Review on the Application of Artificial Intelligence in Bioinformatics

Chuning Liu [1, a], Xuan Liu [2, b], Haotian Shangguan [3, c], Siyan Wen [4, d], Feiyang Zheng [5, e]

[1] Shenzhen Middle School, 18 Shenzhong Street, Renmin Road North, Luohu District, Shenzhen, China

[2] Tilton School, 30 School Street, Tilton NH 03276, USA

[3] Penn State University, 201 Old Main, University Park, Pennsylvania, PA 16801, USA

[4] University of Connecticut, 1 University Pl, Stamford, CT 06901, USA

[5] Xinghai Conservatory of Music, Waihuan Xilu, Panyu District, Guangzhou, China

[a] 2453835242@qq.com, [b] 34310050@qq.com, [c] Haoshanggu19@gmail.com, [d] siyan.wen@uconn.edu, [e] yangyang7bob@163.com

Abstract. Compared with traditional data, biomedical data has the characteristics of less samples, high dimension, unstructured data, more types of data, and huge amount of data. Bioinformatics processing requires the intersection of statistics, mathematics, and computer science. Artificial intelligence technology has been used in genome annotation, drug design, structure prediction and other research, helping biologists to screen, process, interpret and utilize the huge data collected in biological research. This paper reviews the application of artificial intelligence in biological information, introduces the application and research progress of machine learning in gene analysis, disease diagnosis, brain image processing, etc., and finally summarizes the full text.

Keywords: Artificial Intelligence; Bioinformatics; Genome Analysis.

1. Introduction

Nowadays, Artificial Intelligence has already implemented into different kinds of field such as Machine Learning, Robotics, Computer Vision, and Natural Language Processing. Some of AI related technologies are also becomes popular in industry, society, business, and healthcare as well. In the 1940s, researchers tried to develop computer programs that could be used as medical consultants to assist clinicians in diagnosis and treatment. This was the first report of artificial intelligence in medicine]. Now, there are also lots of research studies suggest that AI can perform as well as or better than humans at key healthcare tasks, such as diagnosing disease. In addition, AI has two advantages to human performance which is high efficiency to learn from big data and help clinicians in making decisions in real time and can perform predefined tasks with higher precision. Therefore, AI shows a strong positive effect for different kinds of diseases nowadays. Artificial Neural Networks also helps a lot in disease diagnosis. In 1988, Szolovits et al. proposed to apply deep neural networks to the diagnosis and treatment of diseases. Since AI is also related to some different technology such as computer vision, Natural Language Processing, and Machine Learning, we will talk about how each of them makes connection with disease diagnosis.

In the early period, the most common application use Machine Learning to improve the efficiency is precision medicine. Then, when Machine Learning becomes more powerful, it starts play important role in disease diagnosis. For cancer diagnosis, Machine Learning can play very important role to help professionals to analyze this disease. For example, Machine Learning can analyze gene expression data to determine whether patients have disease or not. In addition, AI also can help to determine the survival rates of cancer patients. As we mentioned above, neural network and deep

learning also very significant for disease diagnosis field. Combine Machine Learning and them together, there is an application which can recognize potential cancerous lesions in radiology images.

Computer Vision for medical imaging also is important class of disease diagnosis. Since we are living post pandemic of COVID-19, it always been an important topic that researcher want to solve. However, there are already several research has been done to uses Computer Vision to against COVID-19. Currently, we are still using RT-qPCR to detect COVID-19 but this kind of test can still cause false reports. Therefore, to acquire image we can uses other alternative method such as CT scan and X-ray Image. In the other way, computer vision also has different function such as thermal imaging, multidimensional imaging, and image segmentation can helps to improve image quality and analysis efficiency.

Lastly, Natural Language Processing also has some important function that can help to detect disease diagnosis. Currently, NLP, such as speech recognition, text analysis, and translation can help whether clinicians or patients to analyze unstructured clinical notes. In addition, there are also some studies have done about how Natural Language Processing to help detect depression. Since depression will affects speech production which we can uses this point try to detect whether patient has depression or not.

However, AI currently can perform as well as or even better than human in this field, there is still have ethical issues can be very important for whether people will accept AI system or not. Nowadays, transparency is one of biggest difficult for AI algorithm especially for deep learning which are impossible to interpret or explain how we get the result. Therefore, giving patients a briefly explanation can be helpful to solve this problem and people can focus on this direction in the future work. In addition, AI could also make mistakes. Therefore, how can we prevent this kind situation or how to provides an alternative solution for this scenario can be a future study need to focus on. Lastly, in the technology side, how to use artificial intelligence technology to process multi-modal medical data, make full use of comprehensive information such as various data, text, and image for disease diagnosis modeling, and improve the accuracy and generalization ability of diagnostic models still need further research, which is also a technical challenge. In addition, due to the particularity of the medical field, the interpretability of diagnostic models is also highly required.

According to the current development of AI-assisted disease diagnosis, the development of AI-assisted disease diagnosis should be promoted from the following aspects in the future. From the aspect of computer technology, aiming at the situation that the generalization ability of the existing model is not strong, and the efficiency is not high, we should study a better algorithm, establish a more stable model, and improve the accuracy of disease diagnosis. In addition, for the situation of low accuracy, we should start with the extraction of disease information data to find better features to label disease information.

2. Application of Artificial Intelligence in Genome Analysis

Genome is a word normally use to describe all the inherited information of an organism; it is being saved as the form of nucleotide sequences in cells and it often but not always manifest as DNA.

Genetic Diagnosis & (Mutation Classification). Genome sequence is a laboratory method that being used to decipher the order of DNA bases. The purpose for decipher the DNA is to discover the changes that occurs in the genetic code that may indicate one's health status. These changes can also be called as variants and mutations. Genome sequences may also help people to predict their risks of getting a specific disease due to a gene defect. However, it only works based on genes that are already known, and until recently, what was known was still limited. There were already three generations of gene sequencing method have been developed in the pass. They are Sanger Sequencing (First Generation), Next generation sequencing or NGS (Second Generation), and single molecule sequencing (Third Generation). Among those three generations of genome sequence, the Sanger sequencing is still the gold standard method, with such a high precision it possesses, it also becomes a method being used to validate the results of other methods.

Artificial intelligence can be used as a tool for genomic diagnosis. The uses of artificial intelligence in genomic diagnostics included directions like: Image diagnosis and electronic health record diagnosis. In these areas, AI shows a great performance in some tasks, sometimes even better than human beings. Image diagnosis, for example, uses computer-aided facial phenotyping tools like DeepGestalt, which is a CNN based facial image analysis algorithm, could extract features from photos and analyze those features in comparison to 1007 distinct terms of different facial abnormalities and an accuracy higher than human dysmorphologies. These 1007 different facial abnormality terms are associated with 4526 different diseases and 2142 genes, and the AI is even accurate enough to distinguish molecular diagnoses that map to the same clinical diagnosis.

Genome assembly. Genome assembly can basically be separated in two kinds of approaches, which is comparative assembly and de novo assembly. Comparative assembly, which is self-explanatory, is a type of assembly method that uses already existing genome as a reference to assemble target genes, which usually has relatively good results (if the reference genome is available). In contrast, de novo assembly does not require a reference genome, but its assembly process can be relatively more difficult than comparative assembly. Current mainstream genome sequence methods like shotgun, which is a method based on Sanger Sequencing, still have some drawbacks, such as the limited number of nucleotides it can read in each DNA fragment. However, the NGS method have used a process of first sequencing, then generating text sequences, and finally combining to produce the complete DNA sequence to circumvented this drawback. Existing automated assemblies include Palmer et al.'s OLC-based Minimus assembler, Zhu et al.'s supervised learning-based PERGA , and Ji et al.'s metaSort assembler, which have more than 99% accuracy in performance evaluation. Like Kleber et al. mentioned, these ML techniques are being used as a strategic way in the assembly process rather than just a plug-in way. However, although these assembly tools are well developed, they still have difficulties in dealing with some special cases (e.g., genetic variants), so there is still not an effective enough solution to these problems.

Artificial intelligence has shown a very wide range of application prospects and great potential in genomic analysis and even in the whole medical field. In recent years, healthcare systems around the whole world have been faced a great challenge due to pandemic. However, there is no doubt that the human medical system has been gradually transformed, and we will usher in a new era. As a lot of people mentioned, artificial intelligence is the next step towards precision pathology. Currently, there are many studies related to COVID-19 (virus) as well as cancer diagnosis, and these may become the main research directions in the future. For example, the study of genomic sequence analysis of COVID-19 and similar viruses using artificial intelligence by Imran et al. [10] showed us the feasibility of using AI for genome sequence analysis and classification of viruses. The case of COVID-19 has also appeared in different variants with high frequency when it was rampant around the world, so the analysis of different variants of the same virus using artificial intelligence may also become a valuable research direction in the future. Another representative example for cancer diagnosis may be the study that Coudray et al. have done about lung cancer using neural networks. They used CNN as a tool digitally scanned the samples from the cancer genome atlas to diagnose the samples to either LUAD, LUSC or normal lung tissue. Other potentially valuable directions may include atrial fibrillation prediction, diabetes, stroke and other disease prediction, etc. Overall the application of AI in precision medicine is very promising, although it still has some limitations and challenges, but it is believed that in the future, AI will have more beneficial impact on healthcare.

3. Artificial Intelligence in Brain Images

Artificial intelligence (AI) is a division of computer science that is capable of analyzing complex biomedical data. AI have many main branches including machine learning, representation learning and deep learning. AI has great potential in both brain imaging data processing and brain disease diagnosis. Increasing number of machine learning algorithms have been developed to address the pre-processing of brain images data including segmentation, normalization, smoothing. In addition, there

are more and more deep learning frameworks have been proposed to help with the brain diseases diagnosis, treatment, and prediction. All these cutting-edge innovations will lay the foundation for advances in biomedical brain imaging.

Before the application of artificial intelligence into brain images, one thing is indispensable, that is data. Brain images data could be divided into the two categories: structural images and functional images. Structural brain images are, as the name implies, a realistic quantification of the physical structure of the brain. In neuroimages, scientist usually use voxel to represent the minimize unit of the brain structure. At here, for the structural image's detection, the scanner will detect the brain from three special direction called sagittal, axial, and coronal, respectively. Scientist will need a very high-resolution equipment in order to get detail information of subject's brain structure such as gray matter, white matter tracts or vasculature. Functional images were usually used to study a specific type of brain capability, often including a time series along with huge amount of brain images data. It could detect the task-related activity and connectivity of the brain. Every 2 second or so, whole brain will be scan. Functional images are capable to reveal the brain structure and brain dynamic, but it also comes with an enormous amount of relatively low-resolution brain images data. Therefore, it is indisputable that every brain imaging test will be accompanied by a large amount of imaging data, both for structural and functional images.

Machine learning application in brain image data processing could be easily understand by illustrating a brain images database call SRPBS database. This database use resting-state functional magnetic resonance imaging (rs-fMRI) to examine the relationship between neuroimaging and psychiatric disorders. The whole database comprises rs-fMRI and structural images from 993 psychiatric disorder patients and 1421 healthy individuals. The psychiatric disorder in this study included autism spectrum disorder, major depression disorders, obsessive compulsive disorder, schizophrenia, chronic pain and others. This database contain subject from a large geographical scale across the whole country of Japan, and then finish the experiment from 8 different sites. The main purpose of doing that is to improve the accuracy of the database. In addition, this study uses nine healthy traveling subjects to visit each sites in order to cancel out the differences between sites. This is called data harmonization.

For the data preprocessing part, this database uses a software package on MATLAB called statistical parametric mapping 8 (SPM8). The actual preprocessing steps are the following: timing correction, realignment, co-registration, segmentation, normalization and spatial smoothing, bandpass filer, and regression. After that, SPM generate the resting-state functional connectivity matrix of each individual.

Deep learning could also be applied into brain image data analysis. A great example of this called patch based hierarchical network (PHN) . This is a deep learning framework for brain image assisted diagnosis of many major psychiatric disorders. In general, this framework includes two stages: pre-attention stage which is stage 1 and local feature extraction and classification stage that is stage 2. Stage 1 is service for stage 2, which could be helpful for reducing the noise information in the brain images data. The key deep learning algorithms within this system is called P-FEN, patch-based feature extraction network. This network could extract the core features that is essential for mental disorder diagnosis from images and then evaluate the possibility of each disorder that this patient could have. Basically, what stage 1 was doing is randomly separate many images' patches with a specific size from input. And then, P-FEN will form a attention map for each patches. After that, stage two will choose the top 10 patches from the attention map and further evaluated then by P-FEN again. Finally, this deep learning system will generate a class prediction probability for each patient to reveal which mental disorder that is person could have.

Both brain imaging studies detailed by the authors are relevant to psychiatric diagnosis. The author here only uses psychiatric diagnosis as an example to demonstrate the importance of machine learning and deep learning for brain imaging studies. Psychiatric disorders are only part of the brain diseases. The authors believe that machine learning and deep learning should be equally important for other brain diseases, such as epilepsy, stroke, and tumors. At the same time, the application of machine

learning and deep learning in brain could not only limit to the MRI data, but also could be extend to other data source such as PET or CT.

4. Conclusion

With the development of information technology and the emergence of various artificial intelligence technologies, the speed and accuracy of biological information processing have been greatly promoted, and the development of bioinformatics has also been promoted. At present, artificial intelligence technology has entered a new stage.. This paper introduces the latest theoretical research results of the intersection of artificial intelligence and biological information processing, and introduces the application of artificial intelligence technology in the fields of bioinformatics and precision medicine.

References

[1] A. D. Goldman and L. F. Landweber, "What is a genome?," PLOS Genetics, vol. 12, no. 7, pp. 1–5, 2016.

[2] K. A. Phillips, J. R. Trosman, R. K. Kelley, M. J. Pletcher, M. P. Douglas, and C. B. Weldon, "Genomic sequencing: Assessing the health care system, policy, and big-data implications," Health Affairs, vol. 33, no. 7, pp. 1246–1253, 2014.

[3] J. M. Heather and B. Chain, "The sequence of sequencers: The history of sequencing DNA," Genomics, vol. 107, no. 1, pp. 1–8, 2016.

[4] Y. Gurovich, Y. Hanani, O. Bar, G. Nadav, N. Fleischer, D. Gelbman, L. Basel-Salmon, P. M. Krawitz, S. B. Kamphausen, M. Zenker, L. M. Bird, and K. W. Gripp, "Identifying facial phenotypes of genetic disorders using Deep Learning," Nature Medicine, vol. 25, no. 1, pp. 60–64, 2019.

[5] K. Padovani de Souza, J. C. Setubal, A. C. Ponce de Leon F. de Carvalho, G. Oliveira, A. Chateau, and R. Alves, "Machine learning meets genome assembly," Briefings in Bioinformatics, vol. 20, no. 6, pp. 2116–2129, 2018.

[6] D. D. Sommer, A. L. Delcher, S. L. Salzberg, and M. Pop, "Minimus: A fast, lightweight genome assembler," BMC Bioinformatics, vol. 8, no. 1, 2007.

[7] X. Zhu, H. C. Leung, F. Y. Chin, S. M. Yiu, G. Quan, B. Liu, and Y. Wang, "Perga: A paired-end read guided de novo assembler for extending contigs using SVM and look ahead approach," PLoS ONE, vol. 9, no. 12, 2014.

[8] Ramesh, A. N., Kambhampati, C., Monson, J. R. T., & Drew, P. J. (2004). Artificial Intelligence in medicine. Annals of The Royal College of Surgeons of England, 86(5), 334–338. https://doi.org/ 10.1308/ 147870804290.

[9] Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, et al. Artificial intelligence in healthcare: past, present, and future. Stroke Vasc Neurol. (2017) 2:230–43. Doi: 10.1136/svn-2017-000101.

[10] Li, Q., Wang, W., Guo, Q., Jiang, L., Qiao, K., Hu, Y., Zhang, X., Wang, Z., Peng, D., Fan, Q., Zhao, M., Fang, Y., Wang, J., Qiu, H., Wang, J., Li, G., Sheng, J., Li, C., & Yang, Z. (2022). Deep Learning System for Brain Image-aided diagnosis of multiple major mental disorders. https://doi. org/ 10.1101/ 2022. 06. 01. 22275855.

[11] Hirsch, G. V., Bauer, C. M., & Merabet, L. B. (2015). Using structural and functional brain imaging to uncover how the brain adapts to blindness. Journal of Psychiatry and Brain Functions, 2(1), 7. https:// doi.org/10.7243/2055-3447-2-7.

[12] Smith, S. M. (2001). Overview of fmri analysis. Functional Magnetic Resonance Imaging, 216–230. https:// doi.org/10.1093/acprof: oso/9780192630711.003.0011.

[13] Tanaka, S. C., Yamashita, A., Yahata, N., Itahashi, T., Lisi, G., Yamada, T., Ichikawa, N., Takamura, M., Yoshihara, Y., Kunimatsu, A., Okada, N., Hashimoto, R., Okada, G., Sakai, Y., Morimoto, J., Narumoto, J., Shimada, Y., Mano, H., Yoshida, W., … Imamizu, H. (2021). A multi-site, multi-disorder resting-state magnetic resonance image database. Scientific Data, 8(1). https://doi.org/10.1038/s41597-021-01004-8.

[14] Chen C, Gao G, Xu Y, Pu L, Wang Q, Wang L, Wang W, Song Y, Chen M, Wang L, Yu F (n.d.) SARSCoV-2–positive sputum and feces after conversion of pharyngeal samples in patients with COVID-19. Ann Intern Med.

[15] Bhargava, A., Bansal, A. Novel coronavirus (COVID-19) diagnosis using computer vision and artificial intelligence techniques: a review. Multimed Tools Appl 80, 19931–19946 (2021). https:// doi.org/ 10.1007/s11042-021-10714-5.

[16] Roniotis, A., Tsiknakis, M. (2018). Detecting Depression Using Voice Signal Extracted by Chatbots: A Feasibility Study. In: Brooks, A., Brooks, E., Vidakis, N. (eds) Interactivity, Game Creation, Design, Learning, and Innovation. ArtsIT DLI 2017 2017. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol 229. Springer, Cham. https://doi. org/10. 1007/ 978-3-319-76908-0_37.