IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

ASSESSMENT OF MACHINE LEARNING APPLICATIONS IN BIOINFORMATICS

Shaheen Hayat
Research Scholar
Centre for Interdisciplinary Research in Basic Sciences
Jamia Millia Islamia
New Delhi-110025

Abstract

Biological processes are inherently complex, requiring an understanding not only of individual components but also of their intricate interconnections. Graphs, or networks, provide a natural framework for representing such relationships, enabling the depiction of both biological entities and their interactions. Recent advancements in high-throughput experimental technologies have led to an exponential increase in the generation of biological network data, which is now readily accessible to the scientific community through online databases facilitated by internet web services. Consequently, there is a growing demand for novel techniques to query, analyze, and process this wealth of data, with the aim of extracting insights into molecular biology, physiology, electronic health records, biological networks, and biomedicine as a whole. Machine learning, owing to its ability to effectively handle large datasets and generate accurate predictions using sophisticated statistical models, has emerged as a powerful tool for meeting these demands, promising rapid growth and widespread utilization in biomedical research.

Keywords: Machine Learning; Biological Network; Bioinformatics Approach

1. Introduction

Many biological processes necessitate knowledge of not only the biological components themselves, but also their interrelationships. A graph, often known as a network, is a natural approach to represent such processes. A graph can be used to represent both components and their interactions. Latest advancements in high-throughput experimental technology have greatly boosted the data output from biological entity relationships, resulting in a massive amount of biological network data and are now available to scientific community. Internet web services grew simultaneously with the rise of these databases, allowing biologists to publish vast amounts of data online for scholarly audiences. As a

result, scientists are looking for new techniques to query, analyse, and process data in order to derive knowledge about molecular biology, physiology, electronic health records, biological networks and biomedicine in general. Machine learning has ability to grow quickly and be utilised widely due to its unique capacity to manage big datasets and generate predictions on them using accurate statistical models [1, 2].

2. Machine Learning

A famous computer scientist Arthur Samuel originated the term Machine Learning in 1959. Machine learning (ML) is the subset of Artificial Intelligence, but both the words Artificial Intelligence and Machine Learning are often used alternately, though incorrectly; "Artificial Intelligence refers to the comprehensive concept of the 'thinking machine' or automated decision-making whereas he described ML as giving computers the ability to learn without being explicitly programmed" [3, 4, 5]. In its fundamental state, machine learning utilizes predetermined algorithms to enhance their operations through the analysis of incoming data and resulting sets of predictions within an acceptable range. Following are just a few of the applications for these algorithms like junk e-mail filtering [6], customer purchase behaviour detection [7], network intrusion detection [8], credit card fraud detection [9], disease modelling [10], optimising manufacturing process [11], and automated text categorisation [12]. These algorithms have a tendency to generate more precise predictions as more data is inputted into them. There are several approaches to categorising machine learning algorithms based on their intended use and the method of instruction. These approaches can be grouped into three main categories. The three categories are as follows: supervised learning, unsupervised learning, and deep learning [13].

There are several types of Machine learning algorithms in use, which are outside the scope of this chapter. However, some of them are addressed here because of their significance in bioinformatics, biomedical and computational biology research.

2.1 Supervised Learning

In supervised machine learning, main aim is to learn a target function which can be used to predict the values of a class and convert an input to an output on the basis of input output pairs. In this learning process, first approach is to distribute the dataset. Function is deduced from labelled training data comprising a set of training examples. The input dataset can be divided into training and testing dataset. The training dataset contains an output variable that needs to be predicted or classified. In the context of classification or prediction tasks, algorithms acquire patterns from the training dataset and subsequently utilise them to make predictions on the test dataset [14]. The most common example used in supervised learning is training a model to discriminate among different kind of fruits like apples, oranges and lemons. Each fruit labelling is first supplied to the algorithm with some features such as colour, weight, shape and size and the characterization of fruits is done by learning the features mixture by algorithm. On the basis of this, new unlabelled fruit can be predicted by the model [5, 15]. It can be understood by taking a Figure 1 for three diabetic patients by using abstract

dataset, which explain the categorization of diabetic and non-diabetic patients. The problem of classification and regression suits well with supervised machine learning algorithm. In the context of classification problems, the fundamental output variable is characterised by its discrete nature. A discrete variable is characterised by its ability to be categorised into distinct groups or classes, such as "red" or "black," or "diabetic" and "nondiabetic." However, in the context of regression issues, the dependent variable of interest is a continuous real number, such as the quantification of an individual's risk for developing cardiovascular disease [13].

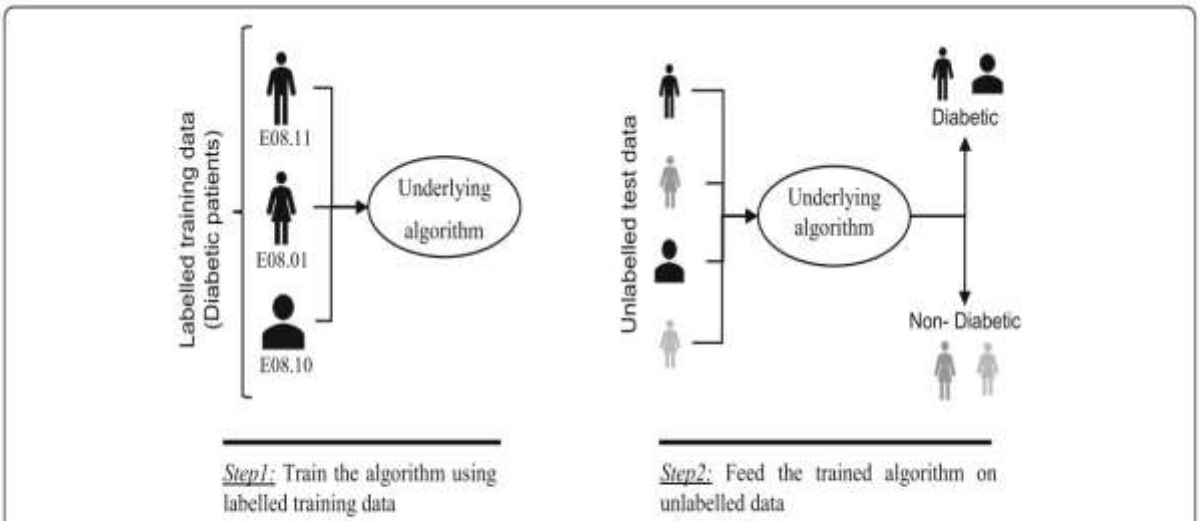

Fig. 1 An illustration of how supervised machine learning algorithms work to categorise diabetic and non-diabetic patients based on abstract data

2.1.1 Support Vector Machine

SVM algorithm exhibits commendable efficacy in the classification of data sets, encompassing both linear and non-linear patterns by building a classifier. By making a decision boundary known as “hyperplane” it separates or categorizes data into different groups on the basis of discernible patterns of information pertaining to said observations or data, commonly known as features. and by using this hyperplane, the most probable label for unseen data can be determined. The coordinates reference the features on the basis of their relationship to each other and makes the support vectors [16]. Even with minimal examples, SVM works well and has good accuracy [17].

2.1.2 Decision Tree

Decision Tree is one of the most popular methods of classification which is widely used in prediction-based problems. A decision tree algorithm models the test data and classifies very large data by forming a tree like structure [18]. The representation of DT takes the form of a hierarchical structure, resembling a tree, composed of multiple interconnected nodes. The foremost and uppermost node within this structure is referred to as the root node. The manifestation of tests on input variables or attributes is exhibited by every internal node. The classification algorithm, in accordance with the test result, directs its path towards the appropriate child node, thereby initiating a cyclical process of testing and branching. This iterative procedure continues until the algorithm ultimately arrives at the

terminal leaf node [19]. The terminal nodes (leaf) represent the decision outcomes. As decision trees are very easy and quick to understand, they are widely used in various biological and medical diagnostic techniques [20]. The comprehensive collection of test results obtained at each node throughout the traversal of the tree provides ample data to formulate rational hypotheses regarding the classification of a given sample [13].

2.1.3 Random Forest

The esteemed Dr. Breiman introduced the random forest algorithm in the year 2001, which has since garnered considerable acclaim for its exceptional performance as both a classification and regression technique. The utilisation of an algorithm that amalgamates multiple randomised decision trees and subsequently aggregates their predictions through the process of averaging has demonstrated commendable performance in scenarios characterised by a substantial disparity between the quantity of variables and observations. This approach operates on the foundation of statistical learning theory, employing Bootstrap randomised resampling to derive diverse iterations of the sample sets from the initial training datasets. Subsequently, it constructs a distinct decision tree model for each of these sample sets. For the prediction of classification result, it uses established voting mechanism and the final model aggregates all the result of decision tree [21].

2.1.4 Naive Bayes

Naïve bayes is a simple classification technique of supervised machine learning that uses mathematical Bayes theorem for getting the probability. It classifies a given dataset by calculating a probability. In a given dataset each of the attributes are independent of each other and they independently maximize the probability. The maximized probability is the output of a given example [22, 23].

2.1.5 K-Nearest Neighbour (KNN)

KNN is one of the earliest and simple classification algorithm based on statistical approach. In KNN, K stands for the number of nearest neighbour used, which is calculated using the stated value's upper limit [24]. This method utilises the majority voting approach from its nearest neighbours to make predictions about the class of a new instance. The Euclidean distance is employed to compute the proximity of a given attribute to its neighbouring attributes [25]. It's simple to execute and master, but it has the problem of being noticeably slower as the amount of data in use rises [26].

2.2 Unsupervised Learning

In unsupervised learning technique, targets or responses are not given in the dataset. This technique simply attempts to get the similarities between the input values and classify them on the basis of these similarities [27, 28]. It's mostly applicable for feature reduction and clustering [26].

2.2.1 K-Means Clustering

K-means clustering algorithm is one of the easiest classical unsupervised learning techniques. This technique makes a ‘k’ cluster by dividing the ‘n’ data points on account of similarity measure, with data points in the same cluster having a high similarity to data points in other clusters. It picks the k-centroid at random, and then allocates the data points to such 'k' centroids using a similarity measure. On the basis of the cluster mean’s similarity i.e. the distance between the data points, a data point is assigned to a cluster to each iteration. After this, the current mean calculation is done and this process is repeated for each new data point. The aim of the method is to make dense clusters of related data points with very little similarity with other clusters. The cluster mean, which is also known as the cluster centroid, can be used to describe cluster similarity [29].

2.2.2 Principal Component Analysis

The method of principle component analysis entails the utilisation of a statistical technique to convert a set of observations, which may exhibit correlations among variables, into a set of values that are linearly uncorrelated. This is achieved through the application of an orthogonal transformation, resulting in what are referred to as principal components. The process of reducing the dimensionality of the data in this particular technique confers a notable advantage in terms of expediting and simplifying the computational procedures involved. Utilising linear combinations, one can effectively elucidate the variance-covariance structure of a set of variables. It's a popular approach for reducing dimensionality [26].

2.3 Deep Learning

Deep learning is the newly evolved branch of machine learning techniques that comes first in 2000s and because of its novel prediction performance on big data, it quickly gained attraction in various fields [30, 31, 32]. Classical Artificial Neural Network is the fundamental concept behind the deep learning, which copies the activity and working design of human brain to make algorithms more intelligent and efficient and also reduces the human labour [33, 34, 35]. Deep learning involves the use of numerous hidden neurons and layers which benefits its architecture paired with new training paradigm, as opposed to the classical neural networks. By using a large number of neurons provided for a broad coverage of the raw data, the layer-by-layer workflow of nonlinear combination of their results produce lower dimensional projection of the input space. A higher perceptual level is related to the entire lower-dimensional projection. It results in an effective high-level abstraction of the raw data or images, if the network is ideally weighted. This high degree of abstraction provides automated feature set that would otherwise need hand-crafted or specialized features [36]. Such features may be used in translational bioinformatics to find nucleotide sequences that potentially bind a DNA or RNA strand to a protein [37]. Despite the fact that deep learning is a newly appeared domain of machine learning, it has a wide range of applications in machine vision, speech and signal processing, sequence and text prediction, and computational biology, all of which are defining the current Artificial Intelligence disciplines [37, 38, 39, 40, 41, 42, 43].

Deep learning can be categorised into four primary classifications. The aforementioned designs encompass Deep Neural Networks (DNNs), Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and emergent models [44] and depending on the kinds of layers and their associated learning methods, there are several deep learning networks of DNNs, out of which common examples are deep belief network (DBN), stacked auto-encoder (SAE), Convolutional neural network (CNN) recurrent neural network (RNN). Those models, which have a certain representational property of model structure and training algorithm, are the most extensively applied in biomedical analysis [45].

3. Machine Learning Algorithms in Bioinformatics and Disease Dynamics

Bioinformatics is an interdisciplinary branch of science that deals with the development of computational methods and applied to convert these huge numbers of biological data that are produced by the several biological systems like genomics, systems biology, proteomics, deep sequencing into knowledge and used them in biomedical applications [46]. To understand the function of genes, cell regulation and signalling, drug designing, drug target, diagnosis and prediction of disease, bioinformatics techniques are able to solve these issues [47].

Machine learning methods are widely employed in the field of bioinformatics for various tasks such as prediction, classification, and feature selection. Machine learning (ML) methodologies have demonstrated considerable efficacy in addressing challenges related to differentiating DNA sequences and classifying them. The field of bioinformatics has witnessed a notable rise in the importance of machine learning (ML), particularly with the emergence of deep learning techniques [48].

This section describes the representative works of machine learning algorithms in different areas of bioinformatics and disease dynamics applications which is summarized below:

- Quan Zou et al. [49] conducted a study in which the prediction of Diabetes mellitus was performed using a combination of random forest, decision tree, and neural network algorithms. The findings of this study demonstrated that the random forest algorithm exhibited the highest level of accuracy in predicting the occurrence of this disease.

- Cai Huang et al. [50] have developed an open-source software platform that used SVM in conjunction with typical recursive feature elimination (RFE) method. This innovative approach enables the accurate prediction of personalized drug responses based on gene expression profiles. The models have demonstrated remarkable efficacy in prognosticating the drug responsiveness of diverse cancer cell lines.

- Samaneh Kouchaki et al. [51] have found that logistic regression and gradient tree boosting perform better in predicting mycobacterium tuberculosis (MTB) drug resistance

- Fiannaca et al. [52] identified the 16S short-read sequences by using *k*-mer and deep learning techniques. The result shows that the method is well enough to classify both 16S shotgun (SG) and amplicon (AMP) data very well.

- Amgarten et al. [53] have used a Random Forest method to develop a tool named MARVEL for predicting double-stranded DNA bacteriophage sequences in metagenomics.

- Budach et al. [54] adeptly categorised biological sequences through the acquisition of sequence and structure motifs by applying CNN.

- Tsubaki et al. [55] have used GNN and CNN to predict the compound-protein interaction (CPI). It is helpful in effective drug discovery.

- Kumar G Dinesh et al. [56] predicted cardiovascular disease by using Support Vector Machine, Gradient Boosting, Random Forest, Naive Bayes classifier and logistic regression techniques.

- Liu et al. [57] introduced an innovative approach utilising a dual radial basis function (RBF) kernel technique in the realm of cancer classification. The primary objective of this method is to discern pertinent features from gene expression data. The superfluous and incongruous genes are eliminated through the amalgamation of RBF kernels employing a weighted analysis technique, thereby facilitating the extraction of pertinent feature genes.

4. Integrating Machine Learning in Biological Networks

The study of the complex interactions of biomolecules that involves the structure and function of living cells is known as Network biology. It deals with the modelling of biological systems, involves very complex datasets produced by the infinite number of multi omics programs. Systems biology as a branch of network biology rebuild and interpret extensive endogenous biological networks, and the branch synthetic biology draft and build minor synthetic gene networks. It has been reported that the machine learning approaches has aided in the identification of network design in field of network biology [58].

As different strata of biological systems produce extensive and various types of data, it can be beneficial to apply machine learning techniques due to these large datasets for constructing intricate and biologically plausible network models covering several strata, from the regulation of gene to interspecific relations. In diverse biomedical applications machine learning approaches develop different tools that can help in increasing the application of these network models. Network biology is able to provide us with superior knowledge about the complexity of disease biology. By considering “disease biology” as an instance, here we can understand the confluence of machine learning and network biology as well as ongoing issues and scope of machine learning in network biology. Despite of pointing out and specifying particular aspects of disease, like finding of disease-causing genes,

network biology makes use of an integrated method, thus as a result, it gives us a complete and absolute picture of the factors that guide disease phenotypes and thereby identify networks and subnetworks of vital biomolecular interactions crucial for the appearance of the disease. Applications of machine learning algorithm can help to understand the network related disease mechanism. To understand in detail by an example application we can use a database Bio-GRID for biomolecular interactions and find out how interactions various biomolecules change in healthy to disease state. Beginning with a data from healthy group, we can prepare a learning model by training a deep learning algorithm, like deep neural network to learn the basic features that determine healthy state. And after training the model, the data from a diseased group can be given to the algorithm that is applied to know the difference between the disease and healthy states, finding distinct groups of regulatory interactions and biomolecules that may be confirmed and further investigated. In the area of network inference, similar techniques have been used to identify topological characteristics that may be attributed to variations in phenotypic information at the expression level [58, 59, 60]. It has been reported that, to get a superior understanding of disease and their associated dysregulation of network, and intricate hierarchical structure of biological networks, we can use a "Capsule network" (a next generation ML method) that may be quite valuable. "Capsule networks involve a new type of neural network architecture, where CNNs are encapsulated in interconnected modules [61, 62]." Capsule networks are suitable for disease and network biology, because biological networks are highly modular in nature and numerous biomolecules have their defined layers, although this method enabling each of these layers to interact with one another. In capsule network technique, capsule represents the biological layers including data produced from all of the layers such as, proteomics, metabolomics, and transcriptomics. Some other ensemble deep learning methods such as,
DNN, CNN, and CNN+RNN have been successfully applied in biological network analysis [63].

References

1. Reuter, J. A., Spacek, D. V., & Snyder, M. P. (2015). High-throughput sequencing technologies. Molecular cell, 58(4), 586-597.

2. Chicco, D. (2017). Ten quick tips for machine learning in computational biology. BioData mining, 10(1), 35.

3. Munoz, A. (2014). Machine learning and optimization. Courant Institute of Mathematical Sciences, 1-2.

4. Mitchell, T. M. (1997). "Machine Learning", New York, NY, USA: McGraw-Hill, Inc.

5. Handelman, G. S., et al. (2018) "Doctor: machine learning and the future of medicine." Journal of internal medicine 284.6: 603-619.

6. Sahami, M., Dumais, S., Heckerman, D., & Horvitz, E. (1998). A Bayesian approach to filtering junk e-mail. In Learning for Text Categorization: Papers from the 1998 workshop Vol. 62, pp. 98-105.

7. Kim, E., Kim, W., & Lee, Y. (2003). Combination of multiple classifiers for the customer's purchase behavior prediction. Decision Support Systems, 34(2), 167-175.

8. Sinclair, C., Pierce, L., & Matzner, S. (1999, December). An application of machine learning to network intrusion detection. In Proceedings 15th annual computer security applications conference pp. 371-377. IEEE.

9. Aleskerov, E., Freisleben, B., & Rao, B. (1997). Cardwatch: A neural network based database mining system for credit card fraud detection. In Proceedings of the computational intelligence for financial engineering pp. 220-226. IEEE.

10. Yao, D., Yang, J., & Zhan, X. (2013). A Novel Method for Disease Prediction: Hybrid of Random Forest and Multivariate Adaptive Regression Splines. J. Comput., 8(1), 170-177.

11. Mahadevan, S., & Theocharous, G. (1998). Optimizing production manufacturing using reinforcement learning. In FLAIRS conference Vol. 372, p. 377.

12. Sebastiani, F. (2002). Machine learning in automated text categorization. ACM computing surveys (CSUR), 34(1), 1-47.

13. Uddin, S., Khan, A., Hossain, M. E., & Moni, M. A. (2019). Comparing different supervised machine learning algorithms for disease prediction. BMC medical informatics and decision making, 19(1), 1-16.

14. Mahesh, B. (2020). Machine learning algorithms-a review. International Journal of Science and Research (IJSR), 9(1), 381-386.

15. El Naqa, I., & Murphy, M. J. (2015). What is machine learning? (pp. 3-11). Springer International Publishing.

16. Pisner, D. A., & Schnyer, D. M. (2020). Support vector machine. In Machine learning (pp. 101-121). Academic Press.

17. Almansour, N. A., Syed, H. F., Khayat, N. R., Altheeb, R. K., Juri, R. E., Alhiyafi, J. & Olatunji, S. O. (2019). Neural network and support vector machine for the prediction of chronic kidney disease: A comparative study. Computers in biology and medicine, 109, 101111.

18. Damanik, I. S., Windarto, A. P., Wanto, A., Poningsih, Andani, S. R., & Saputra, W. (2019). Decision tree optimization in C4. 5 algorithms using genetic algorithm. In Journal of Physics: Conference Series Vol. 1255, No. 1, p. 012012. IOP Publishing.

19. Quinlan, J. R. (1986). Induction of decision trees. Machine learning, 1, 81-106.

20. Cruz, J. A., & Wishart, D. S. (2006). Applications of machine learning in cancer prediction and prognosis. Cancer informatics, 2, 117693510600200030.

21. Xu, W., Zhang, J., Zhang, Q., & Wei, X. (2017). Risk prediction of type II diabetes based on random forest model. Third International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics pp. 382-386. IEEE.

22. Bashir, S., Khan, Z. S., Khan, F. H., Anjum, A., & Bashir, K. (2019). Improving heart disease prediction using feature selection approaches. 16th international bhurban conference on applied sciences and technology pp. 619-623. IEEE.

23. Shah, D., Patel, S., & Bharti, S. K. (2020). Heart disease prediction using machine learning techniques. SN Computer Science, 1(6), 1-6.

24. Ali, M. M., Paul, B. K., Ahmed, K., Bui, F. M., Quinn, J. M., & Moni, M. A. (2021). Heart disease prediction using supervised machine learning algorithms: performance analysis and comparison. Computers in Biology and Medicine, 136, 104672.

25. Pouriyeh, S., Vahid, S., Sannino, G., De Pietro, G., Arabnia, H., & Gutierrez, J. (2017). A comprehensive investigation and comparison of machine learning techniques in the domain of heart disease symposium on computers and communications pp. 204-207. IEEE.

26. Mahesh, B. (2020). Machine Learning Algorithms-A Review. International Journal of Science and Research 9, 381-386.

27. Fatima, M., & Pasha, M. (2017). Survey of machine learning algorithms for disease diagnostic. Journal of Intelligent Learning Systems and Applications, 9(01), 1.

28. Aljanabi, M., Qutqut, H. M., & Hijjawi, M. (2018). Machine learning classification techniques for heart disease prediction: A review. International Journal of Engineering & Technology, 7(4), 5373-5379.

29. Tomar, D., & Agarwal, S. (2013). A survey on Data Mining approaches for Healthcare. International Journal of Bio-Science and Bio-Technology, 5(5), 241-266.

30. Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. Science, 313(5786), 504-507.

31. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436-444.

32. Nussinov, R., Tsai, C. J., Shehu, A., & Jang, H. (2019). Computational structural biology: Successes, future directions, and challenges. Molecules, 24(3), 637.

33. Mamoshina P, Vieira A, Putin E, Zhavoronkov A (2016). Applications of Deep Learning in Biomedicine. Molecular Pharmaceutics, 13:1445-1454.

34. Schmidhuber, J. (2015). Deep learning in neural networks: An overview. Neural networks, 61, 85-117.

35. Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., & Hassabis, D. (2015). Human-level control through deep reinforcement learning. Nature, 518(7540), 529-533.

36. Ravì, D., Wong, C., Deligianni, F., Berthelot, M., Andreu-Perez, J., Lo, B., & Yang, G. Z. (2016). Deep learning for health informatics. IEEE journal of biomedical and health informatics, 21(1), 4-21.

37. Alipanahi, B., Delong, A., Weirauch, M. T., & Frey, B. J. (2015). Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. Nature biotechnology, 33(8), 831-838.

38. Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), 115-118.

39. Zhang, S., Zhou, J., Hu, H., Gong, H., Chen, L., Cheng, C., & Zeng, J. (2016). A deep learning framework for modelling structural features of RNA-binding protein targets. Nucleic acids research, 44(4), e32-e32.

40. Bengio, Y., & LeCun, Y. (2007). Scaling learning algorithms towards AI. Large-scale kernel machines, 34(5), 1-41.

41. Libbrecht, M. W., & Noble, W. S. (2015). Machine learning applications in genetics and genomics. Nature Reviews Genetics, 16(6), 321-332.

42. Ching, T., Himmelstein, D. S., Beaulieu-Jones, B. K., Kalinin, A. A., Do, B. T., Way, G. P., & Greene, C. S. (2018). Opportunities and obstacles for deep learning in biology and medicine. Journal of the Royal Society Interface, 15(141), 20170387.

43. Tang, B., Pan, Z., Yin, K., & Khateeb, A. (2019). Recent advances of deep learning in bioinformatics and computational biology. Frontiers in genetics, 10, 214.

44. Min, S., Lee, B., & Yoon, S. (2017). Deep learning in bioinformatics. Briefings in bioinformatics, 18(5), 851-869.

45. Lan, K., Wang, D. T., Fong, S., Liu, L. S., Wong, K. K., & Dey, N. (2018). A survey of data mining and deep learning in bioinformatics. Journal of medical systems, 42(8), 1-20.

46. Yin, Z., Lan, H., Tan, G., Lu, M., Vasilakos, A. V., & Liu, W. (2017). Computing platforms for big biological data analytics: perspectives and challenges. Computational and structural biotechnology journal, 15, 403-411.

47. Babar, M. M., Zaidi, N. U. S. S., Pothineni, V. R., Ali, Z., Faisal, S., Hakeem, K. R., & Gul, A. (2017). Application of bioinformatics and system biology in medicinal plant studies. In Plant Bioinformatics, pp. 375-393. Springer, Cham.

48. Awad, M., & Khanna, R. (2015). Support vector regression. In Efficient learning machines pp. 67-80. Apress, Berkeley, CA.

49. Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., & Tang, H. (2018). Predicting diabetes mellitus with machine learning techniques. Frontiers in genetics, 9, 515.

50. Huang, C., Mezencev, R., McDonald, J. F., & Vannberg, F. (2017). Open source machine-learning algorithms for the prediction of optimal cancer drug therapies. PLoS One, 12(10), e0186906.

51. Kouchaki, S., Yang, Y., Walker, T. M., Sarah Walker, A., Wilson, D. J., Peto, T. E. & Clifton, D. A. (2019). Application of machine learning techniques to tuberculosis drug resistance analysis. Bioinformatics, 35(13), 2276-2282.

52. Hu, H., Zhang, L., Ai, H., Zhang, H., Fan, Y., Zhao, Q., & Liu, H. (2018). HL PI-ensemble: prediction of human lncRNA-protein interactions based on ensemble strategy. RNA biology, 15(6), 797-806.

53. Amgarten, D., Braga, L. P., da Silva, A. M., & Setubal, J. C. (2018). MARVEL, a tool for prediction of bacteriophage sequences in metagenomic bins. Frontiers in genetics, 9, 304.

54. Budach, S., & Marsico, A. (2018). Pysster: classification of biological sequences by learning sequence and structure motifs with convolutional neural networks. Bioinformatics, 34(17), 3035-3037.

55. Tsubaki, M., Tomii, K., & Sese, J. (2019). Compound–protein interaction prediction with end-to-end learning of neural networks for graphs and sequences. Bioinformatics, 35(2), 309318.

56. Dinesh, K. G., Arumugaraj, K., Santhosh, K. D., & Mareeswari, V. (2018). Prediction of cardiovascular disease using machine learning algorithms. International Conference on Current Trends towards Converging Technologies, pp. 1-7. IEEE.

57. Liu, S., Xu, C., Zhang, Y., Liu, J., Yu, B., Liu, X., & Dehmer, M. (2018). Feature selection of gene expression data for cancer classification using double RBF-kernels. BMC bioinformatics, 19(1), 1-14.

58. Camacho, D. M., Collins, K. M., Powers, R. K., Costello, J. C., & Collins, J. J. (2018). Next-generation machine learning for biological networks. Cell, 173(7), 1581-1592.

59. Karr, J. R., Sanghvi, J. C., Macklin, D. N., Gutschow, M. V., Jacobs, J. M., Bolival Jr, B., & Covert, M. W. (2012). A whole-cell computational model predicts phenotype from genotype. Cell, 150(2), 389-401.

60. Karr, J. R., Phillips, N. C., & Covert, M. W. (2014). Whole Cell Sim DB: a hybrid relational/HDF database for whole-cell model predictions

61. Hinton, G. E., Krizhevsky, A., & Wang, S. D. (2011). Transforming auto-encoders. In International conference on artificial neural networks (pp. 44-51). Springer, Berlin, Heidelberg.

62. Sabour, S., Frosst, N., & Hinton, G. E. (2017). Dynamic routing between capsules. arXiv preprint arXiv:1710.09829.

63. Cao, Y., Geddes, T. A., Yang, J. Y. H., & Yang, P. (2020). Ensemble deep learning in bioinformatics. Nature Machine Intelligence, 2(9), 500-508.