrenia Consortium, et al. Uncovering the hidden risk architecture of the schizophrenias: Confirmation in three independent genome-wide association studies. Am J Psychiatry. 2015 Feb 1; 172(2):139–53. http://dx.doi.org/10.1176/appi.ajp.2014.14040435

35. Zhang Z, Pan Z, Ying Y, Xie Z, Adhikari S, Phillips J, et al. Deep-learning augmented RNA-seq analysis of transcript splicing. Nat Methods. 2019;16(4):307–10. http://dx.doi.org/10.1038/s41592-019-0351-9

36. Zhang Y, Liu X, MacLeod J, Liu J. Discerning novel splice junctions derived from RNA-seq alignment: A deep learning approach. BMC Genomics. 2018 Dec 27;19(1):971. http://dx.doi.org/10.1186/s12864-018-5350-1

37. Jha A, Gazzara MR, Barash Y. Integrative deep models for alternative splicing. Bioinforma Oxf Engl. 2017 Jul 15;33(14):i274–82. http://dx.doi.org/10.1093/bioinformatics/btx268

38. Hill ST, Kuintzle R, Teegarden A, Merrill E, Danaee P, Hendrix DA. A deep recurrent neural network discovers complex biological rules to decipher RNA protein-coding potential. Nucleic Acids Res. 2018 Sep 19;46(16):8105–13. http://dx.doi.org/10.1093/nar/gky567

39. Tripathi R, Patel S, Kumari V, Chakraborty P, Varadwaj PK. DeepLNC, a long non-coding RNA prediction tool using deep neural network. Netw Model Anal Health Inform Bioinforma. 2016 Jun 10; 5(1):21. http://dx.doi.org/10.1007/s13721-016-0129-2

40. Yu N, Yu Z, Pan Y. A deep learning method for lincRNA detection using auto-encoder algorithm. BMC Bioinformatics. 2017 Dec 6;18(Suppl 15):511. http://dx.doi.org/10.1186/s12859-017-1922-3

41. Alipanahi B, Delong A, Weirauch MT, Frey BJ. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat Biotechnol. 2015 Aug;33(8):831–8. http://dx.doi.org/10.1038/nbt.3300

42. Schreiber J, Libbrecht M, Bilmes J, Noble WS. Nucleotide sequence and DNaseI sensitivity are predictive of 3D chromatin architecture. bioRxiv. 2018 Jul 15;103614. http://dx.doi.org/10.1101/103614

43. Wang Y, Liu T, Xu D, Shi H, Zhang C, Mo Y-Y, et al. Predicting DNA methylation state of CpG dinucleotide using genome topological features and deep networks. Sci Rep. 2016 Jan 22;6:19598. http://dx.doi.org/10.1038/srep19598

44. Yin Q, Wu M, Liu Q, Lv H, Jiang R. DeepHistone: A deep learning approach to predicting histone modifications. BMC Genomics. 2019 Apr 4;20(Suppl 2):193. http://dx.doi.org/10.1186/s12864-019-5489-4

45. Singh R, Lanchantin J, Robins G, Qi Y. DeepChrome: Deep-learning for predicting gene expression from histone modifications. Bioinforma Oxf Engl. 2016 01;32(17):i639–48. http://dx.doi.org/10.1093/bioinformatics/btw427

46. Sekhon A, Singh R, Qi Y. DeepDiff: DEEP-learning for predicting DIFFerential gene expression from histone modifications. Bioinforma Oxf Engl. 2018 01;34(17):i891–900. http://dx.doi.org/10.1093/bioinformatics/bty612

47. Zhou X-X, Zeng W-F, Chi H, Luo C, Liu C, Zhan J, et al. pDeep: Predicting MS/MS spectra of peptides with deep learning. Anal Chem. 2017 05;89(23):12690–7. http://dx.doi.org/10.1021/acs.analchem.7b02566

48. Tran NH, Zhang X, Xin L, Shan B, Li M. De novo peptide sequencing by deep learning. Proc Natl Acad Sci U S A. 2017 Jul 18; http://dx.doi.org/10.1073/pnas.1705691114

49. Kim M, Eetemadi A, Tagkopoulos I. DeepPep: Deep proteome inference from peptide profiles. PLoS Comput Biol. 2017 Sep;13(9):e1005661. http://dx.doi.org/10.1371/journal.pcbi.1005661

50. Spencer M, Eickholt J, Jianlin Cheng null. A deep learning network approach to ab initio protein secondary structure prediction. IEEE/ACM Trans Comput Biol Bioinform. 2015 Feb;12(1):103–12. http://dx.doi.org/10.1109/TCBB.2014.2343960

51. Kobayashi N, Hattori Y, Nagata T, Shinya S, Güntert P, Kojima C, et al. Noise peak filtering in multi-dimensional NMR spectra using convolutional neural networks. Bioinforma Oxf Engl. 2018 Dec 15; 34(24):4300–1. http://dx.doi.org/10.1093/bioinformatics/bty581

52. Date Y, Kikuchi J. Application of a deep neural network to metabolomics studies and its performance in determining important variables. Anal Chem. 2018 06;90(3):1805–10. http://dx.doi.org/10.1021/acs.analchem.7b03795

53. Asakura T, Date Y, Kikuchi J. Application of ensemble deep neural network to metabolomics studies. Anal Chim Acta. 2018 Dec 11;1037:230–6. http://dx.doi.org/10.1016/j.aca.2018.02.045

54. Ashley EA. Towards precision medicine. Nat Rev Genet. 2016 16;17(9):507–22. http://dx.doi.org/10.1038/nrg.2016.86

55. Chen R, Snyder M. Promise of personalized omics to precision medicine. Wiley Interdiscip Rev Syst Biol Med. 2013 Feb;5(1):73–82. http://dx.doi.org/10.1002/wsbm.1198

56. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med. 2019 04;380(14):1347–58. http://dx.doi.org/10.1056/NEJMra1814259

57. Eraslan G, Avsec Ž, Gagneur J, Theis FJ. Deep learning: New computational modelling techniques for genomics. Nat Rev Genet. 2019 Apr 10; http://dx.doi.org/10.1038/s41576-019-0122-6

58. Weinstein JN, Collisson EA, Mills GB, Shaw KM, Ozenberger BA, Ellrott K, et al. The Cancer Genome Atlas Pan-cancer analysis project. Nat Genet. 2013 Oct;45(10):1113–20. http://dx.doi.org/10.1038/ng.2764

59. Danaee P, Ghaeini R, Hendrix DA. A deep learning approach for cancer detection and relevant gene identification. Pac Symp Biocomput Pac Symp Biocomput. 2017;22:219–29. http://dx.doi.org/10.1142/9789813207813_0022

60. Lyu B, Haque A. Deep learning based tumor type classification using gene expression data. bioRxiv. 2018 Jul 11;364323. http://dx.doi.org/10.1101/364323

61. Si Z, Yu H, Ma Z. Learning deep features for DNA methylation data analysis. IEEE Access. 2016;4:2732–7. http://dx.doi.org/10.1109/ACCESS.2016.2576598

62. Chatterjee S, Iyer A, Avva S, Kollara A, Sankarasubbu M. Convolutional neural networks in classifying cancer through DNA methylation. ArXiv180709617 Cs Q-Bio Stat [Internet]. 2018 Jul 24 [cited 2019 May 6]; Available from: http://arxiv.org/abs/1807.09617

63. Chaudhary K, Poirion OB, Lu L, Garmire LX. Deep learning-based multi-Omics integration robustly predicts survival in liver cancer. Clin Cancer Res Off J Am Assoc Cancer Res. 2018 Mar 15;24(6):1248–59. http://dx.doi.org/10.1158/1078-0432.CCR-17-0853

64. Poirion OB, Chaudhary K, Garmire LX. Deep learning data integration for better risk stratification models of bladder cancer. AMIA Jt Summits Transl Sci Proc AMIA Jt Summits Transl Sci. 2018;2017:197–206.

65. Zafeiris D, Rutella S, Ball GR. An artificial neural network integrated pipeline for biomarker discovery using Alzheimer's disease as a case study. Comput Struct Biotechnol J. 2018;16:77–87. http://dx.doi.org/10.1016/j.csbj.2018.02.001

66. Yuan Y, Shi Y, Li C, Kim J, Cai W, Han Z, et al. DeepGene: An advanced cancer type classifier based on deep learning and somatic point mutations. BMC Bioinformatics. 2016 Dec 23;17(Suppl 17):476. http://dx.doi.org/10.1186/s12859-016-1334-9

67. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017 02;542(7639):115–8. http://dx.doi.org/10.1038/nature21056

68. Brinker TJ, Hekler A, Utikal JS, Grabe N, Schadendorf D, Klode J, et al. Skin cancer classification using convolutional neural networks: Systematic review. J Med Internet Res. 2018 Oct 17;20(10):e11936. http://dx.doi.org/10.2196/11936

69. Işın A, Direkoğlu C, Şah M. Review of MRI-based brain tumor image segmentation using deep learning methods. Procedia Comput Sci. 2016 Jan 1;102:317–24. http://dx.doi.org/10.1016/j.procs.2016.09.407

70. Pereira S, Pinto A, Alves V, Silva CA. Brain tumor segmentation using convolutional neural networks in MRI images. IEEE Trans Med Imaging. 2016;35(5):1240–51. http://dx.doi.org/10.1109/TMI.2016.2538465

71. Naceur MB, Saouli R, Akil M, Kachouri R. Fully automatic brain tumor segmentation using end-to-end incremental deep neural networks in MRI images. Comput Methods Programs Biomed. 2018 Nov;166:39–49. http://dx.doi.org/10.1016/j.cmpb.2018.09.007

72. Havaei M, Davy A, Warde-Farley D, Biard A, Courville A, Bengio Y, et al. Brain tumor segmentation with deep neural networks. Med Image Anal. 2017;35:18–31. http://dx.doi.org/10.1016/j.media.2016.05.004

73. Xiao Z, Huang R, Ding Y, Lan T, Dong R, Qin Z, et al. A deep learning-based segmentation method for brain tumor in MR images. In: 2016 IEEE 6th International Conference on Computational Advances in Bio and Medical Sciences (ICCABS). Atlanta, GA: IEEE, 2016. p. 1–6.

74. Litjens G, Sánchez CI, Timofeeva N, Hermsen M, Nagtegaal I, Kovacs I, et al. Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. Sci Rep. 2016 23;6:26286. http://dx.doi.org/10.1038/srep26286

75. Xie J, Liu R, Luttrell J, Zhang C. Deep learning based analysis of histopathological images of breast cancer. Front Genet. 2019;10:80. http://dx.doi.org/10.3389/fgene.2019.00080

76. Korbar B, Olofson AM, Miraflor AP, Nicka CM, Suriawinata MA, Torresani L, et al. Deep learning for classification of colorectal polyps on whole-slide images. J Pathol Inform. 2017;8:30. http://dx.doi.org/10.4103/jpi.jpi_34_17

77. Roth HR, Lee CT, Shin H-C, Seff A, Kim L, Yao J, et al. Anatomy-specific classification of medical images using deep convolutional nets. 2015 IEEE 12th Int Symp Biomed Imaging ISBI. 2015 Apr. p. 101–4. http://dx.doi.org/10.1109/ISBI.2015.7163826

78. Roth HR, Farag A, Lu L, Turkbey EB, Summers RM. Deep convolutional networks for pancreas segmentation in CT imaging. ArXiv150403967 Cs. 2015 Mar 20;94131G. http://dx.doi.org/10.1117/12.2081420

79. Wolterink JM, Leiner T, de Vos BD, van Hamersvelt RW, Viergever MA, Išgum I. Automatic coronary artery calcium scoring in cardiac CT angiography using paired convolutional neural networks. Med Image Anal. 2016;34:123–36. http://dx.doi.org/10.1016/j.media.2016.04.004

80. Carneiro G, Nascimento JC. Combining multiple dynamic models and deep learning architectures for tracking the left ventricle endocardium in ultrasound data. IEEE Trans Pattern Anal Mach Intell. 2013 Nov;35(11):2592–607. http://dx.doi.org/10.1109/TPAMI.2013.96

81. Lekadir K, Galimzianova A, Betriu A, Del Mar Vila M, Igual L, Rubin DL, et al. A convolutional neural network for automatic characterization of plaque composition in carotid ultrasound. IEEE J Biomed Health Inform. 2017;21(1):48–55. http://dx.doi.org/10.1109/JBHI.2016.2631401

82. Biswas M, Kuppili V, Edla DR, Suri HS, Saba L, Marinhoe RT, et al. Symtosis: A liver ultrasound tissue characterization and risk stratification in optimized deep learning paradigm. Comput Methods Programs Biomed. 2018;155:165–77. http://dx.doi.org/10.1016/j.cmpb.2017.12.016

83. Nasr-Esfahani E, Samavi S, Karimi N, Soroushmehr SMR, Ward K, Jafari MH, et al. Vessel extraction in X-ray angiograms using deep learning. Conf Proc IEEE Eng Med Biol Soc. 2016;2016:643–6. http://dx.doi.org/10.1109/EMBC.2016.7590784

84. Lee JH, Kim KG. Applying deep learning in medical images: The case of bone age estimation. Healthc Inform Res. 2018 Jan;24(1):86–92. http://dx.doi.org/10.4258/hir.2018.24.1.86

85. Spampinato C, Palazzo S, Giordano D, Aldinucci M, Leonardi R. Deep learning for automated skeletal bone age assessment in X-ray images. Med Image Anal. 2017;36:41–51. http://dx.doi.org/10.1016/j.media.2016.10.010

86. Gurovich Y, Hanani Y, Bar O, Nadav G, Fleischer N, Gelbman D, et al. Identifying facial phenotypes of genetic disorders using deep learning. Nat Med. 2019;25(1):60–4. http://dx.doi.org/10.1038/s41591-018-0279-0

87. LeCun Y, Boser BE, Denker JS, Henderson D, Howard RE, Hubbard WE, et al. Handwritten digit recognition with a back-propagation network. In: Touretzky DS, editor. Advances in neural information processing systems 2 [homepage on the Internet]. Morgan-Kaufmann; 1990 [cited 2019 May 16]. p. 396–404. Available from: http://papers.nips.cc/paper/293-handwritten-digit-recognition-with-a-back-propagation-network.pdf

88. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. In: Pereira F, Burges CJC, Bottou L, Weinberger KQ, editors. Advances in neural information processing systems 25 [homepage on the Internet]. Curran Associates, Inc.; 2012 [cited 2019 May 16].

p. 1097–1105. Available from: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

89. Szegedy C, Wei Liu, Yangqing Jia, Sermanet P, Reed S, Anguelov D, et al. Going deeper with convolutions. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) [homepage on the Internet]. Boston, MA, USA: IEEE; 2015 [cited 2019 May 16]. p. 1–9. Available from: http://ieeexplore.ieee.org/document/7298594/

90. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. ArXiv14091556 Cs [homepage on the Internet]. 2014 Sep 4 [cited 2019 May 16]; Available from: http://arxiv.org/abs/1409.1556

91. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas, NV: IEEE, 2016. p. 770–8.

92. Huang G, Liu Z, Maaten L v d, Weinberger KQ. Densely connected convolutional networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI: IEEE, 2017. p. 2261–9.

93. Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, et al. The Cancer Imaging Archive (TCIA): Maintaining and operating a public information repository. J Digit Imaging. 2013 Dec;26(6): 1045–57. http://dx.doi.org/10.1007/s10278-013-9622-7

94. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. ChestX-Ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI: IEEE, 2017. p. 3462–71.