VIEWPOINT

Biological data studies, scale-up the potential with machine learning

Raj Rajeshwar Malinda[1] ✉

European Journal of Human Genetics (2023) 31:619–620; https://doi.org/10.1038/s41431-023-01361-5

MACHINE LEARNING, AN ADVANCED TOOL TO COLLABORATE

Machine learning has emerged as one of the impactful artificial intelligence (AI)-based tools of computational sciences applications. In the recent past, this approach was consistently advancing and broadened the horizon into many interdisciplinary fields, including the biological sciences and medicine. To exhibit the research impact, scientific collaborations among various disciplines have increased with computer sciences, and this trend has also been noticed dramatically in sub-branches of biological sciences including genomics and genetics. Even though, associations with machine learning in biology field are exploring through various dimensions, however, the potential and demand of machine learning have yet to be extensively explored. Surprisingly, this AI-based approach was often overlooked in the field of biological sciences, assuming that, first, most researchers need to be made aware or relatively uninterested in applying rather different approach than traditional in their primary field to address the research questions, perhaps for various reasons. Second, it might be possible that, previously the direct use of machine learning was limited in their research of interest areas compared to others. Whatever the possible reasons, it is interesting to see how biological researchers consider using this method in the near future. As both field, biology and computer science have advanced, the need and/or desire to use machine learning models in unraveling complex biological data interpretations has recently increased, however, it is still limited, and therefore more researchers from both sides are getting interested and involved lately [1].

From the biological perspective, this system is rather complex and dynamic, and has various dimensions to understand, and is somewhat more challenging than it appears. Obliviously, there is none of single approach that can fully discover the complexity of biological phenomena. Therefore, it is requisite that, in combination with other techniques, machine learning methods may be adopted to extract significant relevant information from the biological system. It is somehow misunderstood and often comes into attention that, machine learning methodology seems to overlap with statistics, however, they both have their domains to process the data. Machine learning can be a significantly helpful tool for understanding the behavior of complex data studies genetics and genomic sciences, and interestingly results can be improved over time once the application algorithms have gained experience with sufficient data inputs. The hardcore technical background of machine learning and statistical methods can be reviewed from other sources available [2, 3]. In this opinion-based piece, I discuss about the latest observations, behavior and impact of machine learning in the field of biological sciences, after that, further narrowing down my extended thoughts to the genetics field. I was emphasized to give umbrella overview of recent updates and potential of machine learning applications to scale-up the potential in biological data studies.

UNDERSTANDING OF MACHINE LEARNING IN BIOLOGICAL DATA SYSTEM

As the understanding of a biological system advances, an enormous amount of data is generated on a daily basis. This vast data come from various input sources, for example, imaging data via high-throughput microscopic analysis in cell and developmental biological field and large-scale genomic-wide association studies, and so on [4]. Though manually handling these large numbers of data increases the risk of being biased, inefficient, costly and can produce errored results. In the analysis of genomics datasets, various practical aspects of machine learning algorithms are adopted, for example, analysis of DNA/RNA-binding proteins and other gene regulatory regions. However, the importance and challenges of machine learning in the analytical research of genomics, proteomics, and metabolomics fields are still considered seriously [5].

In general, machine learning algorithms are trained (or learned) with a sufficient amount of known data (labeled or tagged) so that outcomes of the unknown input data from the experiments can be predicted or interpreted. Based on the training module, these algorithm models predict the results as, for example, right or not right, favorable or unfavorable, in the given scenario. There are several number of tasks that can be performed using these trained algorithms. Therefore, I briefly explain the two broad categories of machine learning typically used when analyzing biological data. First, Supervised learning model, in which algorithms are trained with enough labeled datasets, and then used to predict the outcomes of the experiments as explained above. Second, Unsupervised learning model, here algorithms are not developed based on labeled datasets but instead trained to identify the unlabeled parts in the data and hoping to find something new. In such given scenarios, this model can help to discover the potential novel genetic elements in the genomic datasets.

Both computational and biological researchers have recently taken machine learning-based projects together and handshake for more interdisciplinary collaborations [1], therefore, machine

[1]Independent Investigator, Kobe, Hyogo, Japan. ✉email: contact@rrmalinda.com

Received: 5 February 2023 Revised: 31 March 2023 Accepted: 4 April 2023
Published online: 10 April 2023

learning-based approaches are, now, widely used to annotate the functions of several genes. Such an advanced level of work gives strong confidence to discover the potential roles and new other features of the annotated genes in the study; further, it will help to understand the locations and possible structural analysis of the gene in the whole genomic database [6]. In recent times, machine learning applications are aggressively invading in biology and medicine, and undoubtedly, revolutionizing the outcomes with significant and fruitful results around the globe.

Recent evidence suggests, machine learning-based tools are employed in biological studies, and their results have achieved significantly. Some examples are, PlasFlow is designed with machine learning's advanced neural network algorithms and understand the bacterial plasmid sequences from the environmental samples, and as described, the accuracy in identifying the genomic signatures is leveled-up to 96 percent [7]. MetaBCC-LR in metagenomic binning studies, is developed based on k-mer coverage histograms and oligonucleotide composition [8]. Machine learning algorithms are expanding their use to estimate genetic relatedness using mitochondrial DNA (mtDNA) in humans, and this prediction is mainly based on the analysis of hypervariable region I sequences from African, Asian, and Caucasian genetic databases [9]. Because of the nature of this article and limited space, it is, indeed, not possible to list every biological study using machine learning approaches, as mentioned in just a few of them.

CONCLUSION

As widely accepted now, machine learning has great potential to process biological information as the system is adequately trained with data and delivers significant outcomes with less effort. In my opinion, incorporating the computational understanding and experience of machine learning will further enhance the performance and better understanding of complex biological systems. To do this, one way is to provide initial training on computer language programming, statistics, and/or computer science related techniques to all graduate researchers in their initial years of registration at the university. Because researchers can get a fundamental understanding and handling of data-driven research, and, indeed, it will help them to execute the large-scale data if it comes up in their research at any point. In some cases, without sufficient knowledge of machine learning concepts, it is nearly exhausting, inefficient, and biased to manually work on available massive datasets, especially in genomic-wide studies, next-generational sequencing data, medicinal trial studies, and behavioral studies. In addition to using other advanced tools (depending on experiments and methods), machine learning indeed adds up the capabilities to produce more-efficient, high-quality, and timely-mannered results in such those big data studies. If the machine learning algorithms are adequately trained with related information, it will lead to achieving higher accuracy and sensitivity in predicting or interpreting the outcomes. However, such incredibly super-competent algorithm models are still challenging and open questions for researchers in computational biology [5, 6]. I with others, therefore, point out that such algorithm models are developed with the researcher's prior knowledge of biological data; otherwise, incorrect machine learning models would eventually lead to false positive results and substantially provide the wrong information. Machine learning in medicine, genomics, and genetics holds a significant perspective but is more complicated when practical aspects are thoroughly applied. It is noted that collaborations between computational and biological sciences have seen slightest interest in the past; researchers were probably hesitated to initiate significant collaborations possibly because of their understanding of each other's expertise. However, recent trends have seen dramatic increase in collaboration, and desire to expand their knowledge and may be securing fundings from various resources [10]. I, further, believe that researchers from both fields will be more involved in implementing machine learning models rather than just talking a bit in their presentations [1]. The advantage of using machine learning will be more impactful in biological sciences in the near future, and undoubtedly new era will be beginning.

REFERENCES

1. Littmann M, Selig K, Cohen-Lavi L, Frank Y, Honigschimd P, Kataka E, et al. Validity of machine learning in biology and medicine increased through collaborations across fields of expertise. Nat Mach Intell. 2020;2:18–24. https://doi.org/10.1038/s42256-019-0139-8.
2. Mitchell T Machine learning (McGraw-Hill, 1997).
3. Hastie T, Tibshirani R, Friedman J The elements of statistical learning: data mining, inference and prediction (Springer, 2001).
4. Uffelmann E, Huang QQ, Munung NS, De Vries, J, Okada, Y, Martin, A R, et al. Genome-wide association studies. Nat Rev Methods Prim. 2021;1:59.
5. Libbrecht M, Noble W. Machine learning applications in genetics and genomics. Nat Rev Genet 2015;16:321–32. https://doi.org/10.1038/nrg3920.
6. Huang K, Xiao C, Glass LM, Critchlow CW, Gibson G, Sun J. Machine learning applications for therapeutic tasks with genomics data. Patterns. 2021;2:100328. https://doi.org/10.1016/j.patter.2021.100328.
7. Krawczyk P, Lipinski L, Dziembowshi A. PlasFlow: predicting plasmid sequences in metagenomic data using genome signatures. Nucleic Acids Res. 2018;46:e35. https://doi.org/10.1093/nar/gkx1321.
8. Wickramarachchi A, Mallawaarachchi V, Rajan V. MetaBCC-LR: metagenomics binning by coverage and composition for long reads. Bioinformatics. 2020;36 Suppl 1:i3–i11. https://doi.org/10.1093/bioinformatics/btaa441.
9. Govender P, Fashoto SG, Maharaj L, Adeleke MA, Mbunge E, Olamijuwon J, et al. The application of machine learning to predict genetic relatedness using human mtDNA hypervariable region I sequences. PLoS ONE 2022;17:e0263790. https://doi.org/10.1371/journal.pone.0263790.
10. Cechova M. Ten simple rules for biologists initiating a collaboration with computer scientists. PLoS Comput Biol. 2020;16:e1008281. https://doi.org/10.1371/journal.pcbi.1008281.

ACKNOWLEDGEMENTS

Author acknowledges Prof. Hiroaki Kawashima, University of Hyogo, Kobe, Japan for his valuable and constructive feedback on the manuscript.

AUTHOR CONTRIBUTIONS

RRM- conceptualized, drafted, edited and finalized this article.

FUNDING

None.

COMPETING INTERESTS

The author declares no competing interests.

ADDITIONAL INFORMATION

Correspondence and requests for materials should be addressed to Raj Rajeshwar Malinda.

Reprints and permission information is available at http://www.nature.com/reprints

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