Data Mining in Bioinformatics

Aisha Hagar ahaga077@uottawa.ca

Menna Allah Mohammed mfath041@uottawa.ca

Hosna Eltarras helta037@uottawa.ca

Yomna Abdelsattar yabde005@uOttawa.ca

Abstract

Recently, we have been identifying our world as a world of data, due to the explosion in the amount of data we're encountering on a daily basis. Especially in the data we were born with? This is where bioinformatics comes to action, to help us understand more about our biological functions. But bioinformatics can't deal alone with such massive data. It needs more of a systematic methodology for analysing the data and drawing conclusions from it. And this is where data mining comes to assist. In this paper, we shall be introducing the main regions of intersection between data mining and bioinformatics along with giving examples for each topic. Different preprocessing techniques and algorithms from traditional machine learning (including classification and clustering) to deep learning shall be discussed. Eventually, we'll talk about the challenges that are still facing this field.

Keywords: Bioinformatics, classification, clustering, phylogenetic trees, imbalanced data, curse of dimentionality

1. Introduction

In this digital world, millions of data -if not more- is received every day, which might seem like an explosion of data. That's why it was urged that new technologies should emerge so that we can harness this tremendous amount of data, harnessing the data means something like analyzing it to draw insights and patterns that may assist in understanding the surrounding and maybe even taking decisions upon such insights. That's what we know as `data mining` (**?**). The applications of data mining firstly appeared widely in the industry; marketing, customer experience, and so on. However, there has always been another type of data, a massive one that we needed more than improving customer experience or marketing. Data that was nowhere far from us, on the contrary, this data was directly within us. Data that we have always encompassed and yielded to how we're shaped like this, why I'm 160 cm long not 165 or 155, even more crucial why would someone be struck with cancer and not the other, what caused cancer in the first place? All of this information is enclosed inside our bodies, our genes, waiting for us to explore it. Waiting for us to take our specialized equipment and understand, analyze, draw conclusions from it and this is the field of `bioinformatics`. From what is mentioned, it seems undoubtedly that these two fields need to be linked together, a bridge has to be established between these two disciplines. `Bioinformatics`, which is concerned with extensive biological data that is not concretely understood, opens the door immensely for `data mining` methods to unravel the secrets of biological information (2).

Hence, we now know that the ultimate goal of bioinformatics is to be able to find the biological patterns and information hidden in the wide ocean of biological data, and then use this information to improve the understanding of important mechanisms. After scientists have cracked and discovered the genetic codes, they concluded that protein sequences, are tremendously similar to human language in terms of formulation.

There are basic alphabets that comprise the sequence and different combinations of them refer to different meanings. Thus, this necessitates more the need of data and also text mining in this field (3). [3]

The remainder of this paper will go as follows. Section. 2 reviews some of the preprocessing techniques in bioinformatics. Section. 3 gives a brief about clustering in bioinformatics. Section. 4 introduces some classification algorthims. Section. 5 discovers the employment of deeplearning in this field. Section. 6

takes us to the current challenges in bioinformatics. Section 7 finally concludes the paper.

2. Pre-processing

Data preprocessing is a fundamental step in data mining to improve the efficiency of the data. Also, data preprocessing affects the analysis outcome and the modeling algorithms outcome. So, we need a clean, transformed, and reduced data to have a correct and beneficial data.Why we need preprocessing on bioinformatics data?. The reason is the data of this filed mostly has a lack in attribute values or containing only aggregated data. we have three steps in data preprocessing which are 1.Data cleaning 2.Data transformation 3.Data reduction.

2.1. Data cleaning

Most of the time the data has missing and irrelevant values. To handle these problems, we use data cleaning methods to handle missing, incorrect, and inappropriate values. For example, Gene Filtering is a data cleaning method, is aimed at removing the undesirable-genes that contain outliers and too much missing expression values. Data cleaning involves handling of missing data, and this happen because of the data come from different sources so after integration maybe some values were missed or while record the data. Also, data cleaning involves handling of noisy data which is a meaningless data that the machine may misunderstands it and can't interpret it. For example, the inconsistency with values of other dataset's attributes and this happen in a very recent time, we could find that the same gene has two names as two scientists or labs discovered it at the same time in two different continents and each of them gave it a different name.(9; 10)

2.2. Data transformation

To make appropriate analysis on data we should transform it. Normalization is one of the methods of data transformation where the data scaled within specified range. Another method attribute selection in this method we generate new attributes from the given set of attributes to help the mining process. One example of attribute selection is the detection of cancer using RNA, there are some RNA which not causing cancer and there are others that causes the cancer, so by attribute selection we select the effective RNA. Another data transformation method is the generalization method, while the low-level data are replaced by higher level concepts by using the hierarchical concept.(9; 10)

2.3. Data Reduction

The mining on huge amount of data and complex analysis may take a long time. So, we need data reduction to reduce the dataset without losing the important information in the data, increase the storage efficiency and analysis cost. For example, the gene sequence is very large, and we can't visualize it easily, so we use data reduction to reduce the high-dimensional genomics and visualize it easily. Data cube aggregation is one of the data reduction methods it is applied to the data in the construction of data cube, it merges two or more attributes into single attribute. Furthermore, there is dimensionality reduction that make a dimension reduction on the attributes by encoding mechanisms to reduce the size of the data. Also, there is numerosity reduction which the data is replaced or estimated by alternative smaller data representation like a regression model for example.(9; 10)

3. Clustering

Clustering techniques are a means of capturing the hidden patterns in underlying data that may not be feasible to be maintained by human beings. These techniques have shown their efficiency in many fields like the one discussed here: bioinformatics

Bioinformatics and clustering: we have lots of genes that we know their sequences and functions, nevertheless, we have much more ones that we know nothing about their functions even if know about their sequences. Therefore, a main application for clustering in bioinformatics is to take different groups of genes; some with known functions and others without. Then, check to which cluster did our targeted gene go to. The cluster of genes it belonged to probably share some function with it, thus we can infer the function of this gene from its cluster. Also, we may need to find any negative, positive and nonlinear correction between different genes, since genes don't work separately from each other, rather they interact together to perform some function. (4; 2)

3.1. Examples for clustering Algorithms

Here, we will try to demystify this general section with two examples.

3.1.1. Fuzzy Clustering

Fuzzy clustering, also called non-hard clustering, a clustering technique where each data point can be assigned to more than one cluster (5). As an application in bioinformatics: a specific gene may relate to one cluster of genes from a point like its function in one cell and another cluster from another point as its interaction with another gene to perform another function (2).

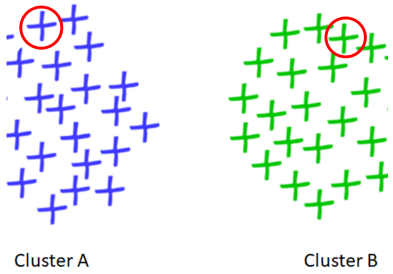


Figure 1: Suppose the two red-cirlcled-Xs are the same data point, it can be assigned to different clusters at the same time

3.1.2. Hierarchical Clustering

Another example of clustering is the Hierarchical/Agglomerative clustering. In bioinformatics this is used as a basis to build phylogenetic trees (2). Phylogenetic trees have different applications, one of them is where we have set of genes; some from human beings, others from animal or plants and we want to find the degree of similarity between each of them.

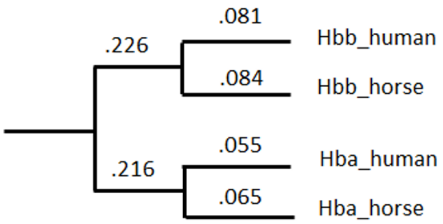


Figure 2: Phylogentic tree showing the similarity between different genes (Hba and Hbb) from human and animal

This is helpful sometimes to find a replacement or a cure for some malfunctioned gene in a human by identifying the most nearby gene to it, where also, clinical trials can be done on the most similar gene from animals instead of experimenting directly on human (6). Something like this is tremendously effective with diseases that are new and we don't know what they do under the hood or where did they come from so that we can treat its patients.

COVID-19, for instance, was a new disease. Now, we know that it belongs to a family of coronaviruses that usually cause specific sort of illnesses like the common cold, but how did we know that? Clinical trials that found this new virus clustered with this family. Then what about covid's origin, How did we know that it came from bats ? From Looking at the phylogenetic tree, scientists knew that a bat coronavirus is the closest relative to SARS-CoV-2 (covid 19), sharing around 96/100 of their genomes (7).

These were only two examples to unravel the employment of clustering in the field of bioinformatics and how beneficial it is. More can be found here (2; 4)

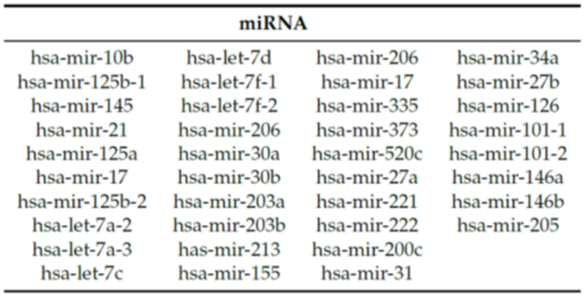
4. Classification

4.1. Motives

Classification algorithms play an important role in the bioinformatics field. Classification can be applied in protein structure prediction, gene classification(22), identifying the gene signature of a disease, classifying patients' data for medical diagnosis and much more. There are several classification algorithms that can be used to perform the mentioned tasks such as KNN, SVM, Decision trees or ensemble methods such as Random forest. There are to main objectives of applying supervised learning in bioinformatics, which are building accurate classifiers or predictive tools and deriving inferences from the results. For classifiers and predictive tools to be considered accurate, they should be able to discriminate between different phenotypes that are under analysis. Biologists are not only interested in accurate predictions, but also, they look for insights that couldn't't be extracted through simple statistical analysis(23). To achieve the first objective, research and experiments have been conducted to compare the performance of classification algorithms for highly dimensional bioinformatics data. This is to determine methodological recommendations for applying these classification algorithms(24). The second objective is necessary for analysis as it allows scientists to make sense of huge biological datasets through recognizing patterns and predictions. For example, scientists analyse protein structure prediction, gene classification, cancer classification, identification of gene expression, protein-protein interactions etc.(22) We will take a look at an experiment that shows the importance of the second objective.

4.2. Validation of miRNAs as Breast Cancer Biomarkers

Researchers have applied classification algorithms along with feature selection methods to analyse the miRNAs that are identified as possible biomarkers for breast cancer detection. Scientists have found that there is a relationship between microRNAs and cancer. But it has been challenging to identify which miRNAs are important for cancer detection because some of then take part in cancer development while others don't, and they differ according to several factors such as the type and location of the cancer being considered. Therefore, machine learning approaches are being used on large datasets of miRNAs for cancer detection. The miRNAs data used was a dataset of miRNAs that have been clinically verified as possible biomarkers for breast cancer detection. The dataset consists of 1207 patient samples with 1881 miRNA features, containing 1110 cancerous samples and 104 healthy samples.



miRNA			
hsa-mir-10b	hsa-let-7d	hsa-mir-206	hsa-mir-34a
hsa-mir-125b-1	hsa-let-7f-1	hsa-mir-17	hsa-mir-27b
hsa-mir-145	hsa-let-7f-2	hsa-mir-335	hsa-mir-126
hsa-mir-21	hsa-mir-206	hsa-mir-373	hsa-mir-101-1
hsa-mir-125a	hsa-mir-30a	hsa-mir-520c	hsa-mir-101-2
hsa-mir-17	hsa-mir-30b	hsa-mir-27a	hsa-mir-146a
hsa-mir-125b-2	hsa-mir-203a	hsa-mir-221	hsa-mir-146b
hsa-let-7a-2	hsa-mir-203b	hsa-mir-222	hsa-mir-205
hsa-let-7a-3	has-mir-213	hsa-mir-200c	
hsa-let-7c	hsa-mir-155	hsa-mir-31	

Figure 3: Clinically Verified MiRNAs

In the flow of the experiment, we can see that in the pre-processing stage, several feature selection methods were performed to select a subset of miRNAs. MiRNAs were grouped into subsets of (3, 5, 10) to test their effectiveness in cancer detection using different feature selection techniques along with different classification algorithms. The validation technique used is 10-fold cross validation on 10% of the data. Then Random forest and SVM were applied to determine if the target is cancerous or not.

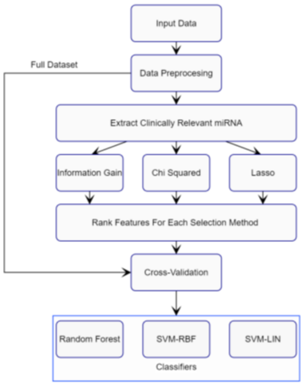
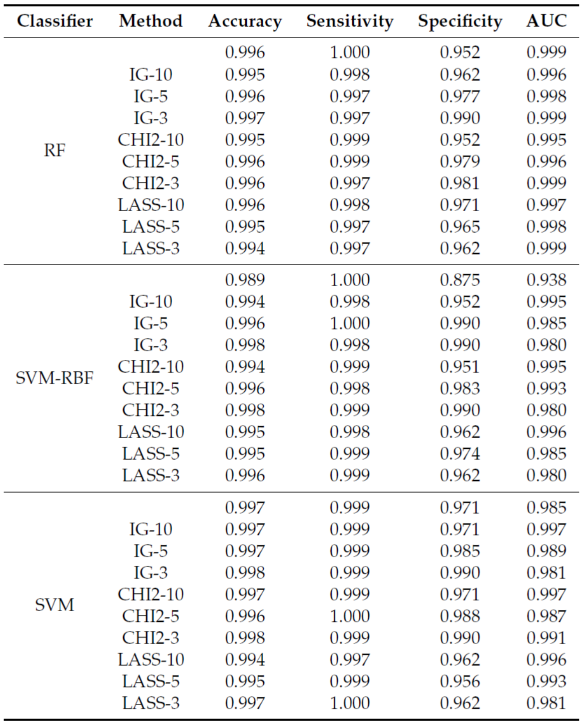


Figure 4: Schematics for Cancer Detection with Machine Learning

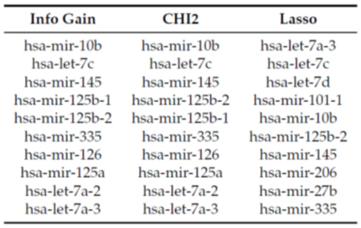
The evaluation of the classification didn't solely rely on accuracy as the data was imbalanced. So, relying on accuracy metric only will be misrepresentative for the model performance evaluation.



Classifier	Method	Accuracy	Sensitivity	Specificity	AUC
RF		0.996	1.000	0.952	0.999
	IG-10	0.995	0.998	0.962	0.996
	IG-5	0.996	0.997	0.977	0.998
	IG-3	0.997	0.997	0.990	0.999
	CHI2-10	0.995	0.999	0.952	0.995
	CHI2-5	0.996	0.999	0.979	0.996
	CHI2-3	0.996	0.997	0.981	0.999
	LASS-10	0.996	0.998	0.971	0.997
	LASS-5	0.995	0.997	0.965	0.998
	LASS-3	0.994	0.997	0.962	0.999
SVM-RBF		0.989	1.000	0.875	0.938
	IG-10	0.994	0.998	0.952	0.995
	IG-5	0.996	1.000	0.990	0.985
	IG-3	0.998	0.998	0.990	0.980
	CHI2-10	0.994	0.999	0.951	0.995
	CHI2-5	0.996	0.998	0.983	0.993
	CHI2-3	0.998	0.999	0.990	0.980
	LASS-10	0.995	0.998	0.962	0.996
	LASS-5	0.995	0.999	0.974	0.985
	LASS-3	0.996	0.999	0.962	0.980
SVM		0.997	0.999	0.971	0.985
	IG-10	0.997	0.999	0.971	0.997
	IG-5	0.997	0.999	0.985	0.989
	IG-3	0.998	0.999	0.990	0.981
	CHI2-10	0.997	0.999	0.971	0.997
	CHI2-5	0.996	1.000	0.988	0.987
	CHI2-3	0.998	0.999	0.990	0.991
	LASS-10	0.994	0.997	0.962	0.996
	LASS-5	0.995	0.999	0.956	0.993
	LASS-3	0.997	1.000	0.962	0.981

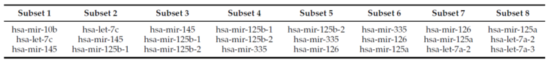
Figure 5: Performance Metrics of Classifiers with Different Feature Selection Methods Over MiRNAs Subsets (3, 5, 10)

In this experiment positive class means cancerous and negative class means non-cancerous. The performance was evaluated according to: Specificity which shows the amount of non-cancerous that has been correctly identified, sensitivity which shows the proportion of cancerous samples that have been predicted correctly, accuracy which shows the proportion of correct classifications of both cancerous and non-cancerous samples, AUC which the ability of the classifiers to differentiate between the two classes. Due to the imbalanced dataset, it is important to focus on specificity metric to know the amount of the non-cancerous samples that have been correctly identified as they fall in the minority class. By examining the results, it was seen that it may be sufficient to focus on a few miRNAs to diagnose patients. The experiment was repeated, and feature selection techniques were applied to determine the top ranked miRNAs. Then the top 24 ranked miRNAs were divided onto subsets in their same ranking order and the subsets were used in Random Forest and SVM classifiers.



Info Gain	CHI2	Lasso
hsa-mir-10b	hsa-mir-10b	hsa-let-7a-3
hsa-let-7c	hsa-let-7c	hsa-let-7c
hsa-mir-145	hsa-mir-145	hsa-let-7d
hsa-mir-125b-1	hsa-mir-125b-2	hsa-mir-101-1
hsa-mir-125b-2	hsa-mir-125b-1	hsa-mir-10b
hsa-mir-335	hsa-mir-335	hsa-mir-125b-2
hsa-mir-126	hsa-mir-126	hsa-mir-145
hsa-mir-125a	hsa-mir-125a	hsa-mir-206
hsa-let-7a-2	hsa-let-7a-2	hsa-mir-27b
hsa-let-7a-3	hsa-let-7a-3	hsa-mir-335

Figure 6: Top Ranked Features Under Different Feature Selection Techniques



Subset 1	Subset 2	Subset 3	Subset 4	Subset 5	Subset 6	Subset 7	Subset 8
hsa-mir-10b hsa-let-7c hsa-mir-145	hsa-let-7c hsa-mir-145 hsa-mir-125b-1	hsa-mir-145 hsa-mir-125b-1 hsa-mir-125b-2	hsa-mir-125b-1 hsa-mir-125b-2 hsa-mir-335	hsa-mir-125b-2 hsa-mir-335 hsa-mir-126	hsa-mir-335 hsa-mir-126 hsa-mir-125a	hsa-mir-126 hsa-mir-125a hsa-let-7a-2	hsa-mir-125a hsa-let-7a-2 hsa-let-7a-3

Figure 7: Subsets of Ranked miRNAs

The evaluation was done by looking at the specificity metric of the classifiers' predictions using the miRNAs subsets. There was a decline in the specificity metric as the subsets go from 1 to 8.

Through feature selection and classification, researchers were able to discover that a minimum of three miRNAs as breast cancer biomarkers can be used instead of 1881 miRNAs(25). This experiment shows the impact machine learning can have in the diagnosis and analysis of breast cancer.

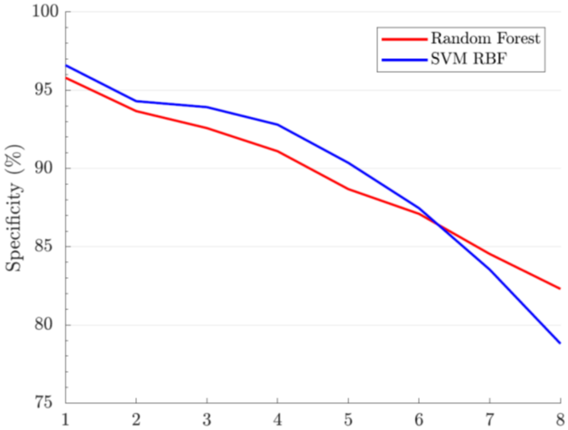


Figure 8: Specificity Across Different Clinical miRNA Subsets

5. Deep Learning

5.1. Motives

The proper performance of conventional machine learning algorithms relies heavily on data representations (extracted features). However, features are typically designed by human engineers with extensive domain expertise, and identifying which features are more appropriate for the given task remains difficult. Deep learning has recently emerged based on big data, the power of parallel and distributed computing, and sophisticated algorithms. To advance from hand-designed to data-driven features, representation learning through deep learning has shown great promise.

Representation learning can discover effective features as well as their mappings from data for given tasks. Furthermore, deep learning can learn complex features by combining simpler features learned from data.Thus, hierarchical representations of data can be discovered with increasing levels of abstraction(11). In other words, due to the enormous amount of biological data in bioinformatics applications, as well as the need for compute capacity, and the challenge in extracting features, deep learning is proved to be a good approach for various bioinformatics problems.

5.2. Deep learning architectures

5.2.1. Deep Neural Network (DNN)

According to the classification of (11), DNNs can be classified as Multilayer perceptron (MLP), stacked auto-encoder (SAE), or deep belief networks (DBN).

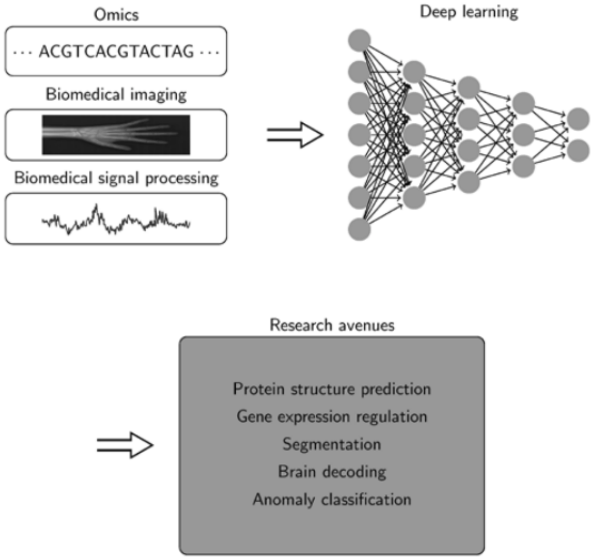


Figure 9: Deep learning applications in bioinformatics

Since the training method is a process of optimization in high-dimensional parameter space, MLP is typically used when a large number of labeled data are available. MLP represent one of the widely used and effective machine learning methods currently applied to diagnostic classification based on high-dimensional genomic data. Since the dimensionalities of the existing genomic data often exceed the available sample sizes by orders of magnitude, the DNN performance may degrade owing to the curse of dimensionality and over-fitting, and may not provide acceptable prediction accuracy. However, some researches provided optimization schemes to ease the curse of dimensionality as in this deep belief network research(12).

5.2.2. Convolutional neural network (CNN)

CNNs are architectures that consist of convolution layers, nonlinear layers, and pooling layers. Recently, CNNs have been adopted rapidly in biomedical imaging studies for its outstanding performance in computer vision and concurrent computation with GPUs. Usually convolution-pooling structure can better learn imaging features from CT scans and MRI images from head trauma, stroke diagnosis and brain EPV (enlarged perivascular space) detection (13). Furthermore, CNN can be combined with other deep learning models, such as RNN to predict imaging content, where CNN encodes an image and RNN generates the corresponding image description(14).

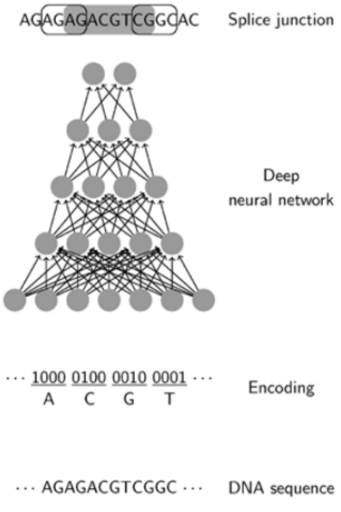


Figure 10: MLP applications in bioinformatic

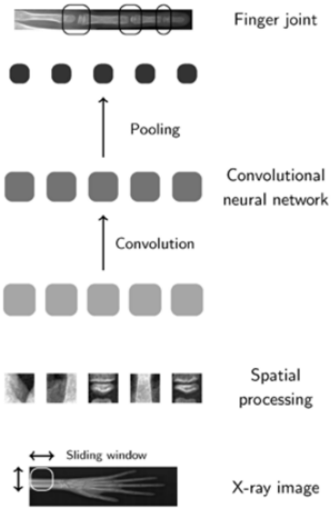


Figure 11: CNN applications in bioinformatic

5.2.3. Recurrent neural network (RNN)

RNNs are designed to utilize sequential information of input data with cyclic connections among building blocks like perceptrons, long short-term memory units (LSTMs), or gated recurrent units (GRUs). Since omics data and biomedical signals are typically sequential and often considered languages of nature, the capabilities of RNNs for mapping a variablelength input sequence to another sequence or fixed-size prediction are promising for bioinformatics research. It's very suitable for long and sequential data, such as DNA array and genomics sequence(11).

RNN usually performs worse than Convolutional Neural Network (CNN) in terms of fine-tuning. But frequently it is ensembled with CNN in diverse applications, such as biological data dimension reduction and medical imaging processing.

But RNN cannot interact with hidden neurons far from the current one. To construct an efficient framework of recalling deep memory, many improved algorithms have been proposed, like Bidirectional-RNN (BRNN) in protein secondary structure prediction (15), and Multi-dimensional-RNN (MD-RNN) in analyzing electron microscopy and MRIs of breast cancer samples(16).

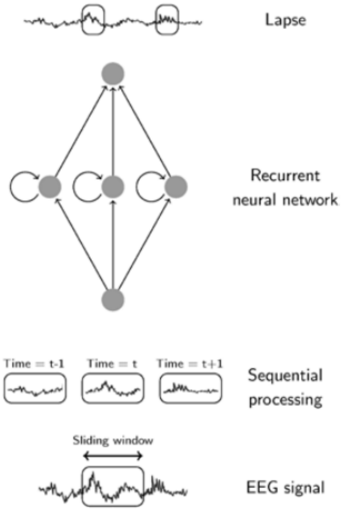


Figure 12: RNN applications in bioinformatic

5.2.4. Auto encoders (AE)

Through an unsupervised manner, autoencoder is another typical artificial neural network, designed to precisely extract coding or representation features using data-driven learning. For high-dimensional data such as biological data, it is time-consuming and infeasible to load all raw data into a network, thus dimension reduction or compression is a necessity in preprocessing of raw data.

Autoencoder can compress and encode information from the input layer into a short code, then after specific processing, it will decode into the output closely matching the original input. Similar to traditional PCA in dimension reduction to some extent, but autoencoder is more robust and effective in extracting data features for its non-linear transformation in hidden layers.

Multiple autoencoders can be stacked to act as a deep autoencoder. Typically, stacked sparse autoencoder (SSAE) was proposed to analyze high-resolution histopathological images in breast cancer(17). Also, by using SAE with three iterations, it was reported to achieve successful prediction of protein secondary structure, local backbone angles, and solvent accessible surface area (18).

5.2.5. Convolutional auto encoder (CAE

CAEs are designed to utilize the advantages of both AE and CNNs so that it can learn good hierarchical representations of data reflecting spatial information and be well regularized by unsupervised training. In training of AEs, reconstruction error is minimized using an encoder and decoder, which extract feature vectors from input data and recreate the data from the feature vectors, respectively. In CNNs, convolution and pooling layers can be regarded as a type of encoder. Both convolution and pooling can compress data while preserving the most representative features in two different ways. Therefore, the CNN encoder and decoder consisting of deconvolution and unpooling layers are integrated to form a CAE and are trained in the same manner as in AE (11).

5.3. Ensemble deep learning

Deep learning models are not without shortcomings: they often exhibit high variance and may fall into local loss minima during training. Indeed, empirical results of ensemble methods that combine the output of multiple deep learning models have been shown to achieve better generalizability than a single model. In addition to simple ensemble approaches such as averaging output from individual models, combining heterogeneous models enables multifaceted abstraction of data, and may lead to better learning outcomes.

For example in genome analysis, the use of supervised CNN and LSTM models allows both global and local sequential features to be captured, and further integration with unsupervised convolutional autoencoders which enables unsupervised pre-training, an effective component for handling small sample size (19).

5.4. Transfer learning

Transfer learning is frequently discussed in the deep learning fields for its great applicability and performance. Ensembled with CNN, transfer learning can attain greater prediction performance of interstitial lung disease CT scans (20). It was also used as a ligament between the multi-layer LSTM and conditional random field (CRF), and the result showed that the LSTM-CRF approach outperformed the baseline methods on the target datasets (21).

6. Challenges

A glimpse of the grand challenges that are still surrounding the field of bioinformatics and need the incorporation of data mining is shown in this table (2; 8).

Table 1: Grand challenges

Main Aspect	Issues
Data Integration	1. Heterogeneous biological data 2. Diversity of data without standard scheme
Data Management	1. High-Memory for distribution 2. Curse of dimentionality and latency
Data Preprocessing	1. Imbalanced data 2. Complex features
Models	1. Select appropriate models 2. Interpretability of models 3. visualize high dimensions

7. Conclusion

Bioinformatics is an active area of research that tries to find answers to biomedical questions. Since it encompasses a great amount of data, it requires the assistance of data mining to help understanding our life more. Data mining, with its advancement, has incorporated with bioinformatics in the preprocessing, machine and deep learning algorithms including classification and clustering which we have discussed a glimpse of them and saw how this opened a wider world for discovery along side with yielding solutions to our concurrent problems. Meanwhile, there still a lot of challenges in this field that entails the endeavours of scientists from both areas to reach our aspired level of understanding our biological life.

References

[1] Mohammed J Zaki, George Karypis, and Jiong Yang, Data Mining in Bioinformatics (BIOKDD), licensee BioMed Central Ltd

[2] Lan, K., Wang, Dt., Fong, S. et al. A Survey of Data Mining and Deep Learning in Bioinformatics. J Med Syst 42, 139 (2018). https://doi.org/10.1007/s10916-018-1003-9

[3] Zhiqiang Zeng, Hua Shi, Yun Wu,et al. (2016) Survey of Natural Language Processing Techniques in Bioinformatics, College of Computer and Information Engineering, Xiamen University of Technology, Xiamen 361024, China

[4] Second: Muhammad Ali Masood, M. N. A. Khan, (2015) Clustering Techniques in Bioinformatics, Shaheed Zulfikar Ali Bhutto Institute of Science and Technologies, Islamabad, Pakistan

[5] Fuzzy clustering - Wikipedia

[6] Garcés-Ayala, F., Araiza-Rodríguez, A., Mendieta-Condado, E. et al. Full genome sequence of the first SARS-CoV-2 detected in Mexico. Arch Virol 165, 2095–2098 (2020). https://doi.org/10.1007/s00705-020-04695-3

[7] Marco Cascella; Michael Rajnik; Abdul Aleem; Scott C. Dulebohn; Raffaela Di Napoli. (2021) Features, Evaluation, and Treatment of Coronavirus (COVID-19)

[8] Manisha Mathur (2018) Biomedical Challenges: A review, Post Graduate Institute of Veterinary Education and Research, Rajasthan University of Veterinary and Animal Science, Jaipur, Rajasthan, India

[9] A Survey on Data Preprocessing Techniques for Bioinformatics and Web Usage Mining 1A. Sivakumar and 2R.Gunasundari 1Department of Computer Science, Karpagam University, Coimbatore. sivamgp@gmail.com 2Department of Information Technology, Karpagam University, Coimbatore.

[10] Jamshed, H., Ali Khan, M. S., Khurram, M., Inayatullah, S. y Athar, S. (2019). Data Preprocessing: A preliminary step for web data mining, Pakistan.

[11] Min, S., Lee, B., and Yoon, S. (2016). Deep learning in bioinformatics. Briefings in Bioinformatics, bbw068.

[12] Ghasemi, F., Mehridehnavi, A., Fassihi, A., and Pérez-Sánchez, H. (2018). Deep neural network in QSAR studies using deep belief network.

[13] Chilamkurthy, S., Ghosh, R., Tanamala, S., Biviji, M., Campeau, N. G., Venugopal, V. K., et al. (2018). Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study. Lancet 392, 2388–2396. doi: 10.1016/S0140-6736(18)31645-3

[14] Angermueller, C., Pärnamaa, T., Parts, L., and Stegle, O. (2016). Deep learning for computational biology. Mol. Syst. Biol. 12:878. doi: 10.15252/msb.20156651

[15] Baldi, P., Brunak, S., Frasconi, P., Soda, G., and Pollastri, G. (1999). Exploiting the past and the future in protein secondary structure prediction. Bioinformatics 15:937. doi: 10.1093/bioinformatics/15.11.937

[16] Kim, Y., Sim, S. H., Park, B., Lee, K. S., Chae, I. H., Park, I. H., et al. (2018). MRI assessment of residual breast cancer after neoadjuvant chemotherapy: relevance to tumor subtypes and MRI interpretation threshold. Clin. Breast Cancer 18, 459–467.e1 doi: 10.1016/j.clbc.2018.05.009

[17] Xu, J., Xiang, L., Liu, Q., Gilmore, H., Wu, J., Tang, J., and Madabhushi, A. (2016). Stacked sparse autoencoder (SSAE) for nuclei detection on breast cancer histopathology images. IEEE Trans. Med. Imaging 35, 119–130. doi: 10.1109/TMI.2015.2458702

[18] Heffernan, R., Paliwal, K., Lyons, J., Dehzangi, A., Sharma, A., Wang, J., et al. (2015). Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning. Sci. Rep. 5:11476. doi: 10.1038/srep11476

[19] Cao, Y., Geddes, T. A., Yang, J. Y. H., and Yang, P. (2020). Ensemble deep learning in bioinformatics. Nature Machine Intelligence, 2(9), 500–508. https://doi.org/10.1038/s42256-020-0217-y

[20] Anthimopoulos, M., Christodoulidis, S., Ebner, L., Christe, A., and Mougiakakou, S. (2016). Lung pattern classification for interstitial lung diseases using a deep convolutional neural network. IEEE Trans. Med. Imag. 35, 1207–1216. doi: 10.1109/TMI.2016.2535865

[21] Giorgi, J. M., and Bader, G. D. (2018). Transfer learning for biomedical named entity recognition with neural networks. Bioinformatics 34, 4087–4094. doi: 10.1093/bioinformatics/bty449

[22] Singh, Pushpa & Singh, Narendra & Bajaj, G. (2020). Role of Data Mining Techniques in Bioinformatics. International Journal of Applied Research in Bioinformatics. 11. 51-60. 10.4018/IJARB.2021010106.

[23] Dua, S., & Chowriappa, P. (2012). Data Mining for Bioinformatics (1st ed.). CRC Press. https://doi.org/10.1201/b13091

[24] Bichindaritz, Isabelle. “Methods in Case-Based Classification in Bioinformatics: Lessons Learned.” Advances in Data Mining. Applications and Theoretical Aspects, Springer Berlin Heidelberg, pp. 300–13, doi:10.1007/978-3-642-23184-1_23.

[25] Rehman O, Zhuang H, Muhamed Ali A, Ibrahim A, Li Z. Validation of miRNAs as Breast Cancer Biomarkers with a Machine Learning Approach. Cancers. 2019; 11(3):431. https://doi.org/10.3390/cancers11030431