Human Genomics

REVIEW Open Access

A review of deep learning applications in human genomics using next-generation sequencing data

Wardah S. Alharbi and Mamoon Rashid*

Abstract

Genomics is advancing towards data-driven science. Through the advent of high-throughput data generating technologies in human genomics, we are overwhelmed with the heap of genomic data. To extract knowledge and pattern out of this genomic data, artificial intelligence especially deep learning methods has been instrumental. In the current review, we address development and application of deep learning methods/models in different subarea of human genomics. We assessed over- and under-charted area of genomics by deep learning techniques. Deep learning algorithms underlying the genomic tools have been discussed briefly in later part of this review. Finally, we discussed briefly about the late application of deep learning tools in genomic. Conclusively, this review is timely for biotechnology or genomic scientists in order to guide them why, when and how to use deep learning methods to analyse human genomic data.

Keywords: Human genomics, Deep learning applications, Disease variants, Gene expression, Epigenomics, Pharmacogenomics, Variant calling, NGS

Introduction

Understanding the genomes of diverse species, specifically, the examination of more than 3 billion base-pairs of Homo sapiens DNA, is a crucial aim of genomic studies. Genomics takes a comprehensive view that implicates all the genes within an organism, including protein-coding genes, RNA genes, *cis*- and *trans*-elements, etc. It is a data-driven science involving the high-throughput technological development of next-generation sequencing (NGS) that generates the entire DNA data of an organism. These techniques include whole genome sequencing (WGS), whole exome sequencing (WES), transcriptomic and proteomic profiling [1–5]. With the recent rapid accumulation of these omics data, increased attention has been paid to bioinformatics and machine learning (ML) tools with established superior performance in several genomics implementations [6]. These implementations involve finding a genotype–phenotype correlation, biomarker identification and gene function prediction, as well as mapping the biomedically active genomic regions, for example, transcriptional enhancers [7–10].

Machine learning (ML) has been deliberated as a core technology in artificial intelligence (AI), which enables the use of algorithms and makes critical predictions based on data learning and not simply following instructions. It has broad technology applications; however, standard ML methods are too narrow to deal with complex, natural, highly dimensional raw data, such as those of genomics. Alternatively, the deep learning (DL) approach is a promising and exciting field currently employed in genomics. It is an ML derivative that extracts features by applying neural networks (NN)

*Correspondence: rashidmamoon@gmail.com; rashidma@ngha.med.sa

Department of AI and Bioinformatics, King Abdullah International Medical Research Center (KAIMRC), King Saud Bin Abdulaziz University for Health Sciences (KSAU-HS), King Abdulaziz Medical City, Ministry of National Guard Health Affairs, P.O. Box 22490, Riyadh 11426, Saudi Arabia

© The Author(s) 2022. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

automatically [11–14]. Deep learning has been effectively applied in fields such as image recognition, audio classification, natural language processing, online web tools, chatbots and robotics. In this regard, the utilisation of DL as a genomic methodology is totally apt to analyse a large amount of data. While it is still in its infant stages, DL in genomics holds the promise of updating arenas such as clinical genetics and functional genomics [15]. Undoubtedly, DL algorithms have dominated computational modelling approaches in which they are currently regularly expanded to report a variety of genomics questions ranging from understanding the effects of mutations on protein–RNA binding [16], prioritising variants and genes, diagnosing patients with rare genetic disorders [17], predicting gene expression levels from histone modification data [18] and to identifying trait-associated single-nucleotide polymorphisms (SNPs) [19].

Although the first concept of the DL theory originated in the 1980s was based on the perceptron model and neuron concept [20], within the last decade, DL algorithms have become a state-of-the-art predictive technology for big data [21–23]. The initial efficient implementation of DL prediction models in genomics was in the 2000s (Fig. 1) [24]. The difficulty associated with the requirement of DL models to train an enormous amount of training datasets and the need for powerful computing resources limited their applications until the introduction of modern hardware, such as the high-efficiency graphical processing units (GPUs) with equivalent structures. Now, the architectures of DL models (also known as

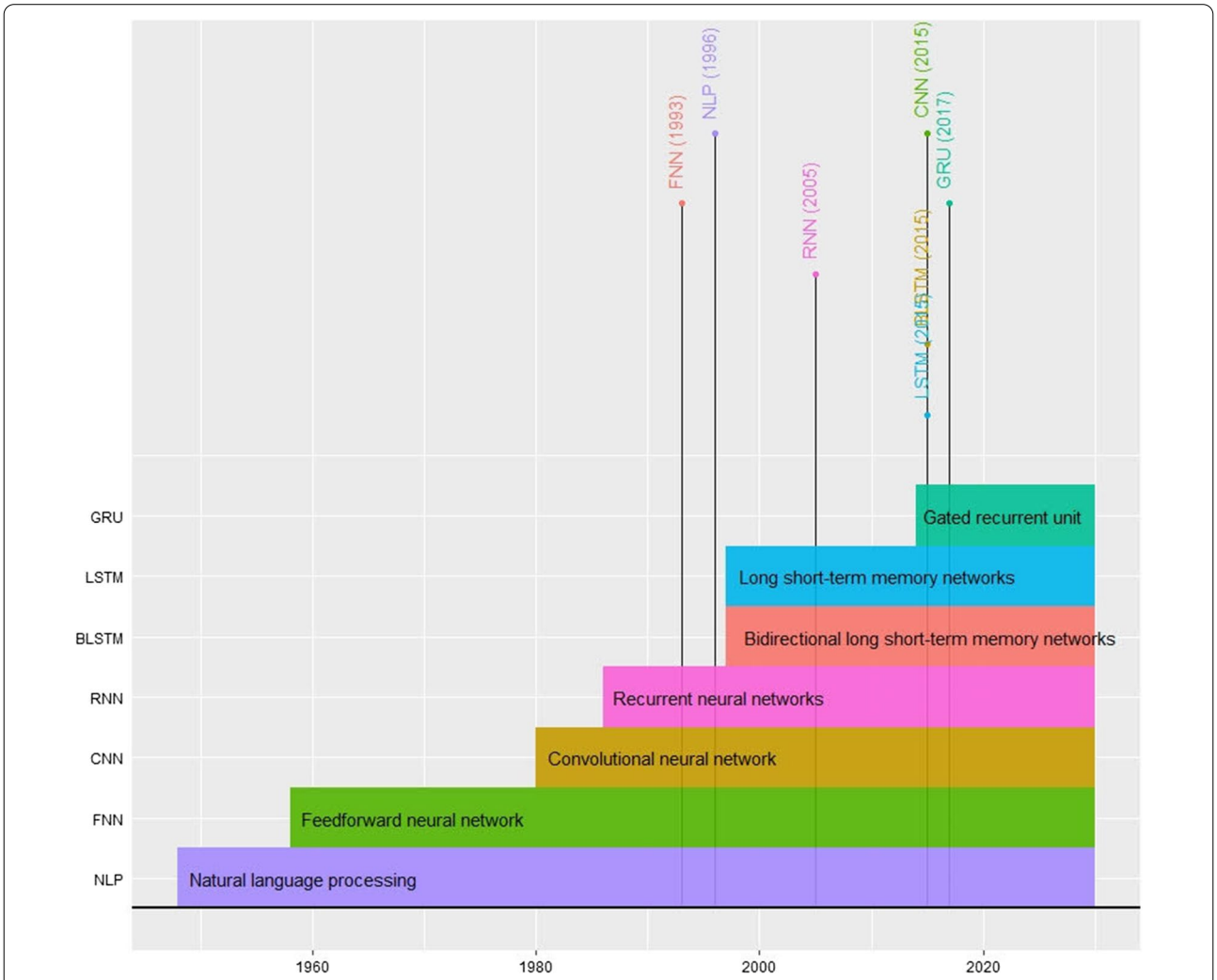

Fig. 1 Timeline of implementing deep learning algorithms in genomics. This timeline plot demonstrated the delay of implementing DL tools in genomics; for example, both (LSTM) and (BLSTM) algorithms have been invented in 1997 and the first genomic application was implemented in 2015. Similar observations are for the rest of the deep learning algorithms (Table 6)

DNNs) are implemented in diverse areas, as mentioned earlier. Classical neural networks consist of only two to three hidden layers; however, DL networks extend this up to 200 layers. Thus, the word "deep" reflects the number of layers that the information passes through. However, DL requires superior hardware and substantial parallelism to be applicable [25]. Due to overwhelmed hardware limitations and demanding resources, several DL packages and resources were introduced to facilitate DL model implementation (discussed in section deep learning resources for genomics).

The evolution of software, hardware (GPUs) and big data in genomics has facilitated the development of deep learning-based prediction models for the prediction of functional elements in genomes. These genetic variants from NGS data predict splice sites in genomic DNA, predict the transcription factor binding sites (TFBSs) via classification tasks, classify the pathogenicity of missense mutations and predict drug response and synergy [26–31]. An example of a technological evolution that has enhanced DL implementation is cloud platforms, which provide GPU resources as a DL solution. GPUs can considerably escalate the training speed as the neural network training style can be more adaptable in certain model architecture situations, thus permitting fast mathematical processes through the use of larger processing unit numbers and high-memory capacities. Primary examples of cloud computing platforms include Amazon Web Services, Google Compute Engine and Microsoft Azure. However, these elucidations still require users to implement model codes [32].

For all ML models, the evaluation metrics are essential in understanding the model performance. Basically, these metrics are crucial to be considered in case of genomic datasets which generate naturally a highly imbalanced classes that makes them demanding to be applied by ML and DL models. A sufficient number of solutions usually applied in this case such as transfer learning [33] and Matthews correlation coefficient (MCC) [34]. In common sense, every ML task can be divided into a regression task (e.g. predicting certain outcomes/effects of a disease) or a classification task (e.g. predicting the presence/absence of a disease); additionally, multiple measurement metrics are obtained from those tasks. Generally, some, but not all, performance metrics used in ML regression-based methods include: mean absolute error (MAE), mean squared error (MSE), root-mean-squared error (RMSE) and coefficient of determination (R^2). In contrast, the performance metrics in ML classification-based methods include: accuracy, confusion matrix, area under the curve (AUC) or/and area under receiver operating characteristics (AUROC) and F1-score. The classification tasks are most commonly applied to problems in research areas in genomics and for comparing different models' performance. For example, AUC is the most widely used metric for evaluating the model performance ranging from [0, 1]. It measures the true-positive rate (TPR) or sensitivity, true-negative rate (TNR) or specificity and the false-positive rate (FPR). Additionally, the F1-score is used to test the model accuracy in highly imbalanced dataset and is the harmonic mean between the precision and recall (also ranging from [0, 1]). For both AUC and F1-score, a greater value reflects better model performance. Also, the confusion matrix describes the complete model performance by measuring the model accuracy to calculate true-positive values plus true-negative values and dividing the sum over the total number of samples [35, 36]. For a greater understanding of the ML evaluation metrics—purpose, calculation, etc.—recommended papers include Handelman et al. (2019) and England and Cheng (2019).

This article reviews deep learning tools/methods based on their current applications in human genomics. We began by collecting recent (i.e. published in 2015–2020) DL tools in five main genomics areas: variant calling and annotation, disease variants, gene expression and regulation, epigenomics and pharmacogenomics. Then, we briefly discussed DL genomics-based algorithms and their application strategies and data structure. Finally, we mentioned DL-based practical resources to facilitate DL adoption that would be extremely beneficial mostly to biomedical researchers and scientists working in human genomics. For further information on the field of DL applications in genomics, we recommend: [37–39].

Deep learning tools/software/pipelines in genomics

Multiple genomic disciplines (e.g. variant calling and annotation, disease variant prediction, gene expression and regulation, epigenomics and pharmacogenomics) take advantage of generating high-throughput data and utilising the power of deep learning algorithms for sophisticated predictions (Fig. 2). The modern evolution of DNA/RNA sequencing technologies and machine learning algorithms especially deep learning opens a new chapter of research capable of transforming big biological data into new knowledge or novel findings in all subareas of genomics. The following sections will discuss the latest software/tools/pipelines developed using deep learning algorithms in various genomics areas.

Variant calling and annotation

This first section discusses the applications of the latest DL algorithms in variant calling and annotation. We provided a short list of tools/algorithms for variant calling and annotation with their source code links, if available

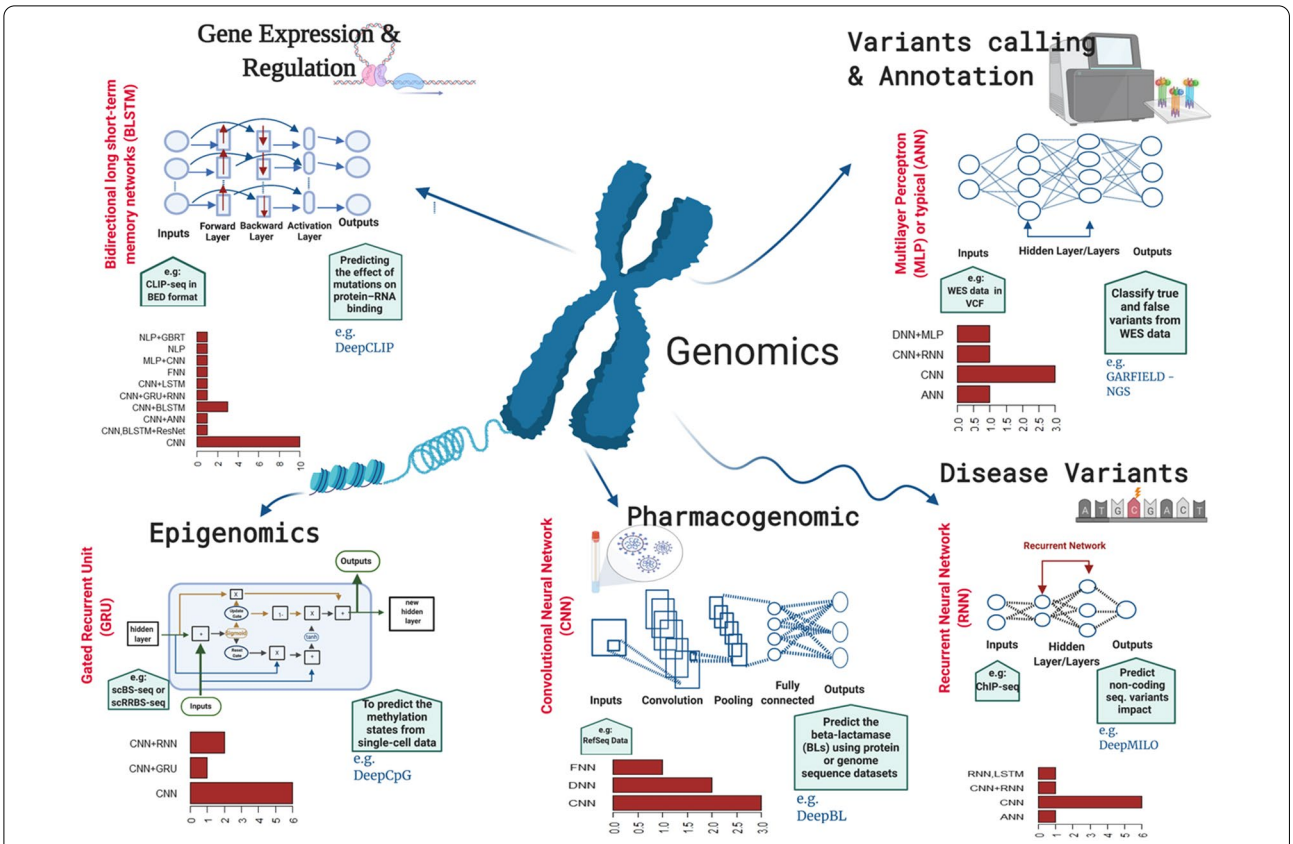

Fig. 2 Deep learning applications in genomics. This figure represents the application of deep learning tools in five major subareas of genomics. One example deep learning tool and underlying network architecture has been shown for each of the genomic subareas, and its input data type and the predictive output were mentioned briefly. Each bar plot depicts the frequency of most used deep learning algorithms underlying deep learning tools in that subarea of genomics (Tables 1, 2, 3, 4, 5)

(Table 1), to facilitate the selection of the most suitable DL tool for a particular data type.

NGS, including whole genome or exome, sets the stage for early developments in personalised medicine, along with its known implications in Mendelian disease research. With the advent of massively parallel, high-throughput sequencing, sequencing thousands of human genomes to identify genetic variations has become a

Table 1 Genomic tools/algorithm based on deep learning architecture for variant calling and annotations

Tools	DL model	Application	Input/Output	Website Code Source	References
Clairvoyante	CNN	To predict variant type, zygosity, alternative allele and Indel length	BAM/VCF	https://github.com/aquaskyline/Clairvoyante	[145]
DeepVariant	CNN	To call genetic variants from next-generation DNA sequencing data	BAM,CRAM/VCF	https://github.com/google/deepvariant	[30]
GARFIELD-NGS	DNN + MLP	To classify true and false variants from WES data	VCF/VCF	https://github.com/gedoardo83/GARFIELD-NGS	[146]
Intelli-NGS	ANN	To define good and bad variant calls from Ion Torrent sequencer data	VCF/xlsx	https://github.com/aditya-88/intelli-ngs	[147]
DAVI (Deep Alignment and Variant Identification)	CNN + RNN	To identify variants in NGS reads	FASTQ/VCF	N/A	[116]
DeepSV	CNN	To call genomic deletions by visualising sequence reads	BAM/VCF	https://github.com/CSuperlei/DeepSV	[52]

routine practice in genomics, including cancer research. Sophisticated bioinformatics and statistical frameworks are available for variant calling.

The weakness of high-throughput sequencing procedures is represented by significantly high technical and bioinformatics error rates [40–42]. Numerous computational problems have originated due to the enormous amounts of medium or low coverage genome sequences, short read fragments and genetic variations among individuals [43]. Such weaknesses make the NGS data dependent on bioinformatics tools for data interpretation. For instance, several variant calling tools are broadly used in clinical genomic variant analyses, such as genome analysis toolkit (GATK) [44], SAMtools [45], Freebayes [46] and Torrent Variant Caller (TVC; [47]). However, despite the availability of whole genome sequencing, some actual variants are yet to be discovered [48].

Contemporary deep learning tools have been proposed in the field of next-generation sequencing to overcome the limitations of conventional interpretation pipelines. For example, Kumaran et al. demonstrated that combining DeepVariant, a deep learning-based variant caller, with conventional variant callers (e.g. SAMtools and GATK) improved the accuracy scores of single-nucleotide variants and Indel detections [49]. Implementing deep learning algorithms in DNA sequencing data interpretation is in its infancy, as seen with the recent pioneering example, DeepVariant, developed by Google. DeepVariant relies on the graphical dissimilarities in input images to perform the classification task for genetic variant calling from NGS short reads. It treats the mapped sequencing datasets as images and converts the variant calls into image classification tasks [30]. However, this model does not provide details about the variant information, for example, the exact alternative allele and type of variant. As such, it is classified as an incomplete variant caller model [50].

Later, several DL models for variant calling and annotation were introduced. For instance, Cai et al. (2019) introduced DeepSV, a genetic variant caller that aims to predict long genomic deletions (>50 bp) extracted from sequencing read images but not other types of structural variants, such as long insertions or inversions. It processes the BAM format or VCF files as inputs and outputs the results in the VCF form. In terms of evaluating DeepSV, it was compared with another eight deletion calling tools and one machine learning-based tool called Concod [51]. The results reveal that although Concod has shorter training times in the case of fewer trained samples, DeepSV shows a higher accuracy score and fewer training losses using the same dataset [52]. Another genomic variant filtering tool, GARFIELD-NGS, can be applied directly to the variant caller outputs. It relies on an MLP algorithm to investigate the true and false variants in exome sequencing datasets generated from the Ion Torrent and Illumina platforms. It represents a robust performance at low coverage data (up to 30X) by handling standard VCF file, resulting in another VCF file. Ravasio et al. (2018) observed that the GARFIELD-NGS model recorded a significant reduction in the false candidate variants after applying a canonical pipeline for the variant prioritisation of disease-related data [53].

The Clairvoyante model was introduced to predict variant type (SNP or Indel), zygosity, allele alternative and Indel length. Thus, it overcomes the DeepVariant model's drawback of lacking the full variant details, including the precise alternative allele and variant type. The Clairvoyante model was specifically designed to utilise long-read sequencing data generated from SMS technologies (e.g. PacBio and ONT), although it is commonly applicable for short read datasets as well [50]. Another variant caller and annotation model, Intelli-NGS, was introduced by Singh and Bhatia (2019). One variant calling was based on artificial neural network (ANN), which utilises the data generated from the Ion Torrent platform to identify true and false effectively. Intelli-NGS takes any number of VCF files as batch inputs and processes them in order. The processed data results in an excel sheet related to each VCF file containing the HGVS codes of all variants [54]. All in all, several studies confirmed the capabilities of deep learning in genetic variant calling and annotation from sequencing data.

Disease variants

Deep learning-based models for the prediction of pathogenic variants, their application and input/output formats with source codes (if available) are listed in Table 2.

Considering extra data from patient relatives or relevant cohorts, medical geneticists frequently prioritise and filter the observed genetic variants after variant calling and annotation (Müller et al. [55]). Variant prioritisation is a method of determining the most likely pathogenic variant within genetic screening that damages gene function and underlying the disease phenotype [56]. Variant prioritisation involves variant annotation to discover clinically insignificant variants, such as synonymous, deep-intronic variants and benign polymorphisms. Subsequently, the remaining variants, such as known variants or variants of unknown clinical significance (VUSs), become attainable [57]. Furthermore, complications in interpreting rare genetic variants in individuals, for example, and understanding their impacts on disorder risk influence the clinical capability of diagnostic sequencing. For example, the numerous and infrequent VUSs in rare genetic diseases represent a challenging obstacle in sequencing implementation for personalised

Table 2 Genomic tools/algorithm based on deep learning architecture for disease variants

Tools	DL model	Application	Input/Output	Website Code Source	References
DeepPVP (PhenomeNet Variant Predictor)	ANN	to identify the variants in both whole exome or whole genome sequence data	VCF / VCF	https://github.com/bio-ontology-research-group/phenomenet-vp	[61]
ExPecto	CNN	Accurately predict tissue-specific transcriptional effects of mutations/functional SNPs	VCF/ CSV	https://github.com/FunctionLab/ExPecto	[138]
PEDIA (Prioritisation of exome data by image analysis)	CNN	To prioritise variants and genes for diagnosis of patients with rare genetic disorders	VCF / CSV	https://github.com/PEDIA-Charite/PEDIA-workflow	[148]
DeepMILO (Deep learning for Modeling Insulator Loops)	CNN + RNN	to predict the impact of non-coding sequence variants on 3D chromatin structure	FASTA / TSV	https://github.com/khuranalab/DeepMILO	[119]
DeepWAS	CNN	To identify disease or trait-associated SNPs	TSV / TSV	https://github.com/cellmapslab/DeepWAS	[19]
PrimateAI	CNN	To classify the pathogenicity of missense mutations	CSV / CSV + txt	https://github.com/Illumina/PrimateAI	[27]
DeepGestalt	CNN	To Identifying facial phenotypes of genetic disorders	Image / txt	Is available through the Face-2Gene application, http://face2gene.com	[149]
DeepMiRGene	RNN, LSTM	To predict miRNA precursor	FASTA / Cross-Validation (CV)-Splits file	https://github.com/eleventh83/deepMiRGene	[150]
Basset	CNN	To predict the causative SNP with sets of related variants	BED, FASTA/ VCF	https://github.com/davek44/Basset	[151]

medicine and healthy population assessment (Sundaram et al., 2018). Although statistical methods, such as GWAS, have had huge success in combining genetic variants to disorders, they still require heavy sampling to distinguish rare genetic variants and cannot deliver information about de novo variants (Fu et al., 2014). Thus, current annotation approaches, such as PolyPhen [58], SIFT [59] and GERP [60], represent beneficial methods for prioritising the causative variants, despite facing some drawbacks. For such problems, DL-based models have been implemented to enable a powerful method for exploiting the deep neural network (DNN) architecture to prioritise variants, for instance, the Basset model, a variant annotator, that relies on a CNN algorithm and is designed to predict the causative SNP exploiting DNase I hypersensitivity sequencing data as an input (Kelley, Snoek and Rinn, 2016).

The clinical and molecular validations cannot be replaced by in silico prediction models; however, in a sense, they can contribute to decrease waiting times for results and can prioritise variants for further functional analysis. These predictable models are mainly suitable when several poorly understood candidate variants convey certain phenotypes [27]. Medical genetics has been significantly transformed following the proposition of NGS technology, particularly with WGS because of its power to interpret genomic variations in both coding and non-coding fragments within the entire human genome. Recently, several ML-based methods have offered to prioritise non-coding variants; still, the recognition of disease-associated variants in complex traits, such as cancers, is challenging. Plus, the majority of positive variants associated with a certain phenotype is required to predict general and precise novel correlations (Schubach et al., 2017). Lately, several DL approaches have been proposed to overcome these challenges. For example, the DeepWAS model relies on a CNN algorithm that allows regulatory impact prediction of each variant on numerous cell-type-specific chromatin features. The key result of the DeepWAS model is the direct determination of the disease-associated SNPs with a common effect on a certain chromatin trait in the related tissue. The DeepWAS model demonstrated the ability to detect the disease-relevant, transcriptionally active genomic position after combining the expression and methylation quantitative-trait loci data (eQTL and meQTL, respectively) of various resources and tissues [19]. Nevertheless, several deep learning algorithms have been described as discovering novel genes. For this reason, deep learning approaches are particularly suited for variant investigation for genes not yet related to specific disease phenotypes [61, 62].

Gene expression and regulation

In this section, we focused on the most efficient deep learning-based tools in the area of gene expression and regulation in the genome. We listed several models applying various deep learning algorithms and summarised the information and source codes mostly in splicing and gene expression applications, if available (Table 3).

Gene expression involves the initial transcriptional regulators (e.g. pre-mRNA splicing, transcription and polyadenylation) to functional protein production [63]. The high-throughput screening technologies that test thousands of synthetic sequences have provided rich knowledge concerning the quantitative regulation of gene expression, although with some limitations. The main limitation is that huge biological sequence regions cannot be explored using experimental or computational techniques [64]. Although recent NGS technology has provided great knowledge in the gene-regulation field, the majority of natural mRNA screening approaches still utilise chromatin accessibility, ChIP-seq and DNase-seq information; they focus on studying promoter regions. Therefore, a robust method is required to understand the relationship between various regions of gene regulatory structures and their networks expression connection [65]. Likewise, the current technology in RNA sequencing has empowered the direct sequencing of single cells, identified as single-cell RNA sequencing (scRNA-seq), that permits querying biological systems at unique intention. For example, the data of scRNA-seq produce valuable information into cellular heterogeneity that could expand the interpretation of human diseases and biology [66, 67]. Its major applications of scRNA-seq data understanding involved in detecting the type and state of the cells [68, 69]. However, the two main computational questions include how to cluster the data and how to retrieve them [70].

Deep learning has empowered essential progress for constructing predictive methods linking regulatory sequence elements to the molecular phenotypes [71–74]. Just recently, Gundogdu and his colleagues (2022) demonstrate an excellent classification model based on deep neural networks (DNNs). It constricted numerous types of previous biological information on functional networks between genes to understand a biological significant illustration of the scRNA-seq data [70]. Moreover, Li et al. (2020) present a DESC an unsupervised deep learning algorithm implemented based on python, which understands iteratively representation of cluster-specific gene expression and the scRNA-seq analysis cluster tasks [75]. Further, deep learning model has also been applied for single-cell sequencing data. Its deep neural network (DNN) model designed to measure the immune infiltration in both colorectal and breast cancers bulk scRNA-seq data. This approach permits quantifying a particular type of immune cells such as CD8 + and CD4Tmem plus the general population of lymphocytes together with Stromal content and B cells [76].

Recently, Jaganathan et al. (2019) constructed SpliceAI, a deep residual neural network that predicts splice function using only pre-mRNA transcript sequencing as inputs. An architecture contained a 32-dilated convolutional layer employed to identify sequence determinates crossing enormous genomic gaps since there are tens of thousands of nucleotides separated splice-donors and splice-acceptors [71].

Many experimental datasets, such as the ChIP-seq and DNase-seq assays, do not measure the effects on gene expression directly; however, they are an ideal complement to deep neural network methods. For instance, Movva et al. (2019) introduced the MPRA-DragoNN model, based on CNN architecture for prediction and analysis of the transcription regulatory activity of non-coding DNA sequencing data measured from (MPRAs) data. Approximately 16 K distinct regulatory regions in K562 and HepG2 cell lines of 295 bp *cis*-regulatory elements cloned upstream of either minimal-promoter or strong-promoter used in the Sharpr-MPRA evaluation [77]. A very contemporary DL model, introduced by Agarwal and Shendure, named the Xpresso model, a deep convolutional neural network (CNN), conjointly models the promoter sequence and its related mRNA stability features to predict the gene expression levels of mRNA. Interestingly, Xpresso models are simple to train at several arbitrary cell types, even when they lack experimental information, such as ChIP and DNase [73]. Zhang Z. et al. (2019) developed a deep learning-based model called DARTS; deep learning augmented RNA-seq analysis of transcript splicing, that use a wide-ranging RNA-seq resources of a various alternative splicing. It consists of two main modules: deep neural network (DNN) and Bayesian hypothesis testing (BHT) [78]. More DL-based models (specifically, four different CNN architectures) designed by Bretschneider et al. (2018), named the competitive splice site model (COSSMO), which adapts to various quantities of alternative splice sites and precisely estimates them via genome-wide cross-validation. The frameworks consist of convolutional layers, communication layers, long short-term memory (LSTM) and residual networks, correspondingly, to discover related motifs from DNA sequences. In every putative splice site, the used model inputs are DNA and RNA sequences with 80 nucleotide-wide windows around the alternative splice sites and opposite constitutive splice sites together with the intron length. The outputs of

Table 3 Genomic tools/algorithm based on deep learning architecture for gene expression regulation

Tools	DL model	Application	Input/Output	Website Code Source	References
DanQ	CNN + BLSTM	To predict DNA function directly from sequence data	.mat /.mat	https://github.com/uci-cbcl/DanQ	[152]
SPEID	CNN + LSTM	For enhancer–promoter interaction (EPI) prediction	.mat /.mat	https://github.com/ma-compbio/SPEID	[153]
EP2vec	NLP + GBRT	To predict enhancer–promoter interactions (EPIs)	CSV / CSV	https://github.com/wanwenzeng/ep2vec	[154]
D-GEX (deep learning for gene expression)	FNN	To understand the expression of target genes from the expression of landmark genes	.cel, txt, BAM / txt	https://github.com/uci-cbcl/D-GEX	[155]
DeepExpression	CNN	To predict gene expression using promoter sequences and enhancer–promoter interactions	.txt /.txt	https://github.com/wanwenzeng/DeepExpression	[156]
DeepGSR	CNN + ANN	To recognise various types of genomic signals and regions (GSRs) in genomic DNA (e.g. splice sites and stop codon)	FASTA /.txt	https://zenodo.org/record/1117159#.Xp4B4y2B1p8	[157]
SpliceAI	CNN	To identify splice function from pre-mRNA sequencing	VCF / VCF	https://github.com/Illumina/SpliceAI	[71]
SpliceRover	CNN	For splice site prediction	FASTA /.txt	N/A	[158]
Splice2Deep	CNN	For splice site prediction in Genomic DNA	FASTA /.txt	https://github.com/SomayahAlbaradei/Splice_Deep	[29]
DeepBind	CNN	To characterise DNA- and RNA-binding protein specificity	FASTA /.txt	https://github.com/MedChaabane/DeepBind-with-PyTorch	[111]
Gene2vec	NLP	To produce a representation of genes distribution and predict gene–gene interaction	.txt /.txt	https://github.com/jingcheng-du/Gene2vec	[130]
MPRA-DragoNN	CNN	To predict and analyse the regulatory DNA sequences and non-coding genetic variants	N/A	https://github.com/kundajelab/MPRA-DragoNN	[77]
BiRen	CNN + GRU + RNN	For enhancers predictions	BED, BigWig /CSV	https://github.com/wenjiegroup/BiRen	[159]
APARENT (APA REgression NeT)	CNN	To predict and engineer the human 3′ UTR Alternative Polyadenylation (APA) and annotate pathogenetic variants	FASTA / CSV	https://github.com/johli/aparent	[72]
LaBranchoR (LSTM Branchpoint Retriever)	BLSTM	To predict the location of RNA splicing branchpoint	FASTA / FASTA	https://github.com/jpaggi/labranchor	[160]
COSSMO	CNN, BLSTM + ResNet	To predict the splice site sequencing and splice factors	TSV, CSV /CSV	http://cossmo.genes.toronto.edu/	[79]
Xpresso	CNN	To predict gene expression levels from genomic sequence	FASTA /.txt	https://github.com/vagarwal87/Xpresso	[73]
DeepLoc	CNN + BLSTM	To predict subcellular localisation of protein from sequencing data	FASTA/ prediction score	https://github.com/JJAlmagro/subcellular_localization	[161]
SPOT-RNA	CNN	To predict RNA Secondary Structure	FASTA /.bpseq,.ct, and. prob	https://github.com/jaswindersingh2/SPOT-RNA/	[162]
DeepCLIP	CNN + BLSTM	For predicting the effect of mutations on protein–RNA binding	FASTA /.txt	https://github.com/deepclip/deepclip	[163]

Table 3 (continued)

Tools	DL model	Application	Input/Output	Website Code Source	References
DECRES (DEep learning for identifying Cis-Regulatory ElementS)	MLP + CNN	To predict active enhancers and promoters across the human genome	FASTA /.txt	https://github.com/yifeng-li/DECRES	[74]
DeepChrome	CNN	For prediction of gene expression levels from histone modification data	Bam / TSV	https://github.com/QData/DeepChrome	[164]
DARTS	DNN + BHT	Deep learning augmented RNA-seq analysis of transcript splicing	.txt	https://github.com/Xinglab/DARTS	

the model are predictions of percent selected index (PSI) distribution of every putative splice-site. All of COSSMO model's performance exceeds MaxEntScan; however, there were large performance variances among the four frameworks, in which recurrent LSTM reached the best accuracy over the communication networks, which did not consider the splice-site ordering [79]. However, to learn the automated relationships among heterogeneous datasets in imperfect biological situations, deep learning models offer unprecedented opportunities.

Epigenomics

This section discusses some epigenomics challenges and summarises up-to-date deep learning models in epigenomics, their implementation, data types and source code (Table 4). Modifications in phenotypes that are not based on genotype modifications are referred to as epigenetics. It is defined as the study of heritable modifications in gene expressions which does not include DNA sequence modifications [80]. Epigenomic mechanisms, including DNA methylation, histone modifications and non-coding RNAs, are considered fundamental in understanding disease developments and finding new

Table 4 Genomic tools/algorithm based on deep learning architecture for epigenomics

Tools	DL model	Application	Input/Output	Website Code Source	References
DeepSEA	CNN	To predict multiple chromatin effects of DNA sequence alterations	N/A	https://github.com/Team-Neptune/DeepSea	[165]
FactorNet	CNN + RNN	For predict the cell-type specific transcriptional binding factors (TF)	BED / BED, gzipped bedgraph file	https://github.com/uci-cbcl/FactorNet	[120]
DeMo (Deep Motif Dashboard)	CNN + RNN	For transcription factor binding site perdition (TFBS) by classification task	FASTA / txt	https://github.com/const-ae/Neural_Network_DNA_Demo	[166]
DeepCpG	CNN + GRU	To predict the methylation states from single-cell data	TSV / TSV	https://github.com/cangermueller/deepcpg	[83]
DeepHistone	CNN	To accurately predict histone modification sites based on sequences and DNase-Seq (experimental) data	txt, CSV / CSV	https://github.com/ucrbioinfo/DeepHistone	[84]
DeepTACT	CNN	To predict 3D chromatin interactions	CSV / CSV	https://github.com/liwenran/DeepTACT	[167]
Basenji	CNN	To predict cell-type-specific epigenetic and transcriptional profiles in large mammalian genomes	FASTA / VCF	https://github.com/calico/basenji	[114]
Deopen	CNN	To predict the chromatin accessibility from DNA sequence/ Downstream analysis also included QTL analysis	BED, hkl /hkl	https://github.com/kimmo1019/Deopen	[31]
DeepFIGV (Deep Functional Interpretation of Genetic Variants)	CNN	To predicts impact on chromatin accessibility and histone modification	FASTA / TSV	http://deepfigv.mssm.edu	[62]

treatment targets. Although in clinical implementations, epigenetics has yet to be completely employed. Recently, complications initiated in developing data interpretation tools to advances in next-generation sequencing and microarray technology to produce epigenetic data. The insufficiency of suitable and efficient computational approaches has led current research to focus on a specific epigenetic mark separately, although several mark interactions and genotypes occurred in vivo [81]. Several previous studies have disclosed the fundamental applications of deep learning models in epigenomics. They reached unlimited success in predicting 3D chromatin interactions, methylation status from single-cell datasets and histone modification sites based on DNase-Seq data [62, 82–84].

Liu et al. (2018) introduced a hybrid deep CNN model, Deopen, which was applied to predict chromatin accessibility within a whole genome from learned regulatory DNA sequence codes. In order to analytically evaluate Deopen's function in capturing the accessibility codes of a genome, a series of experiments were conducted from the perspective of binary classification [31]. As an example of Deopen applications, in the androgen-sensitive human prostate adenocarcinoma cell lines (LN-CaP), the EGR1 recovered by the Deopen model is assumed to play a critical role as a treatment target in gene therapy for prostate cancer [31, 85]. Recently, Yin et al. (2019) proposed the DeepHistone framework, a CNN-based algorithm to predict the histone modifications to various site-specific markers. For precise predictions, this model combines DNA sequence data with chromatin accessibility information. It has revealed the capability to discriminate functional SNPs from their adjacent genetic variants, thus having the possibility to be utilised for investigating functional impacts of putative disorder-related variants [84]. Hence, efficient deep learning models are necessary for genome research to elucidate the epigenomic modifications' impact on the downstream outputs.

Pharmacogenomics

We listed the most deliberated deep learning pharmacogenomics models, their common purposes, input/output formats and the source of code (Table 5). Although there has been a great interest in deep learning approaches in the last few years, until very recently, deep learning tools have been rarely employed for pharmacogenomics problems, such as to predict drug response [86]. Knowledge concerning the association between genetic variants in enormous gene clusters up to whole genomes and the impacts of varying drugs is called pharmacogenomics [87]. A key challenge in modern therapeutic methods is understanding the underlying mechanisms of variability. Sometimes the medication response distribution through a certain population is evidently bimodal, proposing a dominant function for one variable, which is usually genetic. Nonetheless, an understanding of the underlying mechanisms of pharmacokinetics or pharmacodynamics could be utilised to detect candidate genes, wherein the function of those gene variants could explicate various drug reactions (88). The clinical experiments generate various errors during the investigation of drug combination efficiency, which is time- and cost-intensive. Besides, it could expose the patient to excessive risky therapy [89, 90]. In order to identify alternative drug synergy strategies without harming patients, high-throughput screening (HTS) using several concentrations of a

Table 5 Genomic tools/algorithm based on deep learning architecture for pharmacogenomics

Tools	Function	DL model	Application	Input/Output	Website Code Source	References
DeepDR	Drug Repositioning	DNN	To translate pharmacogenomics features identified from in vitro drug screening to predict the response of tumours	txt / txt	https://github.com/ChengF-Lab/deepDR	[97]
DNN-DTI (Drug–target interaction prediction)	Database	DNN	To predict drug-target interaction	txt / txt	https://github.com/JohnnyY8/DNN-DTI	[168]
DeepBL	Antibiotic Resistance	CNN	To predict the beta-lactamase (BLs) using protein or genome sequence datasets	FASTA / CSV	http://deepbl.erc.monash.edu.au	[98]
DeepDrug3D	Binding Site for drugs	CNN	To characterise and classify the protein 3D binding pockets	pdb / txt	https://github.com/pulimeng/DeepDrug3D	[115]
DrugCell	Drug response and synergy for cancer cells	CNN	To predict drug response and synergy	txt / txt	https://github.com/idekerlab/DrugCell	[26]
DeepSynergy	Anticancer drug synergy	FNN	To predict anticancer drug synergy	CSV / CSV	https://github.com/KristinaPreuer/DeepSynergy	[95]

couple of drugs employed to a cancer cell line is utilised [91]. Utilising existing HTS synergy datasets allowed the use of accurate computational models to investigate an enormous synergistic space. Such reliable models would provide direction for both in vitro and in vivo studies, and they are great steps towards personalised medicine, for instance, prediction approaches of anticancer synergic, systems biology [92], kinetic methods [93] and in silico-based models of gene expression screening after single-drug and dose-reaction treatments [94]. Nonetheless, these approaches are limited to particular targets, pathways or certain cell lines and sometimes need a particular omics dataset of treated cell lines with specific compounds [95].

To investigate these pharmacogenomics associations, statistical, such as the analysis of variance (ANOVA) test, is utilised. This can identify, for example, oncogenic changes that occur in patients, which are indicators of drug-sensitivity variances in cell lines. In order to move beyond the drug's relations to the actual drug reaction predictions, numerous statistical and machine learning methods can be employed, from linear regression models to nonlinear ones, such as kernel methods, neural networks and SVM. A central weakness of these approaches is the massive number of inputs feature alongside the low sampling, such as in standard gene expression analysis, and the total number of input genes (or features) exceeds the sample number. An up-to-date strategy to overcome the low sampling number issue is to engage multitasking models [96].

Deep learning methods are reportedly well suited to treatment response prediction tasks based on cell-line omics datasets [95, 97]. One of the examples is, DrugCell, a visible neural network (VNN) interpretation model for the structure and function of human cancer cells in therapy response. It pairs the model's central mechanisms to the human cell-biology structure. Permitting the prediction of any drug response within any cancer then smartly plans the successful combination of treatments. DrugCell was developed to capture both elements of therapy response in an explainable model with two divisions, the VNN-integrating cell genotype and the artificial neural network (ANN)-integrating drug design. The first VNN model inputs comprise text files of the hierarchal association between molecular sub-systems in human cells, which contain 2086 biological process standards in the Gene Ontology (GO) database. The second ANN model inputs were conventional ANN integrating text files of the Morgan fingerprint of medicine, the chemical structure of a canonical vector symbol. The outputs from these two divisions were combined into a single layer of neurons that produced the response of a given genotype to a certain therapy. The prediction accuracy of each drug separately revealed a drug sub-population with significant accuracy. This, in turn, competes with the state-of-the-art regression methods applied in previous models to predict the drug response. Additionally, comparing DrugCell with a parallel neural network model trained merely on drug design and labelled tissue extremely outperformed the tissue-based model. This means that DrugCell has learned data from somatic mutations exceeding the tissue-only method [26]. Another recent model called DeepBL is based on deep learning architecture executed based on Small VGGNet structure (a type of CNNs) and TensorFlow library. This approach detects the beta-lactamases (BLs) and their varieties that provide resistance to beta-lactam antibiotics, with protein sequences as inputs. It is based on well-interpreted massive RefSeq datasets covering > 39 K BLs extracted from the NCBI database. Comparing this model with the other conventional machine learning-based algorithms, including SVM, RF, NB and LR, DeepBL outperformed them after evaluation on an independent test set comprising more than 10 K sequences [98]. Until very recently, deep learning applications in pharmacogenomics remained under consideration.

Deep learning algorithms/techniques used in genomics

The accomplishment of the recent, attainable models mentioned in deep learning tools/software/pipelines in genomics section suggests that deep learning is a powerful technique in genomic research. Here, we focus on deep learning algorithms recently applied in genomic applications: convolutional neural networks (CNNs), feedforward neural networks (FNN), natural language processing (NLP), recurrent neural networks (RNNs), long short-term memory networks (LSTMs), bidirectional long short-term memory networks (BLSTMs) and gated recurrent unit (GRU; Table 6; Fig. 1).

Deep learning is a contemporary and rapidly expanding subarea of machine learning. It endeavours to model concepts from wide-ranging data by occupying multi-layered DNNs, hence creating data logic, such as pictures, sounds and texts. Generally, deep learning has two features: first, the structure of nonlinear processing parts is multiple layers, and second, the feature extraction fashion on each layer is either the supervised or unsupervised method [99]. In the 1980s, the initial deep learning architecture was constructed on artificial neural networks (ANNs) [100], but the actual power of deep learning developed outward in 2006 [101, 102]. Since then, deep learning has been functional in various arenas involving genomics, bioinformatics, drug discovery, automated speech detection, image recognition and natural language processing [6, 13, 103].

Table 6 Deep learning algorithms in genomics and their original development and applications

ANN Algorithms	Natural Language Processing (NLP)	Feedforward neural network	Convolutional neural network (CNN)	Recurrent neural networks (RNNs)	Bidirectional long short-term memory networks (BLSTMs)	Long short-term memory networks (LSTMs)	Gated recurrent unit (GRU)
Algorithm Inventor	Applied in dictionary look-up system developed at Birkbeck College, London	Frank Rosenblatt	It was named as "neocognitron" by Fukushima	Rumelhart, Hinton and Williams	Schuster and Paliwal	Hochreiter and Schmidhuber	Cho et al
Year of Development	1948	1958	1980	1986	1997	1997	2014
Year of Initial Genomics' Function	1996	1993	2015	2005	2015	2015	2017
First User in Genomics	Schuler et al	S Eskiizmililer	Alipanahi et al	Maraziotis, Dragomir and Bezerianos	Quang and Xie	Quang and Xie	Angermueller et al
First Genomic Application	Entrez databases	Karyotyping architecture based on Artificial Neural Networks	DeepBind	Predicting the complicated causative associations between genes from microarray datasets based on recurrent neuro-fuzzy technique	DanQ model	DanQ model	DeepCpG
Genomic Function Exemplar(s)	Genetic counsellors AI-based chatbots and EPIs prediction	Karyotyping, Prenatal diagnostic for early detection of aneuploidy syndrome	Prediction of variant impacts on expression and disease risk, predicting drug response of tumours from genomic profiles, and pharmacogenomics	Predicting transcription factor binding sites, for Alignment and SNV identification	DNA function predictions and prediction of protein localisation, predict miRNA precursor	Enhancer–promoter interaction (EPI) prediction	Enhancers and methylation states predictions
Landmark References	[128, 169, 170]	[171–173]	[97, 111, 174–176]	[24, 116, 118, 177, 178]	[122, 123, 179, 180]	[16, 121, 123]	[126, 181]

Artificial neural networks (ANNs) were motivated by the human brain's neurons and their networks [104]. They consist of clusters of fully connected nodes, or neurons, demonstrating the stimulus circulation of synapses in the brain through the neural networks. This architecture of deep learning networks is utilised for feature extraction, classification, decreased data dimensions or sub-elements of a deeper framework such as CNNs [105].

Multi-omics study generates huge volumes of data, as mentioned earlier, basically because of the evolution that has been pursued in genomics and improvements in biotechnology. Symbolic examples involve the high-throughput technology, which extent thousands of gene expression or non-coding transcription, such as miRNAs. Moreover, the genotyping platforms and NGS techniques and the associated GWAS that generates measurable gene expression reports, such as RNA-Seq, discover numerous genetic variants, together with further genomic modifications in various populations [11]. However, some DL models rely purely on DNA sequence datasets that seemingly lack the power to create predictions of a cell-line-exclusive method due to the identical DNA sequencing of various cell lines. In order to overcome this deficiency, several hybrid deep learning models have been advised and revealed obvious enhancement in certain studies through joining DNA sequencing data with biological experiments information [84].

Feedforward Neural Networks (FNNs) Are a type of artificial neural network that consists of one forward direction network starting from input layers, crossing the hidden layers and reaching to the output layer, without forming loops such as RNNs [106]. It is used in genomics to comprehend the expression of target genes from the expression of landmark genes using the D-GEX model [12]. Moreover, active enhancers and promoters have been predicted across the human genome utilising the DECRES model [107]. Moreover, anticancer drug synergy predictions have been made via the DeepSynergy model [95].

Convolutional Neural Networks (CNNs) Also called ConvNet, CNN is a deep learning algorithm that has a deep feedforward architecture consisting of various building blocks, such as convolution layers, pooling layers and fully connected layers [97, 108]. It illustrates a fully connected network since each node in a single layer is fully connected to the entire node of the next layer. The convolution units in the CNN layers can obtain the input data from units of the earlier one, which all together generate a prediction. The key principle of such deep construction is that massive processing and connection feature represents inferring nonlinear association between both inputs and outputs [109, 110]. The most common analysis uses of CNNs were applied in graphical images and were initially considered a fully automated image network interpreter for classifying handcraft fonts [105].

For genomic functions, CNNs considered the dominant algorithm utilised genomic information (Fig. 2). The primary CNN implementation, DeepBind, was proposed by [111] for binding protein predictions and showed greater prediction power than conventional models (Table 6). More examples of CNN are used as a single algorithm in gene expression, and regulations include the DeepExpression model, which has been effectively used to predict gene expression using promoter sequences and enhancer–promoter interactions [112]. The SpliceAI model was introduced to identify splice function from pre-mRNA sequencing [71]. Further, the SPOT-RNA model was developed for predicting RNA secondary structure [16]. CNN was also used for DNA sequencing in call genetic variants, such as Clairvoyante, IntelliNGS and DeepSV models [52, 54, 113]. In epigenomics, the DeepTACT model was used for predicting the 3D chromatin interactions [82], and the Basenji model was employed for predicting cell-type-specific epigenetic and transcriptional profiles in large mammalian genomes [114]. In disease variants, the ExPecto model was used to predict tissue-specific transcriptional effects of mutations/functions [32], and the DeepWAS model was used to identify disease or trait-associated SNPs [19]. Finally, in pharmacogenomics applications, CNN was utilised to create the DrugCell model for drug response and synergy predictions [26]. Additionally, the DeepDrug3D model was obtained for characterising and classifying the 3D protein binding pockets [115].

Additionally, CNN algorithms were combined with other algorithms to build up efficient approaches in epigenomics, combining CNN with GRU to predict the methylation states from single-cell data [83], while in terms of gene expression and regulation, [74] linked CNN algorithms with MLP in the DECRES model to predict active enhancers and promoters across the human genome. Besides, [116] used CNN with RNN algorithms in a DNA sequencing application to create the DAVI model and identify NGS read variants.

Recurrent neural networks (RNNs) are ANNs with a recurrent layer consisting of typical recurrent layers that enable state updates of past and current inputs with feedback connections. They are distinguished by the internal cycle connections between recurrent layer units and are concerned with sequential datasets [117, 118]. Recurrent neural networks have regularly expended for the task that comprised in learning sequencing datasets, such as translation languages and recognising speech. However, it has not been utilised widely on DNA sequencing data which is the data style

where the order link between bases are crucial for its assessment [119]. Maraziotis et al. [24] initiated RNN implementation in genomics using microarray experimental data based on recurrent the neuro-fuzzy protocol to infer the complicated causative relationship between genes by predicting the time-series of gene expression (Table 6).

Most RNNs are applied in genomics combined with other algorithms, such as CNNs. For example, to identify NGS read variants, the DAVI model introduced the combination of CNN and RNN algorithms [116]. The FactorNet model was designed based on both CNN and RNN algorithms and raised to predict the cell-type-specific transcriptional binding factors (TFBSs) [120]. However, CNN algorithms are perfect at capturing local DNA sequence patterns; contrastingly, RNN derivatives, such as LSTM, are ideal for capturing long-distance dependencies between sequence datasets [119].

Long short-term memory networks (LSTMs) are standard recurrent cells with "gates" to handle long-term dependency tasks [118]. They deliberate to prevent long-term dependency difficulties through their competence in acquiring long-term dependencies. It has a node, input gate, output gate and forget gate as core LSTM unit. The node considers values through certain time gaps, whereas the input and output gates control information flow [121]. The preliminary implementations of LSTM algorithms in genomics advised the SPEID model, which used a pattern of deep learning algorithms utilising both LSTM and CNN for EPI predictions (Table 6; [18]). Park et al.[122] obtained DeepMiRGene, a fusion of the RNN and LSTM models, to predict miRNA precursors.

Bidirectional Long Short-Term Memory Networks (BLSTMs) In BLSTM, two RNNs with two hidden layers (forward and backward layers) can be trained in both time directions in parallel to enable the previous context usage that cannot be accomplished via standard RNNs [118]. Quang et al. [123] expressed the DanQ model, the original employment in genomics that predicted DNA function directly from sequence data developed from CNN and BLSTM constructions (Table 6). Later, [124] presented DeepCLIP, also utilising CNN and BLSTM, to predict the effect of mutations on protein–RNA binding.

Gated Recurrent Unit (GRU) is categorised as a variant of the LSTM algorithm with cell has only "two gates": the update gate and reset gate [118]. It couples neural networks opposing each other. The first network produces artificial, accurate information, while the second estimates the validity of the information [125]. It was initially applied in gene expression and regulation by [126], who presented the BiRen model, an architecture consisting of RNNs, CNNs and GRUs, to predict enhancers (Table 6). After, the DeepCpG model appeared, combining CNN and GRU frameworks to predict the methylation states from single-cell data [83].

Natural Language Processing (NLP) It examines the computers usage to recognise human languages for the purpose of executing beneficial tasks [127]. In the field of NLP, in fact, the "distributed representations" technique is utilised in several state-of-the-art DL models [128]. For example, the word2vec model is an achieved NLP that utilises the distribution representation process, "neural embedding". This is because of the embedding task that is frequently expressed through neural networks beside numerous parameters. The aim of word embedding is to convey linear mapping and then generate a direct advantage of representing a single word, thereby distinguishing vectors in continuous space and hence become open for backpropagation-based methods in neural networks [129]. In terms of deep learning demands in the field of gene expression and regulation, Du et al. (2019) explored the Gene2vec model, an idea of distributed representation of genes. It engages genes' natural contexts and their expression and co-expression patterns from GEO data. The essential layer of a multilayer neural network uses the embedded gene, which predicts gene-to-gene interactions with a 0.72 AUC score. This is an interesting outcome because the initial model input is the names of two genes merely. Thus, the distributed representation of genes technique is burdened with rich indications about gene function [130]. Another NLP implementation in the same field was shown by Zeng et al. (2018), who combined NLP with GBRT and introduced the EP2vec model to EPIs.

Graphical Neural Network (GNN) Due to the emerging biological network data sets in genomics, graph neural network has been evolved as an important deep learning method to tackle these data sets[131]. GNN was proposed by Gori et al. (2005) as a novel neural network model to tackle graph structure data [132]. Out of many applications of GNN in analysing multi-omics data, the few salient ones are disease gene prediction, drug discovery, drug interaction network, protein–protein interaction network and biomedical imaging. GNN is capable of modelling both the molecular structure data [133] and biological network data[134].

Deep learning resources for genomics

We collected the most efficient user-friendly genomic resources developed based on deep learning architectures (Table 7). The adoption of various deep learning solutions and models is still limited, despite the enormous success of these tools in genomics and bioinformatics. One reason for this is the lack of deep learning-based published protocols to adapt to new, heterogeneous datasets requiring significant data engineering [135]. In genomics,

Table 7 Deep learning packages and resources

Resource Name	Category	Application	Date created	Link	Free/paid
Libraries					
Janggu[a]	Python package	facilitates deep learning in the context of genomics	2020	https://github.com/BIMSBbioinfo/janggu	Free
ExPecto[a]	Python-based repository	Contains code for predicting expression effects of human genome variants ab initio from sequence	2018	https://github.com/FunctionLab/ExPecto	Free
Selene[a]	PyTorch-based Library	A library for biological sequence data training and model architecture development	2019	https://selene.flatironinstitute.org/	Free
Pysster[a]	TensorFlow-based Library	Used for learning sequence and structure motifs In biological sequences using convolutional neural networks	2018	https://github.com/budach/pysster	Free
Kipoi[a]	Python package	Kipoi is an API and a repository of ready-to-use trained models for genomics	2019	https://github.com/kipoi/kipoi http://kipoi.org/	Free
Compute platform					
Google Colaboratory (Colab)	PnP GPUs	Colab allows anybody to write and execute arbitrary python code through the browser, and is especially well suited to machine learning, data analysis and education	2017	https://colab.research.google.com/	Free
IBM Cloud	Cloud service	Cloud computing platform; Design complex neural networks, then experiment at scale to deploy optimised learning models within IBM Watson Studio	2011	https://www.ibm.com/cloud	Free tier Cost tier
Google CloudML	PnP GPUs	For extreme scalability in the long run	2008	https://cloud.google.com/ai-platform	Paid
Vertex AI	AI platform	Google Cloud's new unified ML platform	2021	https://cloud.google.com/vertex-ai	
Amazon EC2	Cloud service	A website facility which delivers secure, scalable compute power in the cloud	2006	https://aws.amazon.com/ec2/	Free Paid

[a] These deep learning libraries/packages are specific to Genomic application

high-throughput data (e.g. WGS, WES, RNA-seq, ChIP-seq, etc.) are utilised to train neural networks and have become typical for disease predictions or understanding regulatory genomics. Similarly, developing new DL models and testing current models on new datasets face great challenges due to the lack of inclusive, generalisable, practical deep learning libraries for biology [136]. In this respect, software frameworks and genomic packages are necessary to allow rapid progress in adopting a novel research question or hypothesis, combining original data or investigating using different neural network structures [135]. In order to facilitate the DL model implementation in genomics, the following software packages or libraries could become critical for genomic scientists and biomedical researchers.

Janggu is a deep learning python library based on deep CNN for genomic implementations. It aims at a data-procuring facility and model assessment by supporting flexible neural network prototype models. The Janggu library provides three use cases: transcriptional factor predictions, utilising and enhancing the published deep learning designs and predicting the CAGE-tag count normalisation of promoters. This library offers easy access and pre-processing to convert data from standard file formats (e.g. FASTA, BAM, Bigwig, BED and narrowPeak) to BigWig files [135].

Selene is a deep learning library based on PyTorch for biological sequence data training and model architecture development. Selene supports the prediction of genetic variant effects and visualises the variant scores as a Manhattan plot. It also automatically generates training, testing and validation split from the given input dataset. Further, Selene automatically trains the data and can examine the model on a test set, thereby producing a visualised figure to display the model's performance [137].

ExPecto is a variant prioritisation model for predicting the gene expression levels from a broad regulatory region (~ 40 kb) range of promoter-proximal sequencing regions. It relies on CNN to convert the input sequences into epigenomic features. ExPecto facilitates rare variants or unprecedented variants prediction. This is because of its unique design architecture, which does not utilise any variant information during the training process. ExPecto processes VCF files and outputs CSV files [138].

Pysster is a python library package based on CNN for biological sequencing data training and classification. Pysster provides automatic hyperparameter optimisation and motif visualisation options along with their position and class enrichment information [139].

Kipoi (Greek for "gardens"; pronounced "kípi") is a genomic repository for sharing and reusing trained genome-related models. Kipoi provides more than 2 K distinctly trained models from 22 different studies covering significant predictive genomic tasks. The prediction includes chromatin accessibility determination, transcription factor binding and alternative splicing from DNA sequences [136].

Implementation of these deep learning, genome-based libraries/packages requires accessing the computer power and familiarity with web-based resources (Table 7). Several major cloud-computing platforms have proposed on-demand GPU access in user-friendly manners, including Google CloudML, IBM cloud, Vertex AI and Amazon EC2 [140–142]. User configuration and the installation of the appropriate environments for general GPU coding are required in these cloud-based machines. Concurrently, for users who need to avoid semi-manual setup methods, an expert plug-and-play (PnP) platform GPU access is offered, such as Google Colaboratory (Colab). Google Colab is considered the simplest alternative python-based notebook and provides free K80 GPU utilisation for 12 continuous hours [143, 144]. Links to the resources (packages/libraries and web platforms) for the application of deep learning in genomics are provided in Table 7.

Conclusion

This manuscript catalogues different deep learning tools/software developed in different subareas of genomics to fulfil the predictive tasks of various genomic analyses. We discussed, in detail, the data types in different genomics assays so that readers could have primary knowledge of the basic requirements to develop deep learning-based prediction models using human genomics datasets. In the later part of the manuscript, different deep learning architectures were briefly introduced to genomic scientists in order to help them decide the deep learning network architecture for their specific data types and/or problems. We also briefly discussed the late application of the deep learning technique in genomics and its underlying causes and solutions. Towards the end of the manuscript, various computational resources, software packages or libraries and web-based computational platforms are provided to act as pointers for researchers to create their very first deep learning model utilising genomic datasets. In conclusion, this timely review holds the potential to assist genomic scientists in adopting state-of-the-art deep learning techniques for the exploration of genomic NGS datasets and analyses. This will certainly be beneficial for biomedicine and human genomics researchers.

Abbreviations

NGS: Next-generation sequencing; WGS: Whole genome sequencing; WES: Whole exome sequencing; SMS: Single-molecule sequencing; RNA-seq: RNA sequencing; ChIP-seq: Chromatin immunoprecipitation sequencing; PacBio: Pacific biosciences; ONT: Oxford nanopore technology; MPRAs: Massively parallel reporter assays; miRNA: MicroRNAs; GWAS: Genome-wide association study; PSI: Percent selected index; HGVS: Human genome variation society; IMSGC: International multiple sclerosis genetics consortium; VUS: Variant of uncertain significance; CADD: Combined annotation dependent depletion; GATK: Genomic Analysis ToolKit; BAM: Binary alignment map; VCF: Variant call format; FASTA: Text-based format for either nucleotide sequences or amino acids; BED: Browser extensible data; CSV: Comma-separated values; CAGE: Cap analysis of gene expression; GEO: Gene expression omnibus; EPI: Enhancer–promoter interaction; TFBS: Transcription factor binding sites; DL: Deep learning; ML: Machine learning; DNN: Deep neural network; MLP: Multilayer perceptron; CNN: Convolutional neural networks; RNN: Recurrent neural network; LSTM: Long short-term memory network; BLSTM: Bidirectional long short-term memory network; ANN: Artificial neural network; FNN: Feedforward neural networks; NLP: Natural language processing; GRU: Gated recurrent unit; VGGNet: Visual geometry group networks; GBRT: Gradient boosted regression trees; LR: Linear regression; RF: Random forest; NB: Naive Bayes; DBN: Deep belief networks; SVR: Support vector regression; AUC: Area under the curve; auPR: Area under the precision–recall curve; auROC: Area under the receiver operating characteristic.

Acknowledgements

We duly acknowledge Dr. Mohamed Aly Hussain for his motivation and useful discussion regarding the inception of this review article. We also appreciate Dr. Lamya Alomair for her support during the development of this manuscript.

Author contributions

WA and MR conceptualised this study. WA collected the data and performed investigation. MR supervised this study. WA and MR wrote original draft. All authors read and approved the final manuscript.

Funding
This study is not funded by any funding source.

Availability of data and materials
Not applicable.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 24 November 2021 Accepted: 12 July 2022
Published online: 25 July 2022

References

1. Auffray C, Imbeaud S, Roux-Rouquié M, Hood L. From functional genomics to systems biology: concepts and practices. C R Biol. 2003;326(10–11):879–92.
2. Goldfeder RL, Priest JR, Zook JM, Grove ME, Waggott D, Wheeler MT, et al. Medical implications of technical accuracy in genome sequencing. Genome Med. 2016;8(1):24.
3. Goodwin S, McPherson JD, McCombie WR. Coming of age: Ten years of next-generation sequencing technologies. Nat Rev Genet. 2016;17(6):333–51.
4. Yue T, Wang H. Deep Learning for Genomics: A Concise Overview. 2018
5. Honoré B, Østergaard M, Vorum H. Functional genomics studied by proteomics. BioEssays. 2004;26(8):901–15.
6. Talukder A, Barham C, Li X, Hu H. Interpretation of deep learning in genomics and epigenomics. Brief Bioinform. 2020;2:447.
7. Fulco CP, Munschauer M, Anyoha R, Munson G, Grossman SR, Perez EM, et al. Systematic mapping of functional enhancer–promoter connections with CRISPR interference. Science (80-). 2016;354(6313):769–73.
8. Kulasingam V, Pavlou MP, Diamandis EP. Integrating high-throughput technologies in the quest for effective biomarkers for ovarian cancer. Nat Rev Cancer. 2010;10(5):371–8.
9. Nariai N, Kolaczyk ED, Kasif S. Probabilistic protein function prediction from heterogeneous genome-wide data. PLoS One. 2007;2(3):e337.
10. Ritchie MD, Holzinger ER, Li R, Pendergrass SA, Kim D. Methods of integrating data to uncover genotype–phenotype interactions. Nat Rev Genet. 2015;16(2):85–97.
11. Koumakis L. Deep learning models in genomics; are we there yet? Comput Struct Biotechnol J. 2020;18:1466–73.
12. Cao C, Liu F, Tan H, Song D, Shu W, Li W, et al. Deep learning and its applications in biomedicine. Genom Proteom Bioinform. 2018;16(1):17–32.
13. Telenti A, Lippert C, Chang PC, DePristo M. Deep learning of genomic variation and regulatory network data. Hum Mol Genet. 2018;27(R1):R63-71.
14. Kopp W, Monti R, Tamburrini A, Ohler U, Akalin A. Deep learning for genomics using Janggu. Nat Commun. 2020;11(1):3488.
15. Deep learning for genomics. Nat Genet. 2019;51(1):1–1.
16. Singh J, Hanson J, Paliwal K, Zhou Y. RNA secondary structure prediction using an ensemble of two-dimensional deep neural networks and transfer learning. Nat Commun. 2019;10(1):5407.
17. Hsieh T-C, Mensah MA, Pantel JT, Aguilar D, Bar O, Bayat A, et al. PEDIA: prioritization of exome data by image analysis. Genet Med. 2019;21(12):2807–14.
18. Singh R, Lanchantin J, Robins G, Qi Y. DeepChrome: deep-learning for predicting gene expression from histone modifications. Bioinformatics. 2016;32(17):i639–48.
19. Arloth J, Eraslan G, Andlauer TFM, Martins J, Iurato S, Kühnel B, et al. DeepWAS: multivariate genotype-phenotype associations by directly integrating regulatory information using deep learning. PLOS Comput Biol. 2020;16(2):e1007616.
20. Rosenblatt F. The perceptron: a probabilistic model for information storage and organization in the brain. Psychol Rev. 1958;65(6):386–408.
21. Sarker IH. Machine learning: algorithms, real-world applications and research directions. SN Comput Sci. 2021;2(3):160.
22. Wang C, Tan XP, Tor SB, Lim CS. Machine learning in additive manufacturing: state-of-the-art and perspectives. Addit Manuf. 2020;36:101538.
23. Muzio G, O'Bray L, Borgwardt K. Biological network analysis with deep learning. Brief Bioinform. 2021;22(2):1515–30.
24. Marazioti I, Dragomir A, Bezerianos A. Gene networks inference from expression data using a recurrent neuro-fuzzy approach. In: 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference. IEEE; 2005. p. 4834–7.
25. LeCun Y. 1.1 Deep learning hardware: past, present, and future. In: 2019 IEEE International Solid-State Circuits Conference-(ISSCC). IEEE; 2019. p. 12–9.
26. Kuenzi BM, Park J, Fong SH, Sanchez KS, Lee J, Kreisberg JF, et al. Predicting drug response and synergy using a deep learning model of human cancer cells. Cancer Cell. 2020;38(5):672-684.e6.
27. Sundaram L, Gao H, Padigepati SR, McRae JF, Li Y, Kosmicki JA, et al. Predicting the clinical impact of human mutation with deep neural networks. Nat Genet. 2018;50(8):1161–70.
28. Lanchantin J, Singh R, Wang B, Qi Y. Deep motif dashboard: visualizing and understanding genomic sequences using deep neural networks. World Sci. 2017;3:254–65.
29. Albaradei S, Magana-Mora A, Thafar M, Uludag M, Bajic VB, Gojobori T, et al. Splice2Deep: an ensemble of deep convolutional neural networks for improved splice site prediction in genomic DNA. Gene X. 2020;5:100035.
30. Poplin R, Chang PC, 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.
31. Liu Q, Xia F, Yin Q, Jiang R. Chromatin accessibility prediction via a hybrid deep convolutional neural network. Bioinformatics. 2018;2:1147.
32. Zou J, Huss M, Abid A, Mohammadi P, Torkamani A, Telenti A. A primer on deep learning in genomics. Nat Genet. 2019;51(1):12–8.
33. Al-Stouhi S, Reddy CK. Transfer learning for class imbalance problems with inadequate data. Knowl Inf Syst. 2016;48(1):201–28.
34. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genom. 2020;21(1):6.
35. Handelman GS, Kok HK, Chandra RV, Razavi AH, Huang S, Brooks M, et al. Peering into the black box of artificial intelligence: evaluation metrics of machine learning methods. Am J Roentgenol. 2019;212(1):38–43.
36. England JR, Cheng PM. Artificial intelligence for medical image analysis: a guide for authors and reviewers. Am J Roentgenol. 2019;212(3):513–9.
37. Eraslan G, Avsec Ž, Gagneur J, Theis FJ. Deep learning: new computational modelling techniques for genomics. Nat Rev Genet. 2019;20(7):389–403.
38. Zou J, Huss M, Abid A, Mohammadi P, Torkamani A, Telenti A. A primer on deep learning in genomics. Nat Genet. 2019;51(1):12–8.
39. Pérez-Enciso M, Zingaretti LM. A guide for using deep learning for complex trait genomic prediction. Genes (Basel). 2019;10(7):12258.
40. Abnizova I, Boekhorst RT, Orlov YL. Computational errors and biases in short read next generation sequencing. J Proteom Bioinform. 2017;10(1):400089.
41. Ma X, Shao Y, Tian L, Flasch DA, Mulder HL, Edmonson MN, et al. Analysis of error profiles in deep next-generation sequencing data. Genome Biol. 2019;20(1):50.
42. Pfeiffer F, Gröber C, Blank M, Händler K, Beyer M, Schultze JL, et al. Systematic evaluation of error rates and causes in short samples in next-generation sequencing. Sci Rep. 2018;8(1):10950.
43. Horner DS, Pavesi G, Castrignano T, De Meo PD, Liuni S, Sammeth M, et al. Bioinformatics approaches for genomics and post genomics applications of next-generation sequencing. Brief Bioinform. 2010;11(2):181–97.
44. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The genome analysis toolkit: a mapreduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20(9):1297–303.

45. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25(16):2078–9.
46. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. Science. 2012;7:4458.
47. Hwang S, Kim E, Lee I, Marcotte EM. Systematic comparison of variant calling pipelines using gold standard personal exome variants. Sci Rep. 2015;5(1):17875.
48. Kotlarz K, Mielczarek M, Suchocki T, Czech B, Guldbrandtsen B, Szyda J. The application of deep learning for the classification of correct and incorrect SNP genotypes from whole-genome DNA sequencing pipelines. J Appl Genet. 2020;61(4):607–16.
49. Kumaran M, Subramanian U, Devarajan B. Performance assessment of variant calling pipelines using human whole exome sequencing and simulated data. BMC Bioinform. 2019;20(1):342.
50. Luo R, Sedlazeck FJ, Lam T, Schatz MC, Kong H, Genome H. Clairvoyante: a multi-task convolutional deep neural network for variant calling in single molecule sequencing. Science. 2018;3:7745.
51. Cai L, Chu C, Zhang X, Wu Y, Gao J. Concod: an effective integration framework of consensus-based calling deletions from next-generation sequencing data. Int J Data Min Bioinform. 2017;17(2):153.
52. Cai L, Wu Y, Gao J. DeepSV: accurate calling of genomic deletions from high-throughput sequencing data using deep convolutional neural network. BMC Bioinform. 2019;20(1):665.
53. Ravasio V, Ritelli M, Legati A, Giacopuzzi E. GARFIELD-NGS: genomic vARiants FIltering by dEep learning moDels in NGS. Bioinformatics. 2018;34(17):3038–40.
54. Singh A, Bhatia P. Intelli-NGS: intelligent NGS, a deep neural network-based artificial intelligence to delineate good and bad variant calls from IonTorrent sequencer data. bioRxiv. 2019;12:879403.
55. Müller H, Jimenez-Heredia R, Krolo A, Hirschmugl T, Dmytrus J, Boztug K, et al. VCF.Filter: interactive prioritization of disease-linked genetic variants from sequencing data. Nucleic Acids Res. 2017;45(W1):W567-72.
56. Eilbeck K, Quinlan A, Yandell M. Settling the score: variant prioritization and Mendelian disease. Nat Rev Genet. 2017;18(10):599–612.
57. Roy S, Coldren C, Karunamurthy A, Kip NS, Klee EW, Lincoln SE, et al. Standards and guidelines for validating next-generation sequencing bioinformatics pipelines. J Mol Diagn. 2018;20(1):4–27.
58. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods. 2010;7(4):248–9.
59. Ng PC. SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res. 2003;31(13):3812–4.
60. Cooper GM. Distribution and intensity of constraint in mammalian genomic sequence. Genome Res. 2005;15(7):901–13.
61. Boudellioua I, Kulmanov M, Schofield PN, Gkoutos GV, Hoehndorf R. DeepPVP: phenotype-based prioritization of causative variants using deep learning. BMC Bioinform. 2019;20(1):65.
62. Hoffman GE, Bendl J, Girdhar K, Schadt EE, Roussos P. Functional interpretation of genetic variants using deep learning predicts impact on chromatin accessibility and histone modification. Nucleic Acids Res. 2019;3:5589.
63. Tupler R, Perini G, Green MR. Expressing the human genome. Nature. 2001;409(6822):832–3.
64. Zrimec J, Börlin CS, Buric F, Muhammad AS, Chen R, Siewers V, et al. Deep learning suggests that gene expression is encoded in all parts of a co-evolving interacting gene regulatory structure. Nat Commun. 2020;11(1):6141.
65. Zrimec J, Börlin CS, Buric F, Muhammad AS, Chen R, Siewers V, et al. Deep learning suggests that gene expression is encoded in all parts of a co-evolving interacting gene regulatory structure. Nat Commun. 2020;11(1):6141.
66. Angerer P, Simon L, Tritschler S, Wolf FA, Fischer D, Theis FJ. Single cells make big data: new challenges and opportunities in transcriptomics. Curr Opin Syst Biol. 2017;4:85–91.
67. Falco MM, Peña-Chilet M, Loucera C, Hidalgo MR, Dopazo J. Mechanistic models of signaling pathways deconvolute the glioblastoma single-cell functional landscape. NAR Cancer. 2020;2(2):5589.
68. Poulin J-F, Tasic B, Hjerling-Leffler J, Trimarchi JM, Awatramani R. Disentangling neural cell diversity using single-cell transcriptomics. Nat Neurosci. 2016;19(9):1131–41.
69. Darmanis S, Sloan SA, Zhang Y, Enge M, Caneda C, Shuer LM, et al. A survey of human brain transcriptome diversity at the single cell level. Proc Natl Acad Sci. 2015;112(23):7285–90.
70. Gundogdu P, Loucera C, Alamo-Alvarez I, Dopazo J, Nepomuceno I. Integrating pathway knowledge with deep neural networks to reduce the dimensionality in single-cell RNA-seq data. BioData Min. 2022;15(1):1.
71. Jaganathan K, Kyriazopoulou Panagiotopoulou S, McRae JF, Darbandi SF, Knowles D, Li YI, et al. Predicting splicing from primary sequence with deep learning. Cell. 2019;176(3):535-548.e24.
72. Bogard N, Linder J, Rosenberg AB, Seelig G. A deep neural network for predicting and engineering alternative polyadenylation. Cell. 2019;71:9886.
73. Agarwal V, Shendure J. Predicting mRNA abundance directly from genomic sequence using deep convolutional neural networks. Cell Rep. 2020;31(7):107663.
74. Li Y, Shi W, Wasserman WW. Genome-wide prediction of cis-regulatory regions using supervised deep learning methods. BMC Bioinform. 2018;19(1):202.
75. Li X, Wang K, Lyu Y, Pan H, Zhang J, Stambolian D, et al. Deep learning enables accurate clustering with batch effect removal in single-cell RNA-seq analysis. Nat Commun. 2020;11(1):2338.
76. Torroja C, Sanchez-Cabo F. Digitaldlsorter: deep-learning on scRNA-seq to deconvolute gene expression data. Front Genet. 2019;10:77458.
77. Movva R, Greenside P, Marinov GK, Nair S, Shrikumar A, Kundaje A. Deciphering regulatory DNA sequences and noncoding genetic variants using neural network models of massively parallel reporter assays. PLoS One. 2019;71:466689.
78. 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.
79. Bretschneider H, Gandhi S, Deshwar AG, Zuberi K, Frey BJ. COSSMO: predicting competitive alternative splice site selection using deep learning. In: Bioinformatics. 2018.
80. Lo Bosco G, Rizzo R, Fiannaca A, La Rosa M, Urso A. A deep learning model for epigenomic studies. In: 2016 12th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS). IEEE; 2016. p. 688–92.
81. Cazaly E, Saad J, Wang W, Heckman C, Ollikainen M, Tang J. Making sense of the epigenome using data integration approaches. Front Pharmacol. 2019;19:10.
82. Li W, Wong WH, Jiang R. DeepTACT: predicting 3D chromatin contacts via bootstrapping deep learning. Nucleic Acids Res. 2019;47(10):e60–e60.
83. Angermueller C, Lee HJ, Reik W, Stegle O. DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning. Genome Biol. 2017;18(1):67.
84. Yin Q, Wu M, Liu Q, Lv H, Jiang R. DeepHistone: a deep learning approach to predicting histone modifications. BMC Genomics. 2019;20(2):193.
85. Baron V, Adamson ED, Calogero A, Ragona G, Mercola D. The transcription factor Egr1 is a direct regulator of multiple tumor suppressors including TGFβ1, PTEN, p53, and fibronectin. Cancer Gene Ther. 2006;13(2):115–24.
86. Baptista D, Ferreira PG, Rocha M. Deep learning for drug response prediction in cancer. Brief Bioinform. 2021;22(1):360–79.
87. Lesko LJ, Woodcock J. Translation of pharmacogenomics and pharmacogenetics: a regulatory perspective. Nat Rev Drug Discov. 2004;3(9):763–9.
88. Roden DM. Pharmacogenomics: challenges and opportunities. Ann Intern Med. 2006;145(10):749.
89. Pang K, Wan Y-W, Choi WT, Donehower LA, Sun J, Pant D, et al. Combinatorial therapy discovery using mixed integer linear programming. Bioinformatics. 2014;30(10):1456–63.
90. Day D, Siu LL. Approaches to modernize the combination drug development paradigm. Genome Med. 2016;8(1):115.
91. White RE. High-throughput screening in drug metabolism and pharmacokinetic support of drug discovery. Annu Rev Pharmacol Toxicol. 2000;40(1):133–57.

92. Feala JD, Cortes J, Duxbury PM, Piermarocchi C, McCulloch AD, Paternostro G. Systems approaches and algorithms for discovery of combinatorial therapies. Wiley Interdiscip Rev Syst Biol Med. 2010;2(2):181–93.
93. Sun X, Bao J, You Z, Chen X, Cui J. Modeling of signaling crosstalk-mediated drug resistance and its implications on drug combination. Oncotarget. 2016;7(39):63995–4006.
94. Goswami CP, Cheng L, Alexander P, Singal A, Li L. A new drug combinatory effect prediction algorithm on the cancer cell based on gene expression and dose-response curve. CPT Pharmacometrics Syst Pharmacol. 2015;4(2):80–90.
95. Preuer K, Lewis RPI, Hochreiter S, Bender A, Bulusu KC, Klambauer G. DeepSynergy: predicting anti-cancer drug synergy with deep learning. Bioinformatics. 2018;34(9):1538–46.
96. Kalamara A, Tobalina L, Saez-Rodriguez J. How to find the right drug for each patient? advances and challenges in pharmacogenomics. Curr Opin Syst Biol. 2018;10:53–62.
97. Chiu Y-C, Chen H-IH, Zhang T, Zhang S, Gorthi A, Wang L-J, et al. Predicting drug response of tumors from integrated genomic profiles by deep neural networks. BMC Med Genom. 2019;12(51):18.
98. Wang Y, Li F, Bharathwaj M, Rosas NC, Leier A, Akutsu T, et al. DeepBL: a deep learning-based approach for in silico discovery of beta-lactamases. Brief Bioinform. 2020;7:8859.
99. Yu D, Deng L. Deep learning and its applications to signal and information processing exploratory DSP. IEEE Signal Process Mag. 2011;28(1):145–54.
100. Fukushima K, Miyake S. Neocognitron: A Self-Organizing Neural Network Model for a Mechanism of Visual Pattern Recognition. In 1982. p. 267–85.
101. Hinton GE. Reducing the dimensionality of data with neural networks. Science (80-). 2006;313(5786):504–7.
102. Hinton GE, Osindero S, Teh Y-W. A fast learning algorithm for deep belief nets. Neural Comput. 2006;18(7):1527–54.
103. Shi L, Wang Z. Computational strategies for scalable genomics analysis. Genes (Basel). 2019;10(12):1–8.
104. Nelson D, Wang J. Introduction to artificial neural systems. Neurocomputing. 1992;4(6):328–30.
105. Lecun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proc IEEE. 1998;86(11):2278–324.
106. Zell A. Simulation Neuronaler Netze. London: Addison-Wesley; 1994. p. 73.
107. Zeng W, Wu M, Jiang R. Prediction of enhancer-promoter interactions via natural language processing. BMC Genom. 2018;19(S2):84.
108. Indolia S, Goswami AK, Mishra SP, Asopa P. Conceptual understanding of convolutional neural network-a deep learning approach. Procedia Comput Sci. 2018;132:679–88.
109. Gu J, Wang Z, Kuen J, Ma L, Shahroudy A, Shuai B, et al. Recent advances in convolutional. Neural Netw. 2015;5:71143.
110. Rawat W, Wang Z. Deep convolutional neural networks for image classification: a comprehensive review. Neural Comput. 2017;29(9):2352–449.
111. Alipanahi B, Delong A, Weirauch MT, Frey BJ. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat Biotechnol. 2015;33(8):831–8.
112. Zeng W, Wang Y, Jiang R. Integrating distal and proximal information to predict gene expression via a densely connected convolutional neural network. Bioinformatics. 2019;6:7110.
113. Lysenkov V. Introducing deep learning-based methods into the variant calling analysis pipeline. Science. 2019;6:7789.
114. Kelley DR, Reshef YA, Bileschi M, Belanger D, Mclean CY, Snoek J. Sequential regulatory activity prediction across chromosomes with convolutional neural networks. Science. 2018;71:739–50.
115. Pu L, Govindaraj RG, Lemoine JM, Wu H, Brylinski M. DeepDrug3D: classification of ligand-binding pockets in proteins with a convolutional neural network. PLOS Comput Biol. 2019;15(2):e1006718.
116. Gupta G, Saini S. DAVI: deep learning based tool for alignment and single nucleotide variant identification. Science. 2019;2:1–27.
117. Marhon SA, Cameron CJF, Kremer SC. Recurrent Neural Networks. In 2013. p. 29–65.
118. Yu Y, Si X, Hu C, Zhang J. A review of recurrent neural networks: LSTM cells and network architectures. Neural Comput. 2019;31(7):1235–70.
119. Trieu T, Martinez-Fundichely A, Khurana E. DeepMILO: a deep learning approach to predict the impact of non-coding sequence variants on 3D chromatin structure. Genome Biol. 2020;21(1):79.
120. Quang D, Xie X. FactorNet: A deep learning framework for predicting cell type specific transcription factor binding from nucleotide-resolution sequential data. Methods. 2019;166:40–7.
121. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput. 1997;9(8):1735–80.
122. Park S, Min S, Choi H-S, Yoon S. Deep Recurrent Neural Network-Based Identification of Precursor microRNAs. In: Guyon I, Luxburg U V, Bengio S, Wallach H, Fergus R, Vishwanathan S, et al., editors. Advances in Neural Information Processing Systems. Curran Associates, Inc.; 2017.
123. Quang D, Xie X. DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. Nucleic Acids Res. 2016;44(11):e107–e107.
124. Grønning AGB, Doktor TK, Larsen SJ, Petersen USS, Holm LL, Bruun GH, et al. DeepCLIP: predicting the effect of mutations on protein–RNA binding with deep learning. Nucleic Acids Res. 2020;22:7449.
125. Radford A, Metz L, Chintala S. Unsupervised representation learning with deep convolutional generative adversarial networks. Science. 2015;6:7789.
126. Yang B, Liu F, Ren C, Ouyang Z, Xie Z, Bo X, et al. BiRen: predicting enhancers with a deep-learning-based model using the DNA sequence alone. Bioinformatics. 2017;33(13):1930–6.
127. Deng L, Liu Y. Deep Learning in Natural Language Processing. Singapore: Springer; 2018.
128. Schuler GD, Epstein JA, Ohkawa H, Kans JA. [10] Entrez: Molecular biology database and retrieval system. In 1996. p. 141–62.
129. Mikolov T, Chen K, Corrado G, Dean J. Efficient Estimation of Word Representations in Vector Space. 2013;
130. Du J, Jia P, Dai Y, Tao C, Zhao Z, Zhi D. Gene2vec: distributed representation of genes based on co-expression. BMC Genom. 2019;20(1):82.
131. Zhang X-M, Liang L, Liu L, Tang M-J. Graph neural networks and their current applications in bioinformatics. Front Genet. 2021;12:4799.
132. Gori M, Monfardini G, Scarselli F. A new model for learning in graph domains. In: Proceedings 2005 IEEE International Joint Conference on Neural Networks, 2005. IEEE; p. 729–34.
133. Kwon Y, Yoo J, Choi Y-S, Son W-J, Lee D, Kang S. Efficient learning of non-autoregressive graph variational autoencoders for molecular graph generation. J Cheminform. 2019;11(1):70.
134. Barabási A-L, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. Nat Rev Genet. 2011;12(1):56–68.
135. Kopp W, Monti R, Tamburrini A, Ohler U, Akalin A. Deep learning for genomics using Janggu. Nat Commun. 2020;11(1):3488.
136. Avsec Ž, Kreuzhuber R, Israeli J, Xu N, Cheng J, Shrikumar A, et al. The Kipoi repository accelerates community exchange and reuse of predictive models for genomics. Nat Biotechnol. 2019;37(6):592–600.
137. Chen KM, Cofer EM, Zhou J, Troyanskaya OG. Selene: a PyTorch-based deep learning library for sequence data. Nat Methods. 2019;16(4):315–8.
138. Zhou J, Theesfeld CL, Yao K, Chen KM, Wong AK, Troyanskaya OG. Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. Nat Genet. 2018;50(8):1171–9.
139. Budach S, Marsico A. pysster: classification of biological sequences by learning sequence and structure motifs with convolutional neural networks. Bioinformatics. 2018;34(17):3035–7.
140. Neloy AA, Alam S, Bindu RA, Moni NJ. Machine Learning based Health Prediction System using IBM Cloud as PaaS. In: 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI). IEEE; 2019. p. 444–50.
141. Ciaburro G, Ayyadevara VK, Perrier A. Hands-On Machine Learning on Google Cloud Platform: Implementing smart and efficient analytics using Cloud ML Engine. Packt Publishing; 2018. 500 p.
142. Peng L, Peng M, Liao B, Huang G, Li W, Xie D. The advances and challenges of deep learning application in biological big data processing. Curr Bioinform. 2018;13(4):352–9.
143. Carneiro T, Da Medeiros NRV, Nepomuceno T, Bian G-B, De Albuquerque VHC, Filho PPR. Performance analysis of google colaboratory as a tool for accelerating deep learning applications. IEEE Access. 2018;6:61677–85.

144. Bisong E. Google Colaboratory. In: Building Machine Learning and Deep Learning Models on Google Cloud Platform. Berkeley: Apress; 2019. p. 59–64.
145. Luo R, Sedlazeck FJ, Lam TW, Schatz MC. A multi-task convolutional deep neural network for variant calling in single molecule sequencing. Nat Commun. 2019;10(1):1–11.
146. Ravasio V, Ritelli M, Legati A, Giacopuzzi E. GARFIELD-NGS: genomic vARiants fIltering by dEep learning moDels in NGS. Bioinformatics. 2018;34(17):3038–40.
147. Singh A, Bhatia P. Intelli-NGS: Intelligent NGS, a deep neural network-based artificial intelligence to delineate good and bad variant calls from IonTorrent sequencer data. bioRxiv. 2019;2019:879403.
148. Hsieh T-C, Mensah MA, Pantel JT, Aguilar D, Bar O, Bayat A, et al. PEDIA: prioritization of exome data by image analysis. Genet Med. 2019;21(12):2807–14.
149. 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.
150. Park S, Min S, Choi H, Yoon S. deepMiRGene: deep neural network based precursor microRNA prediction. Science. 2016;71:89968.
151. Kelley DR, Snoek J, Rinn JL. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. Genome Res. 2016;26(7):990–9.
152. Quang D, Xie X. DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. Nucleic Acids Res. 2016;44(11):e107–e107.
153. Singh S, Yang Y, Póczos B, Ma J. Predicting enhancer-promoter interaction from genomic sequence with deep neural networks. Quant Biol. 2019;7(2):122–37.
154. Zeng W, Wu M, Jiang R. Prediction of enhancer-promoter interactions via natural language processing. BMC Genom. 2018;19(S2):84.
155. Chen Y, Li Y, Narayan R, Subramanian A, Xie X. Gene expression inference with deep learning. Bioinformatics. 2016;32(12):1832–9.
156. Zeng W, Wang Y, Jiang R. Integrating distal and proximal information to predict gene expression via a densely connected convolutional neural network. Bioinformatics. 2019;2:7889.
157. Kalkatawi M, Magana-Mora A, Jankovic B, Bajic VB. DeepGSR: an optimized deep-learning structure for the recognition of genomic signals and regions. Bioinformatics. 2019;35(7):1125–32.
158. Zuallaert J, Godin F, Kim M, Soete A, Saeys Y, De Neve W. SpliceRover: interpretable convolutional neural networks for improved splice site prediction. Bioinformatics. 2018;34(24):4180–8.
159. Yang B, Liu F, Ren C, Ouyang Z, Xie Z, Bo X, et al. BiRen: predicting enhancers with a deep-learning-based model using the DNA sequence alone. Bioinformatics. 2017;33(13):1930–6.
160. Paggi JM, Bejerano G. A sequence-based, deep learning model accurately predicts RNA splicing branchpoints. RNA. 2018;24(12):1647–58.
161. Almagro AJJ, Sønderby CK, Sønderby SK, Nielsen H, Winther O. DeepLoc: prediction of protein subcellular localization using deep learning. Bioinformatics. 2017;33(21):3387–95.
162. Singh J, Hanson J, Paliwal K, Zhou Y. RNA secondary structure prediction using an ensemble of two-dimensional deep neural networks and transfer learning. Nat Commun. 2019;10(1):5407.
163. Grønning AGB, Doktor TK, Larsen SJ, Petersen USS, Holm LL, Bruun GH, et al. DeepCLIP: predicting the effect of mutations on protein–RNA binding with deep learning. Nucleic Acids Res. 2020;5:9956.
164. Singh R, Lanchantin J, Robins G, Qi Y. DeepChrome: deep-learning for predicting gene expression from histone modifications. Bioinformatics. 2016;32(17):i639–48.
165. Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning–based sequence model. Nat Methods. 2015;12(10):931–4.
166. Lanchantin J, Singh R, Lin Z, Qi Y. Deep Motif: visualizing genomic sequence classifications. Science. 2016;78:1–5.
167. Li W, Wong WH, Jiang R. DeepTACT: predicting 3D chromatin contacts via bootstrapping deep learning. Nucleic Acids Res. 2019;47(10):e60–e60.
168. Xie L, He S, Song X, Bo X, Zhang Z. Deep learning-based transcriptome data classification for drug-target interaction prediction. BMC Genom. 2018;19(S7):667.
169. Kohut K, Limb S, Crawford G. The changing role of the genetic counsellor in the genomics Era. Curr Genet Med Rep. 2019;7(2):75–84.
170. Zeng W, Wu M, Jiang R. Prediction of enhancer-promoter interactions via natural language processing. BMC Genom. 2018;19(S2):84.
171. Frank H. Guenther. Neural Networks: Biological Models and Applications. In: Smel-ser NJ, Baltes PB editors, editor. Oxford: International Encyclopedia of the Social & Behavioral Sciences; 2001. p. 10534–7.
172. Eskiizmililer S. An intelligent Karyotyping architecture based on Artificial Neural Networks and features obtained by automated image analysis. 1993.
173. Catic A, Gurbeta L, Kurtovic-Kozaric A, Mehmedbasic S, Badnjevic A. Application of neural networks for classification of patau, edwards, down, turner and klinefelter syndrome based on first trimester maternal serum screening data, ultrasonographic findings and patient demographics. BMC Med Genom. 2018;11(1):19.
174. Fukushima K. Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biol Cybern. 1980;36(4):193–202.
175. Sakellaropoulos T, Vougas K, Narang S, Koinis F, Kotsinas A, Polyzos A, et al. A deep learning framework for predicting response to therapy in cancer. Cell Rep. 2019;29(11):3367-3373.e4.
176. Kalinin AA, Higgins GA, Reamaroon N, Soroushmehr S, Allyn-Feuer A, Dinov ID, et al. Deep learning in pharmacogenomics: from gene regulation to patient stratification. Pharmacogenomics. 2018;19(7):629–50.
177. Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature. 1986;323(6088):533–6.
178. Shen Z, Bao W, Huang D-S. Recurrent neural network for predicting transcription factor binding sites. Sci Rep. 2018;8(1):15270.
179. Schuster M, Paliwal KK. Bidirectional recurrent neural networks. IEEE Trans Signal Process. 1997;45(11):2673–81.
180. Almagro Armenteros JJ, Sønderby CK, Sønderby SK, Nielsen H, Winther O. DeepLoc: prediction of protein subcellular localization using deep learning. Bioinformatics. 2017;33(21):3387–95.
181. Chung J, Gulcehre C, Cho K, Bengio Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. Science. 2014;7:44598.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

BMC