Alfarama Journal of Basic & Applied Sciences
Faculty of Science Port Said University
https://ajbas.journals.ekb.eg
ajbas@sci.psu.edu.eg
http://sci.psu.edu.eg/en/
ISSN 2682-275X
January 2021, Volume 2, Issue 1
DOI: 10.21608/ajbas.2020.34160.1023
Submitted: 15-07-2020
Accepted: 23-08-2020
Pages: 105-113

Machine and Deep Learning Approaches in Genome: Review Article

Bossy M. Mostafa[1,*], Noha E. El-Attar[2] , Samy A. Abd-Elhafeez[1] , Wael A. Awad[1,3]
[1]Faculty of Science, Portsaid University, Portsaid, Egypt.
[2]Faculty of Computers and AI, Benha University, Benha, Egypt.
[3]Faculty of Computers and Information, Damietta University, Damietta, Egypt.
***Corresponding author**: eng.bossy@hotmail.com

ABSTRACT

Throughout the years Machine Learning (ML) has increased a lot of consideration on ordinary products as search, filters, recognition and recently genomics. Various strategies incorporate sophisticated artificial neural system designs and are all known as applications of Deep Learning (DL). These days, deep learning could be a current and a fortifying field of machine learning. Deep Learning models have fair been shown prepared for both enhancing data encoding simplicity and prescient design execution over elective methodologies. Also deep learning techniques have been shown to reflect and learn unsurprising relationships in various different types of data and to guarantee that the future of genomics research and precise medicine applications will change. DL applications in genomic field is rapidly developed.

This review presents a clarification for machine learning and deep learning methods utilized in genomics. And the main goal is to show a detailed comprehensive overview on the available ML and DL techniques used in genomics.

Keywords
Genomics, DNA, Neural Network, Deep Learning

1. INTRODUCTION

Genomics, is the examination of the whole genes for any living being as study of the structure, function, and inheritance of the genes known as the genome. By employing a high-performance calculations and mathematical methods called bioinformatics, genomics specialists examine huge amount of DNA-sequence data to discover assortments which influence wellbeing, ailment or medication reaction. In people, it requires more than 23,000 genes to look through around 3 billion units of DNA [1].

The amount of available DNA sequences is growing exponentially. However, while raw data is getting progressively accessible, the biological interpretation of this data is happening at a much slower pace. Therefore, there is a significant need to create ML frameworks that can consequently solve numerous problems related with genomic issues, such as gene classification that utilized in diseases diagnosis and

gene prediction which needs to determine the location of protein-encoding genes within a given DNA sequence [2].

2. MACHINE LEARNING IN GENOMICS

ML draws upon a grouping of areas, including computer science, math, engineering and neuroscience. ML researchers create both the mathematical principals and the practical applications of the data-learning systems. ML focuses on the advancement of PC programs that can get data and utilize it to learn for themselves. The ordinary methodology for ML is consists of the following steps as shown in Figure1; collect data, feature selection, training, evaluating, and testing models [3].

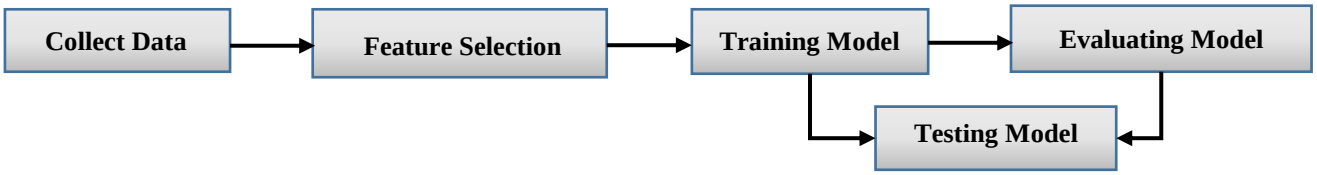

Figure (1): Typical Machine Learning Methodology

Generally, ML has three primary types: supervised, unsupervised, and reinforcement learning. “Supervised learning” is the sort that depends on learning from past models by mapping input-output sets to predict the modern results through labeled training data. It basically concerns with classification and regression problems. A classification is a procedure of categorizing a given set of data into classes, it can be performed on both structured and unstructured data, utilizing different techniques as (Support Vector Machine, Discriminant Analysis, Naïve Bayes, and Nearest Neighbor). Regression models are a predictive goal value based on autonomous variables, using different techniques as (Liner Regression, Decision Trees, Neural Networks, and Ensemble methods). "Unsupervised learning" deduce data patterns without a reliant variable or well-known data, processing huge bulks of data, finding data plans and learning the characteristics that make data focus more or less comparable to one another. It concerns with clustering issues that differentiate between objects on the basis of similarity and disparity between them, utilizing procedures as (K-Means, fuzzy, Hierarchal, Gaussian Mixture, Neural Networks, and Hidden Markov Model). "Reinforcement learning" highlights learning by inquiry and blunder, using incentives or techniques for progress or failure in an assignment as a means of finding fruitful ways to go about it in complex environments. It also known as a semi-supervised learning model in ML [3, 4].

Numerous ML strategies can determine pattern recognition, classification and prediction models from available information and are not depend on rigid suspicions about data-generating systems, making them more viable in a few difficult applications, but less efficient in creating explicit biologically significant models in a few cases [4]. ML methods have been applied successfully in numerous biological and health-related research topics, including the detection and understanding of distinctive gene expressions, binding specificities and splicing effects on cell procedures, interactions of gene-environment, etc. [5].

Recognizing the coding locales in DNA Sequence is a basic step in understanding these sequences. Decision trees is one of a computational strategies for distinguishing between coding and noncoding districts [6]. This methodology cements a few coding measures to supply classifiers with reliably higher accuracies than past techniques, on DNA sequences expanding from 54 base sets to 162 base combines long. Differentially expressed genes (DEGs) were commonly utilized to understand not only the function of genes, but also the subatomic mechanisms underlying different biological types. A few other ML methods such as Logistic Regression, Classification via Regression, Random Forest, Logistic Model Tree (LMT),

and Random Subspace have been used for recognizing the Differentially Expressed Genes (DEGs) that cannot be identified by the conventional Ribonucleic Acid (RNA-sequence) method [7].

RNA-Seq is a dynamic high-throughput procedure that grants analysts to track the genetic make-up of a particular pattern and can aid in determining the regulatory techniques and transcription unit prediction. Research including RNA-Seq data creates gene expression profiles, in which a separate estimate of expression is defined for each remarked gene for that group. Such profiles of gene expression are extracted through pipelines for computational analysis. And McDermaid A. et al. (2018) provided an ML-based method named Gene Expression Quality Control (GeneQC) that can evaluate the reliability of the expression level of each gene deduced from the RNA-Seq dataset. Furthermore it grants researchers with the option of deciding reliable estimates of expression and directing further research on gene expression of adequate quality. It also allows scientists to search ongoing re-alignment strategies to decide more accurate estimates of gene expression for those with low reliability [8].

ML algorithms moreover have expanded prescient capacities for complex disease risk. Host-pathogen protein-protein interactions, which are a physical interactions between two or more molecules resulted from biochemical events, act a crucial part in initiating infections. In this context, Ahmed I. et al. (2018) in [9] have utilized neural network (NN) and compared it with support vector machine (SVM), NN has a significant improvement in overall performance and reaches pleasing precision in human-B.anthracis PPI predictions. In [10] least squares support vector machine (LS-SVM) are utilized by Singh, G., and Samavedham, L. (2015) for clinical diagnosis of neurodegenerative diseases (ND) at person level. Attributes are derived from preprocessed brain MRIs using an unsupervised approach using Kohonen self-organizing map (KSOM). These features are then fed to LSSVM as input for classification purposes [11].The statistical analysis of data was carried out using Excel 2016 according to [18].

3. DEEP LEARNING IN GENOMICS

DL refers to a collection of new techniques that, together, have displayed breakthrough gains over existing ML algorithms for several fields. DL algorithms are a sort of ML algorithms that could make forecasts and perform an efficient dimensional decreasing. The key difference between standard ML and DL strategies utilized in genomics is the high capabilities of DL in dealing with genomics huge data, in addition to their ability in fast adaptation. However, this adaptability is a twofold edged sword. DL can automatically learn attributes and designs with fewer master handcrafting. It needs continuously significant consideration training on and describing the basic biology [11, 12]. DL has many implementation representations like artificial neural network, deep structured learning, and hierarchical learning, which routinely use a group of structured networks to surmise the quantitative characteristics among responses and triggers within a dataset [13]. Figure 2 shows the difference between traditional simple Artificial Neural Network (ANN) and Deep Neural Network (DNN). ANN consists of one or two hidden layers to process data while DNN mainly contains multiple layers between the input and output layers.

Due to the regularly slow learning process associated with a hierarchy of learning data abstractions and representations between layers, some DL algorithms may turn out to be computational-cost when handling high-dimensional image data. Convolutional Neural Networks (CNNs) feasibly scale up knowledge with a large dimensions. CNNs are appropriate for processing a multiple arrays form of information. The general guideline of CNNs architecture design is to lessen the parameters without losing the capacity to learn [12, 13].

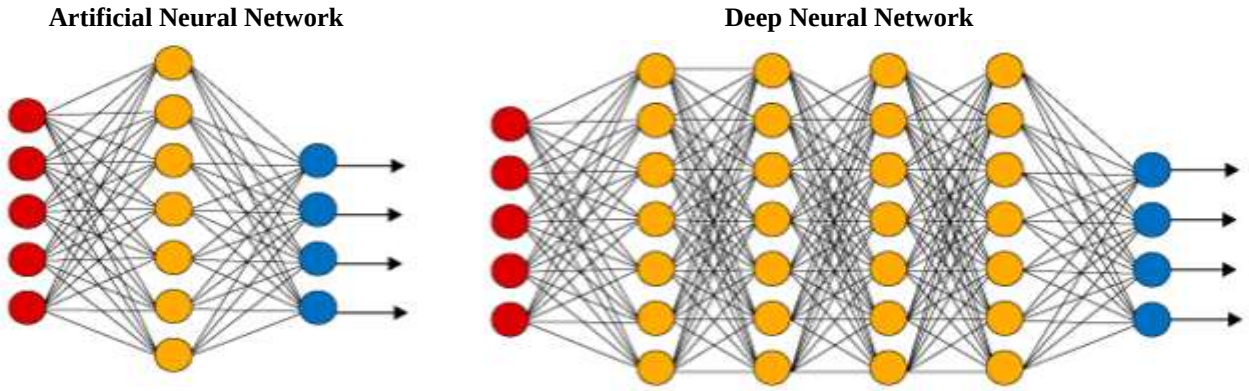

Figure (2): Artificial Neural Network Vs Deep Neural Network [14]

In general, CNN consists of two main layers; convolutional layer and pooling layer which can be repeated until reaching the optimal parameters as shown in Figure 3. The convolution layer can be known as detection layer as it can be used to identify features. Also it contains a set of filters/kernels, in which the input data are convolved with a filter and the outcomes are obtained as a set of feature maps. The pooling layer is to dynamically diminish the spatial size of the representation to reduce the amount of parameters and calculation in the network. It operates on each feature map independently. Most commonly used kind of pooling layer is max pool which compute maximum value of nodes. The result of the previous layers is the feed of the fully-connected neural network that presents its outcome as the final classification decision [15].

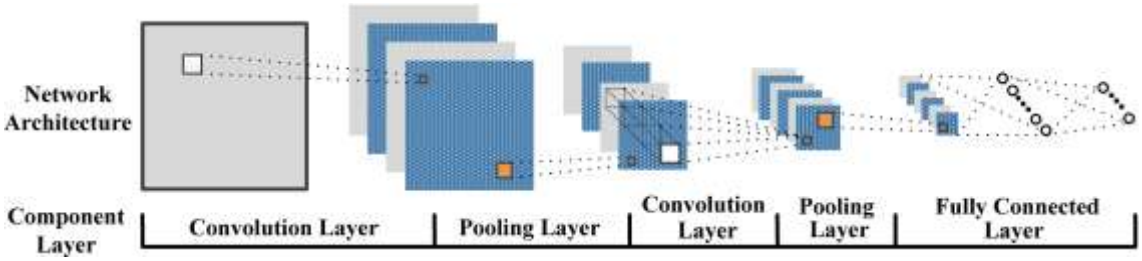

Figure (3): Convolutional Neural Network Architecture [17]

At present, CNNs are one of the best profound learning models inferable from their exceptional ability to analyze spatial data. Because of their advancements in the field of object recognition, CNNs also prove a great potential in bioinformatics and biomedical field. CNN has been broadly utilized for learning DNA sequence datasets identified with regulatory regions and other important features. A great number of researchers have made use of CNNs to handle and process the biological data. But unfortunately, these implementations are usually difficult to reuse to other issues where every biological problem has its own type of characteristics and features [17].

Sequence analysis is the method of a sequence of DNA, RNA or peptide sequence to any of a broad variety of rational techniques to comprehend its landmarks, design, purpose, or evolvement. The used techniques are sequence alignment, searches against biological databases, and others. Sequence alignment is an approach that grew enormously lately in response to the overwhelming burst in data generated by

molecular science activities, it is used to determining the similarity of two sequences. Sequence alignment has two types; Global and Local alignment. In the global alignment, an endeavor is made to join the whole sequences with as many characters as possible, it is usually used in sequence-sequence alignment. While, in the local alignment, highest priority is given to stretches of sequence with highest match density, it is usually used in sequences classification [18]. Regarding the local sequence alignment, Busia A et al., (2018) [19] have proposed a DL approach to forecast the presence/absence of kinds in the metagenomics sample by aligning readings to a rule reference genome data by using a DNN. It can foresee the database-derived labels directly from the query sequence, moreover it can anticipate the types of origin of individual reads more precisely than traditional ML algorithms. They have proved that DNN will be exact and beats read alignment approaches when the query sequences are especially uproarious or vague. Budach S. and Marsico A. (2018) in [20] have implemented a python package (pysster) to make the use of CNN on biological sequence data simpler for researchers. Sequences are categorized by learning sequence and structure forms. The kit provides a computerized method for optimizing the hyper-parameter and options for visualizing learned motifs along with details about their positional and class enrichment.

Another proposed CNN model is implemented to handle the RNA splicing. It is a type of RNA processing in which a newly made antecedent messenger RNA (pre-mRNA) transcript is changed into a develop delivery messenger RNA (mRNA). Introns (Non-coding regions) are separated during splicing, and exons (Coding regions) are combined. Leung M. et al. (2014) in [21] have developed a CNN algorithm for predicting patterns of splicing in isolated tissues and differences in patterns of splicing through tissues. When making predictions, the architecture uses hidden variables that jointly represent features in the genomic sequences and tissue types.

Also, CNN can be applied to obtain a precisely rationed sequence motif from the target sequences. A motif is a sequence pattern that appears in a set of target sequences on more than once. A sequence motif can often involve inserts and deletions between rationed sequences. To accommodate such operations Aoki G. and Sakakibara Y. (2018) in [22], have proposed an application of CNNs for classification of pair sequence alignments for precise sequence clustering and demonstrated the advantages of the CNN input technique for pair alignment for non-coding RNA (ncRNA) sequence clustering and motif discovery.

Another big challenge is the detection of extremely divergent as still unknown viruses. Tampuu A. et al., (2019) in [23] have created ViraMiner contains two parts of sections of Convolutionary Neural Networks designed to classify patterns as well as pattern frequencies on contiguous raw metagenomics. The representation reaches significantly enhanced accuracy in contrast to other methods of ML for classifying the viral genome.

Saving the life of patient is the greatest errand for the specialist particularly when the patient is suffering from chronic disease. Cancer is a one of the heterogeneous disease presenting serious difficulties in early identification and management of the disease. Diverse research efforts are being done since most recent couple of decades to counter the diagnosis challenge [24]. A computer-aided diagnosis system (CAD) has been presented by Sekaran K. et al. (2018) in [25] for mechanized breast cancer diagnosis. For the selection of features this technique used deep neural network (DNN) as the classifier model and recursive feature elimination. A CNN approach has been proposed by Matsubara T. et al. (2019) in [26] which combines spectral clustering data to classify lung cancer. The spectral-CNN based technique developed makes progress in the integration of data from protein interaction networks and profiles of gene expression to classify lung cancer. This proposed technique performs better than that of other methods of ML such as SVM or Random Forest. Table 1 shows the results for the presented learning algorithms used in ML. While Table 2 presented learning algorithms used in DL. Within each category, the algorithms are ordered in ascending order according to the publishing year.

Table (1): Machine Learning Methods

Algorithm	Author	Experimental Study	Accuracy	Disadvantages	Data Size	Year
Decision Tree	Salzberg S., [6]	Distinguishing between coding and noncoding regions in DNA sequence	1-78% for 54 base-pair 2-79% for 108 base-pair 3- 43% for 162 base-pair	1- It can't recognize exons and introns in eukaryotic DNA. 2-The classifier work for a very short DNA sequence	1-290,628 from DNA of 54 base 2- 130,424 from DNA of 108 base 3- 130,424 from DNA of 162 base	2009
Least Squares Support Vector Machine	Singh, G., and Samavedham, L.,[10]	Clinical diagnosis of neurodegenerative disorders on an person basis (NDs)	up to 99%	-	831 T1-weighted MRIs collected from the Parkinson's Progression Markers Initiative (PPMI)	2015
Regression, Random Forest, LMT, Random Subspace	Wang L. et al., [7]	Identify the Differentially expressed genes (DEGs) not identified by traditional RNA-sequence method.	78.6%	Studies will be carried out to increase the efficiency of ML based approaches by the use of more epigenomics data and advanced models in DL	468 features	2018
Machine Learning based tool (GeneQC)	McDermaid A. et al.,[8]	Present a ML-based method called (GeneQC) Gene expression Quality Control, which can reliably evaluate the reliability of the level of expression of each gene derived from an RNA-Seq dataset	50% for RNA-Seq 12.5 for other species	Mapping uncertainty, in modern RNA-Seq analyzes, considers a serious problem. High mapping uncertainty can result in highly biased estimates of expression over fewer genes, while moderate rates of mapping uncertainty on a wider scale as seen in plant species can cause biases on a lesser scale to be estimated at widespread expression.	95 RNA-Seq datasets,	2018
Neural Network	Ahmed I. et al.,[9]	Host-pathogen protein-protein interactions	80.5%	-	554 human-B. anthracis	2018

Table (2): Deep Learning Methods

Algorithm	Author	Experimental Study	Accuracy	Disadvantages	Data Size	Year
CNN	Leung M. et al.,[21]	Predict patterns of splicing in individual tissues, and contrast patterns of splicing through tissues.	85.6%	A decrease in brain-to-heart output of 7% is observed due to the heterogeneity of brain tissues	11019 mouse alternative exons profiled from RNA-Seq data	2014
CNN	Busia A et al., [19]	Predicting the presence / absence of species in the metagenomics sample by aligning readings to a rule Reference genome data problem	81.1%	For commonly used sequencing technologies such as Illumina, this may not be practical.	19,851 16S ribosomal RNA reference sequences	2018
CNN	Aoki G. and Sakakibara Y.,[22]	Classification of pairwise alignments of sequences for accurate clustering of sequences	85%	-	5983 ncRNA genes	2018
CNN	Sekaran K. et al.,[25]	A computer-aided diagnosis system (CAD) to perform mechanized diagnosis for breast cancer	98.62%	The algorithm's long training time since it has been training the neural network in depth.	699 instances with 9 feature variables	2018
CNN	Tampuu A. et al.,[23]	The discovery of highly divergent or yet unknown	92%	-	dataset included 19 different NGS	2019

4. CONCLUSION & FUTURE WORK

DL is becoming the overwhelming focus for international academic and business interests as we enter the significant era of big data. In bioinformatics, where remarkable advances with traditional ML have been made, profound learning is anticipated to yield promising results. In this review we provided an extensive review of bioinformatics research applying ML and DL.

Although DL holds a guarantee it cannot provide incredible outcomes in specially appointed bioinformatics applications. Numerous potential challenges remain, including restricted or imbalanced information, interpretation of the results of DL, and determination of a proper design and hyper-parameters.

Additionally, further analysis is required to fully leverage the capabilities of DL, multimodality, and increasing speed of DL. We accept that this survey will provide important

understanding and fill in as a starting stage for a greater use of DL in future research to advance bioinformatics.

REFERENCES

[1] GRIFFITHS A., U.S "Genomics", available on line at https://www.britannica.com/science/genomics, last access on 1/4/2020 at 9 AM.

[2] NATIONAL HUMAN GENOME RESEARCH INSTITUTE, U.S "A Brief Guide to Genomics", available on line at https://www.genome.gov/about-genomics/fact-sheets/A-Brief-Guide-to-Genomics , last access on 1/4/2020 at 9 AM.

[3] NATIONAL ACADEMY OF SCIENCES; The Royal Society, "The Frontiers of Machine Learning", First Edition, copyright by National Academy of Sciences, 2017.

[4] XU CH. AND JACKSON S. "Machine Learning and Complex Biological Data", Genome Biology, 20 (76), 16, April 2019.

[5] MAZANDU G. ET AL., "Artificial Intelligence Applications in Medicine and Biology", First Edition, copyright by intechopen, 2019.

[6] SALZBERG S., "Locating Protein Coding Regions in Human DNA using a Decision Tree Algorithm", Journal of Computational Biology, 2 (3), 2009.

[7] WANG L. ET AL., "RNA-seq Assistant: Machine Learning Based Methods to Identify More Transcriptional Regulated Genes", BMC Genomics; 19 (546), 2018.

[8] MCDERMAID A. ET AL., "A New Machine Learning-Based Framework for Mapping Uncertainty Analysis in RNA-Seq Read Alignment and Gene Expression Estimation", Frontiers in Genetics, 9 (313), 14, August, 2018.

[9] AHMED I. ET AL., "Prediction of Human-Bacillus Anthracis Protein-Protein Interactions using Multi-Layer Neural Network", Bioinformatics, 34 (24), P. 4159-4164, 2018.

[10] SINGH, G., AND SAMAVEDHAM, L., "Unsupervised Learning Based Feature Extraction for Diff erential Diagnosis of Neurodegenerative Diseases: A Case Study on Early-Stage Diagnosis of Parkinson Disease", Journal of Neurosci Methods, 256, P.30–40, 2015.

[11] GRAPOV D. ET AL., "Rise of Deep Learning for Genomic, Proteomic, and Metabolomic Data Integration in Precision Medicine", OMICS A Journal of Integrative Biology, 22 (10), 2018.

[12] ZOU J. ET AL, "A Primer on Deep Learning in Genomics", Nature Genetics, 51 (1), November, 2018.

[13] TANG B. ET AL., "Recent Advances of Deep Learning in Bioinformatics and Computational Biology", Frontiers in Genetics, 10 (214), 26, March, 2019.

[14] GLOBAL ENGAGE, available on line at http://www.global-engage.com/life-science/deep-learning-in-digital-pathology/ , last access on 1/4/2020 at 9 AM.

[15] AMIN M. ET AL., "Comparative Study of Machine Learning Techniques for Population Genetics", International Journal of Computer Science and Network Security, 19 (6), June, 2019.

[16] LIAO S. ET AL., "Reduced-Complexity Deep Neural Networks Design Using Multi-Level Compression", IEEE Transactions on Sustainable Computing, 4, P. 245-251, February, 2019.

[17] MIN S. ET AL., "Deep Learning in Bioinformatics", Briefings in Bioinformatics, 1, July, 2016.

[18] EL-BONDKLY A., "Biotechnology and Biology of Trichoderma", First Edition, copyright by Elsevier, 2014.

[19] BUSIA A ET AL., "A deep learning approach to pattern recognition for short DNA sequences", BioRxiv, June, 2018.

[20] BUDACH S. AND MARSICO A., "pysster: classification of biological sequences by learning sequence and structure motifs with convolutional neural networks", Bioinformatics, 34 (17), October, 2018.
[21] LEUNG M. ET AL., "Deep learning of the tissue-regulated splicing code", Bioinformatics, 30, P 121–129, 2014.
[22] AOKI G. AND SAKAKIBARA Y., "Convolutional neural networks for classification of alignments of non-coding RNA sequences", Bioinformatics, 34 (13), June, 2018.
[23] TAMPUU A. ET AL., "ViraMiner: Deep learning on raw DNA sequences for identifying viral genomes in human samples", Plos One, 14 (9), September, 2019.
[24] HUSSAIN F. ET AL., "Classifying cancer patients based on DNA sequences using machine learning", Journal of Medical Imaging and Health Informatics, 9 (3), March 2019.
[25] SEKARAN K. ET AL., "Knowledge Computing and Its Applications", First Edition, copyright by Springer Nature Singapore, 2018.
[26] MATSUBARA T. ET AL., "Convolutional neural network approach to lung cancer classification integrating protein interaction network and gene expression profiles", Journal of Bioinformatics and Computational Biology,17 (3), June, 2019.