Nirma Univ J Pharm Sci; 2021, 8(1&2) 1-16
© 2014, Nirma University, Ahmedabad, Gujarat, India
ISSN 2348 –4012

ARTICLE

THE ROLE OF MACHINE LEARNING IN BIOLOGICAL SCIENCES

Shailvi Shah, Lajja Patel, Tejal Mehta*
Department of Pharmaceutics, Institute of Pharmacy, Nirma University,
Ahmedabad, Gujarat, India-382481.
**Corresponding author:*
Dr. Tejal A. Mehta
E-mail: tjshah3@gmail.com; tejal.shah@nirmauni.ac.in

ABSTRACT

In the biological sciences, machine learning (ML) has become an essential technology that is revolutionizing research methods and speeding up discoveries in a variety of fields. A thorough overview of the various uses of ML in biological sciences is discussed in this article, including drug development, protein sciences, vaccines, biosystems, and computational biology. ML models facilitate the rapid discovery of innovative drug candidates with decreased side effects and increased efficacy, hence speeding up the drug development pipeline by utilizing large-scale biological data. ML techniques are improving the prediction of protein interactions, structures, and functions in the field of protein sciences. The design of vaccines, epitope prediction, and antigen selection are all greatly aided by ML techniques. ML models evaluate genetic and proteomic data based on individual immune responses, facilitating the generation of personalized immunizations that are optimal for immunogenicity and vaccine efficacy. Furthermore, by replicating cellular processes, modeling intricate biological networks, and forecasting gene regulatory mechanisms, ML techniques are revolutionizing the study of biosystems. In computational biology, ML is utilized for phenotypic prediction, gene expression profiling, and sequence analysis. ML models facilitate the development of precision medicine techniques, the characterization of medication response patterns, and the identification of disease biomarkers by combining multi-omics data. To fully explore the potential of ML for tackling significant issues in healthcare, computer scientists, biologists, and bioinformaticians must *collaborate across disciplinary boundaries. This review emphasizes the revolutionary impact of ML on biological sciences.*

Keywords: *Machine learning, artificial intelligence, biological sciences.*

INTRODUCTION

Machine learning (ML) is the evaluation of algorithms that allow pattern classification, recognition, and prediction based on models created from readily available data. It is important to acknowledge two features of automation when considering ML in a broad sense. First, the categorization and prediction duties ought to be accomplished by a computer system with the proper programming. That is, ML produces a system for classification that might be used practically with existing technology. Moreover, the classification development process is designed to be highly automated with little need for human intervention [1].

ML addresses the problem of building computers that learn automatically from use. Because it lies at the nexus of data science and artificial intelligence (AI), as well as statistics and computer science, it is one of the technical fields with the fastest development rates in the present day. Recent advances in ML can be attributed to the evaluation of new learning algorithms and theories, the continuous development of low-cost processing, and internet data accessibility. The use of data-intensive ML techniques by technology, science, and business has led to a rise in the application of evidence-based decision-making across a range of industries, including financial modeling, manufacturing, marketing, and healthcare. In the previous ten years, there has been a notable growth in the capacity of mobile and networked computing systems to gather and transfer massive amounts of data; this development is commonly referred to as "Big Data." The researchers and engineers who collect these data mostly utilize ML to address the challenge of obtaining meaningful conclusions, producing accurate projections, and making decisions based on such data sets[2]. A few examples of how historical data are used are historical crime statistics, which are used to help distribute police officers to specific places at certain times, and historical traffic data, which are used to enhance traffic control and minimize congestion. Massive experimental data sets are gathered and vetted to progress the fields of neurology, astronomy, biology, and various other data-intensive empirical sciences [3]. The development of machine-learning tools to assess high throughput experimental data in new ways resulted in a correspondingly broad range of effects in the empirical sciences, including social science, cosmology, and biology. Engineers and data scientists are driven by biology, but they also focus on developing ML methods that solve practical problems. Consequently, models constructed using these techniques often neglect acknowledged physiological constraints. Any biological model is an abstraction, one could say, and even while it doesn't perfectly capture every facet of the living thing, it can still be helpful [4].

Biology is changing to a data-rich field due to the development of high throughput sequencing and “omics” technologies, as well as the consequent exponential increase in the number of measurements of structure, macromolecular sequence, and gene expression. Like physics was before Leibniz and Newton, biology has mostly remained an informative science despite these advancements. ML presently offers some of the most affordable technologies accessible for creating prediction models from biological data. These include techniques for finding genetic markers for diseases, predicting the function of macromolecules, finding functionally important protein sites, classifying new genomic sequences, and figuring out the network of genetic relationships that control important biological processes.

Biology must continue to improve its machine learning (ML) methods to become an engineering discipline. Examples of these methods include learning from highly imbalanced information sets, learning complex structures of class labels (such as labels associated by acyclic graphs that are directed instead of one of multiple mutually exclusive labels), and learning from richly structured data, such as macromolecular DNA sequences along with 3-dimensional molecular structures.[5]

Large-scale genome pattern projects have come out in the availability of hundreds of full genome sequences. More importantly, every 18 months, the amount of nucleic acid patterns in the GenBank database doubles. The structural genomics efforts have also caused a rise in the lot of macromolecular (like proteins) structures. Biologists can currently access over a thousand databases. The advent of high-throughput “omics” techniques, such as those for analyzing the expression of hundreds of genes in response to various perturbations has made system-wide measurements of biological variables conceivable. ML is therefore making discoveries in the biological sciences possible to a greater extent.

A few instances of ML applications in computational and systems biology are as follows: determining a primary (amino acid) sequence of protein, structure, and interaction partners to predict its function(s); identifying the amino acid pattern and if possible its structure to determine functionally significant sites (such as protein-protein, protein-RNA, protein-DNA, and post-translational modification sites); assigning structural classifications to protein sequences (and structures); finding functional modules—groups of genes that act together—and genetic networks using information on gene expression.

ARTIFICIAL INTELLIGENCE REVOLUTION

Recently, deep (multi-layered) neural networks have gained significant prominence in the ML industry. These networks comprise algorithms that are built into multiple (several to hundreds) functional layers and are roughly modeled after the connectivity of a brain of human [6]. Numerous articles in nearly every branch of science explain how deep learning (DL) could be applicable to address any problem for which sufficient data is provided. There are a lot of reasons behind this. First, the technique itself has become simple to use by a reasonably competent programmer with the accessibility of software that has made it accessible to anyone to try out DL experiments. Only a few years ago, even experienced computer scientists would have found it difficult to carry out these experiments.

The use of inexpensive graphics processors that significantly speed up ML is one example where hardware advancements have also played a significant role. Today, one need not even purchase the hardware because it can be used, sometimes for free, through a variety of cloud services. Lastly, there are a ton more training options available, which has led to an enormous rise in job opportunities for AI [7].

While biology has long employed ML techniques, especially neural networks, there has been a notable upsurge in interest in the field recently [8]. The urge to take on problems whose solutions could enhance the health and happiness of millions of people is strong. The conventional academic publication mechanisms in the biomedical sciences are challenged by the widespread popularization of computer-led biological data science [9]. The problem lies in the fact that many of the papers emerging from AI development are not contributing to the field because these methods are not being applied properly. Often, this means that either their experimental design is flawed or they do not offer any advancement over currently used methods [10]. Table 1 shows examples in which AI is used for prediction modeling.

Table 1 Disease detection and prediction modeling with the use of AI in clinical data modality

Diseases	Algorithm	Outcomes	Reference
AMD	ML-based predictive model	The model was highly reliable to predict AMD progression	[11]
COVID-19	Passive Aggressive (PA)	70–80% accuracy achieved	[4]
Alzheimer’s disease	Random Forest, Shapley additive explanations (SHAP)	Accuracy of 93.95% in first layer and 87.08% in second layer was predicted	[12]
Ovarian cancer	Artificial Neural Network	For survival - 93% accuracy In surgical outcomes - 77% accuracy	[13]
Pulmonary cancer	Lung Cancer Prediction-Convolutional Neural Network, Brock model	Compared to the LCP-CNN model was capable to foresee the malignancy of lung nodules with greater accuracy and fewer false-negative outcomes	[14]
Influenza	Innovative Accessible Technology-Back Progression in Neural Network (IAT-BPNN)	For a large population, IAT-BPNN demonstrated high accuracy in predicting influenza-like illness.	[15]

PROBLEM SUITABILITY AND CURRENT DEVELOPMENTS

Initially, one must acknowledge that the efficacy of DL methods has been limited to applications that meet specific data requirements. These requirements include plenty of data, high dimensionality (a sample comprising numerous variables), and well-structuredness (a reference to a graphical interaction among the variables). The images are the perfect sample for DL techniques since they provide a huge number of variables (pixels) that can be precisely categorized into well-defined objects (e.g., the pixels that make up a nose on a face). Text and audio data can also be used. Naturally, a lot of biological data sets—such as text data from sequencing or picture data from microscopes—also satisfy these requirements.

Other, possibly less obvious uses for example, current developments in DL have greatly enhanced our ability to figure out the tertiary structure of proteins from their amino acid sequences. This has been made possible by the ability to view protein folds as 2D maps of interatomic distances that may be analyzed like that of pictures. However, not all biological data sets lend themselves to DL analysis. One of the examples is the analysis of single nucleotide polymorphisms (SNPs) in genomic data. The existence or absence of known SNPs in a data genome is highly dimensional given the millions of SNPs that are known, although the data remain unstructured.

This kind of information is known as categorical data, and it can only be displayed as tables with an unstructured row order. Although SNP data can still be categorized by tagging them with a specific gene or chromosome, this is insufficient to make them suitable for DL, and other kinds of analysis are probably more useful. Due to this limitation on the data applicability, it is crucial to remember that state-of-the-art status should never be derived from its use of DL in an evaluation. Rather, it needs to begin with the contrary assumption and rely on the research presented in the paper to convince the reviewers of the opposite. [10][16].

MACHINE LEARNING IN BIOLOGICAL SCIENCES

ML in Drug Development

The process of developing new medications requires a long time and investment. Indeed, potential medications must go through a rigorous and competitive process to ensure both patient safety and medical efficacy. Phase 0 to Phase IV are the four main stages that comprise drug development. According to the literature, automating certain significant but monotonous data processing and analysis tasks, particularly with robotics and ML techniques is a rather cheap solution. Certainly, many fruitful collaborations have emerged between AI/ML and companies, universities, research centers, and pharmaceutical laboratories gradually decreasing the gap in bioinformatics between applied mathematics, computer sciences, and biology. This would facilitate the drug development pipelines to go more quickly because they could be carried out computationally, autonomously, and with a lower risk of human error [17]. By providing novel approaches to enduring problems, ML is a key factor in transforming the drug development process.

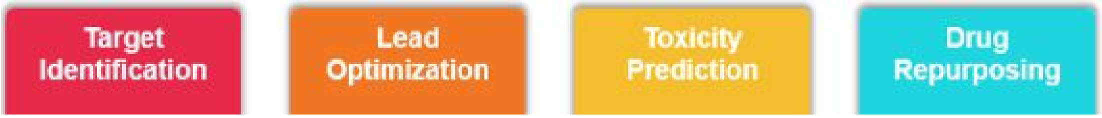

Fig. 1 ML in drug development

Fig. 1 lists the various stages where ML is used in drug development. ML algorithms aid in target identification during the early phases of drug discovery by analyzing large biological datasets to identify targets linked to disease and rank possible therapeutic candidates. Then, by precisely predicting structure-activity relationships (SAR), ML algorithms facilitate lead optimization by accelerating the identification of molecules with favorable pharmacokinetic characteristics. Moreover, ML algorithms improve medication safety profiles by predicting adverse drug responses (ADRs) and identifying off-target effects, which helps with toxicity prediction. Also, ML-driven drug repurposing techniques find new therapeutic indications for authorized medications by utilizing large-scale data integration and current biological knowledge, greatly cutting down on the time and expense of conventional drug discovery procedures. Additionally, by customizing treatment plans to fit each patient's unique profile, anticipating a patient's reaction to a particular therapy, and streamlining treatment schedules, ML facilitates precision medicine approaches. By way of publicly funded programs such as the US National Institutes of Health's Molecular Libraries Screening Centres, these technologies greatly boost the rate and volume of information that can be obtained regarding the influence of chemical compounds, paving the way for the development of massive databases such as PubChem. These databases frequently comprise scores for many compounds on an assay, which represent the outcomes of multiple screens, in addition to details on the targets of particular assays. To evaluate the influence of perturbagens on certain molecular targets and cell behaviors, high content screening and high throughput microscopy are widely used techniques. Classifier training to identify predicted patterns and feature computation to define visual elements are common analysis techniques for high-content displays. One method involves employing clustering algorithms, which don't necessitate prior knowledge of the patterns, to determine compounds with comparable biological effects. Machine-vision techniques, which can extract more accurate data, could be applied to high-content assays. Pattern-unmixing techniques try to cope with the continuous nature of relocation events by calculating the proportion of a target that exists in each subcellular place [18]. Ultimately, there is a lot of potential for the use of ML in drug development to hasten the identification of safe and effective treatments, bringing about the era of precision medicine and better patient outcomes. On the other hand, the integration of multi-view data may allow for the deployment or improvement of precision medicine procedures, which could reduce the cost and time associated with drug discovery while simultaneously making medicines more patient-oriented.

ML in Protein Sciences

Essentially, intermolecular interactions are necessary for proteins to create complexes that regulate biological functions in living cells. Such interactions between proteins that occur in a wide range of dynamic conformational states and levels of inherent disorder are becoming more recognized. Furthermore, the structures' size varies from tiny dynamic biomolecular condensates of 100 nm or more to simple binary complexes. Some of the major issues in molecular biosciences include how such interactions are governed, how they arise, how they influence function, and what happens when they occur inadvertently and cause disease [19]. ML is revolutionizing protein sciences by offering powerful computational tools to predict the actions of proteins shown in Fig.2. ML algorithms excel in predicting protein structures, functions, and interactions, thereby accelerating the pace of biological research (8). By using massive repositories of known protein structures, ML methods predict protein structures accurately from amino acid sequences, offering important information on the folding patterns and functional characteristics of the resulting proteins (9).

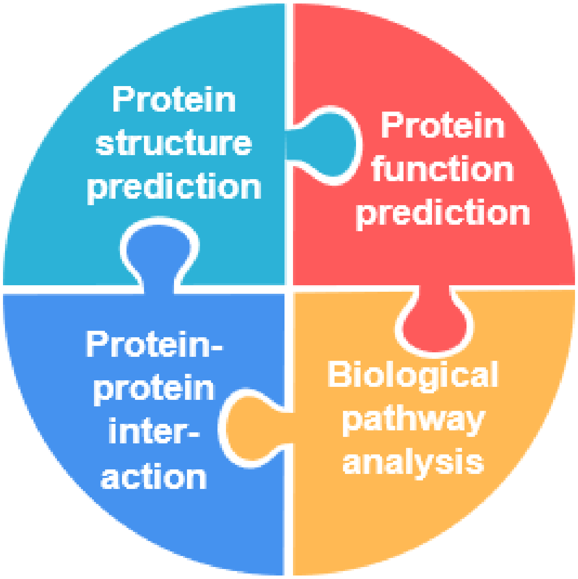
Protein structure prediction
Protein function prediction
Protein-protein inter-action
Biological pathway analysis

Fig.2 ML in protein sciences

Predictions of structural attributes of many proteins, like those related to solvent accessibility, secondary structure, disordered regions, binding sites, functional sites, protein domain boundaries, and transmembrane helices, are 1-D prediction problems. Basic correlation techniques can achieve accuracy levels substantially above random fluctuations, with up to 50% accuracy and a specific amount of data captured. The development techniques have been improved upon. To further enhance secondary structure prediction, PSI-PRED, for example, leverages PSI-BLAST to generate additional profiles based on position-specific score matrices. In an attempt to increase the reliability of prediction by integrating data that exceeds the fixed-size window input of conventional feedforward neural networks, more complex recursive neural network architectures have been created through new algorithmic advancements motivated by the theory of probabilistic graphical models. Numerous neural network ensembles in size have also been employed. Secondary structure prediction accuracy has increased to between 78% and 80% because of new technologies and an increase in protein sequence databases used to create profiles. Furthermore, to enhance secondary structure prediction, hybrid techniques that incorporate homology searches and neural network approaches have been created [20]. Many facets of predicting the 3D structures of proteins, including the development and assessment of the fold recognition model, have been addressed using ML techniques. Fold recognition aims to identify a known protein with a structure believed to be similar to that of an unknown protein. The most efficient methods for predicting 3D structures using templates necessitate the initial identification of structural homologs. First, neural networks were combined with threading for this task. Recently, a universal ML framework based on pairwise similarity characteristics between query and template proteins has been suggested to enhance fold recognition sensitivity and specificity. The framework can be expanded to any other supervised learning technique, even though its current implementation employs support vector machines to find folds. Moreover, by analyzing structural properties, evolutionary conservation patterns, and sequence motifs, ML approaches allow the prediction of protein activities. This makes it easier to identify new drug targets and annotate proteins whose functions are unknown. Additionally, ML-driven approaches play a crucial role in predicting protein-protein interactions [22,22] (shown in Fig. 3), elucidating the intricate networks of molecular interactions underlying various biological processes. Researchers can obtain more detailed knowledge of disease mechanisms, signaling pathways, and protein dynamics by combining multi-omics data and utilizing modern ML techniques [23]. This can facilitate the development of innovative therapies and precision medicine tactics. In general, applying ML to the field of protein sciences has great potential to improve our knowledge of biological systems and hasten the development of novel therapeutics for human illnesses.

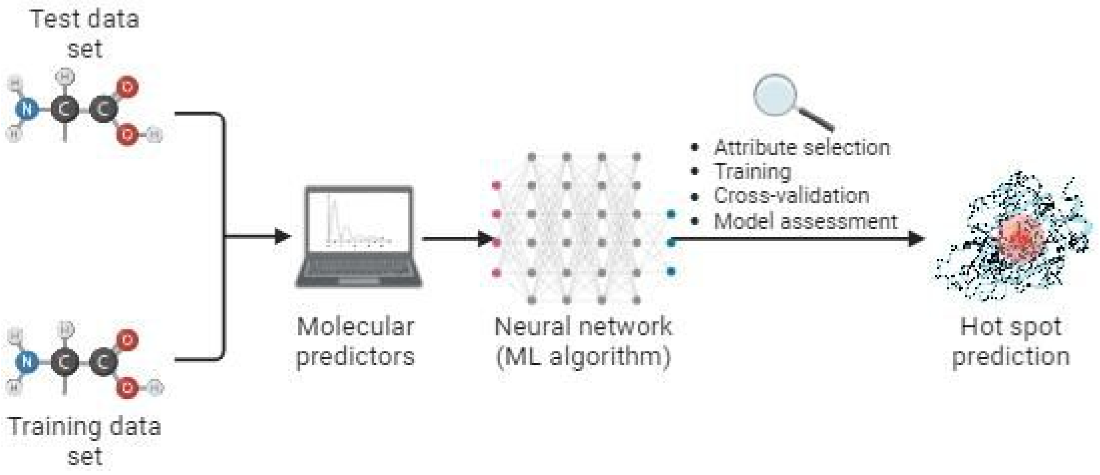

Fig.3 Hotspot prediction in protein-protein interactions

Improving Reverse Vaccinology

ML is increasingly being utilized in various aspects of vaccine development, from antigen selection to vaccine design and optimization as shown in Fig.4.

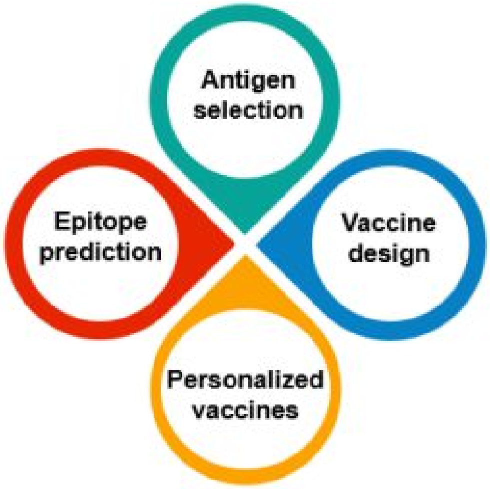

Fig.4 Use of machine learning in vaccine development

Reverse vaccinology (RV), a technique that utilizes bioinformatic algorithms to find putative protective antigens in bacterial proteomes that could serve as prospective vaccine candidates, has gained popularity during the past ten years. This rise has been fueled by the quick collection of information from the full genome sequencing of more than 4000 bacteria, which is utilized to pinpoint the protein-coding genes that serve as the foundation for RV methods. Since there is no need to cultivate bacteria, promising vaccine candidates can be identified quickly and at a lower cost using RV than with conventional vaccinology. Additionally, RV can discover all of a bacterial species' putative protective protein antigens rather than only the most prevalent antigens that are conventionally extracted from bacterial cultures [24]. The most reliable strategy to address the urge for future vaccinology has been described as using recombinant proteins as immunogens. RV has been used to limit the number of vaccine candidates whose immunogenicity is tested in experimental animals before vaccine development and human trials. Comparing subunit vaccines to live vaccines, which have issues with attenuation and reversion to virulence, the benefits encompass improved safety, decreased cost, minimized competition between antigens, and targeted delivery to infectious sites. Delivery methods for subunit vaccinations that align with the RV-predicted antigens include non-pathogenic vectors, DNA vaccines, and non-living antigen delivery systems. The process of choosing potential antigens for immunogenicity tests and the creation of subunit vaccines can be expedited quickly using RV methods [25]. The first supervised, web-based technique for predicting bacterial protective antigens (BPAgs) is Vaxign [26]. Fig.5 shows the use of ML in vaccine development [27]. Overall, ML holds promise for revolutionizing vaccine development by facilitating antigen discovery, epitope prediction, vaccine design, immunogenicity assessment, and personalized vaccine strategies, thereby advancing efforts to combat infectious diseases and emerging pathogens [28].

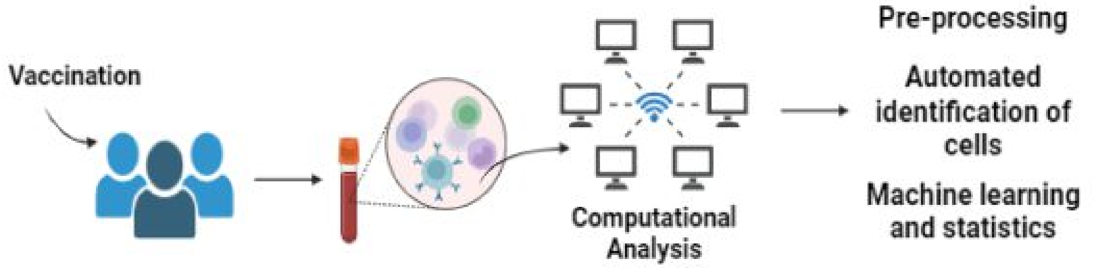
Vaccination
Computational Analysis
Pre-processing
Automated identification of cells
Machine learning and statistics

Fig.5 Computational analysis of vaccination data

Biosystems Design by ML

More and more research is being done on biotechnological applications of biosystems, comprising enzymes, pathways, and entire cells. However, because of the enormous interdependence of biosystems, it is very challenging to design biosystems with the necessary characteristics. The development of high-throughput phenotyping technology has made ML an efficient method for predicting biological system behavior and enabling the learning phase [29]. Designing biosystems with ML techniques integrated has become possible since high throughput technologies like as-omics have developed quickly. By identifying new candidates for the best performance, ML models can improve biosystem design applications by identifying patterns in complicated biological data at various scales of investigation [30]. ML is being used throughout the entire design process for biosystems to find novel technical solutions with fewer variations in design. By seeing trends and patterns in systems with a lot of data, ML has become a promising tool for accelerating success in biosystems design. We want to further close any gaps between computer science and biology researchers because effective research in this field requires their cooperation. The biological audience will gain from a brief explanation of the fundamentals of ML required to comprehend the technical aspects of pertinent work and understand the advantages and disadvantages of ML approaches and the potential applications of ML in other fields [31]. The objective of ML from a computing perspective is to create a function that can convert a specific instance of input data into a desired output. The ML paradigm assumes that the values of input and results of the training data are associated. By using ML algorithms to identify these correlation patterns, instances of unknown input data from the training set can be processed to get the intended output value. The more training data that is provided, the more accurate the learned function will be. This function, which is usually referred to as a model since it may be considered as establishing a model of the data being investigated, can be created in a variety of ways, which is why ML is closely related to statistical models [32].

ML in Computational Biology

Important biological elements of the process of controlling gene expression are transcription factors. To precisely control the spatiotemporal regulation of genes, transcription factors or chromatin regulators bind to specific DNA sequences at specified locations known as transcription factor binding sites. Transcription factors are ubiquitous proteins that are essential for many biological processes. It is vital to anticipate the function and structure of proteins effectively given the rise in protein sequences [33]. Currently, techniques for predicting protein-DNA binding sites are based on both DL and conventional ML algorithms. To anticipate protein-DNA binding sites in the early stages, we typically employ a development method based on a conventional ML algorithm. DL-based techniques for predicting protein-DNA binding sites from sequence data have been incredibly successful in recent years [34]. The function of DNA-binding proteins can be predicted using a variety of statistical and ML techniques, and these techniques are constantly being refined. Convolutional neural networks (CNN), recursive neural networks (RNN), and combined neural networks based on CNN-RNN can be used to classify the current state of deep-learning techniques for protein-DNA-binding site prediction [35].

CHALLENGES AND OPPORTUNITIES

One significant challenge of ML is data integration and quality. ML models heavily rely on large and diverse datasets, and integrating disparate sources of data while ensuring data quality and consistency can be complex and time-consuming. Additionally, maintaining data privacy and security remains a critical concern, particularly in sensitive domains like healthcare. Another challenge is the interpretability and transparency of ML models. As ML algorithms become increasingly complex, understanding how they arrive at decisions or predictions can be difficult. Furthermore, ethical considerations such as bias and fairness in algorithms, unintended consequences of automated decision-making, and the ethical use of data are central concerns. Despite these challenges, ML also presents numerous opportunities. ML algorithms have the potential to uncover patterns and insights in data that humans may overlook, leading to breakthroughs in areas such as drug discovery, disease diagnosis, and personalized medicine. Moreover, ML enables automation and optimization of processes, improving efficiency and productivity across industries. By tackling issues related to data integration, interpretability, and ethics, we can harness the full benefits of ML while minimizing risks and ensuring equitable and ethical outcomes for society as a whole.

SUMMARY AND IMPLICATIONS FOR THE FUTURE

The review article delves into the transformative impact of ML on biological sciences. It highlights ML's pivotal role across various domains, including drug development, protein sciences, vaccine design, biosystems, and computational biology. ML techniques analyze vast biological datasets to identify novel drug targets, predict protein structures and functions, design personalized vaccines, model biological networks, and integrate multi-omics data for diagnosis and treating illness. The future of ML in biological sciences holds huge potential to improve our knowledge of diverse biological systems and address key challenges in healthcare. By continuing to innovate in algorithm development, data integration, and model interpretability, researchers can unlock new insights into disease mechanisms, develop more effective therapies, and optimize bioproduction processes. By fostering interdisciplinary partnerships and promoting accessibility to data and computational resources, ML promises to revolutionize biological sciences, ushering in a new era of innovation and discovery for the betterment of society.

REFERENCES

1. Tarca, A.L., Carey, V.J., Chen, X. wen, Romero, R., and DrÎghici, S. (2007) Machine learning and its applications to biology. *PLoS computational biology*, **3**.

2. Majaj, N.J., and Pelli, D.G. (2018) Deep learning-Using machine learning to study biological vision. *Journal of Vision*, **18** (13), 1–13.

3. Torrisi, M., Pollastri, G., and Le, Q. (2020) Deep learning methods in protein structure prediction. *Computational and Structural Biotechnology Journal*, **18**, 1301–1310.

4. Alkady, W., ElBahnasy, K., Leiva, V., and Gad, W. (2022) Classifying COVID-19 based on amino acids encoding with machine learning algorithms. *Chemometrics and Intelligent Laboratory Systems*, **224**, 104535.

5. Gilpin, W., Huang, Y., and Forger, D.B. (2020) Learning dynamics from large biological data sets: Machine learning meets systems biology. *Current Opinion in Systems Biology*, **22**, 1–7.

6. Ghosh, S., and Dasgupta, R. (2022) *Machine Learning in Biological Sciences: Updates and Future Prospects*.

7. Xu, Y., Liu, X., Cao, X., Huang, C., Liu, E., Qian, S., Liu, X., Wu, Y., Dong, F., Qiu, C.W., Qiu, J., Hua, K., Su, W., Wu, J., Xu, H., Han, Y., Fu, C., Yin, Z., Liu, M., Roepman, R., Dietmann, S., Virta, M., Kengara, F., Zhang, Z., Zhang, L., Zhao, T., Dai, J., Yang, J., Lan, L., Luo, M., Liu, Z., An, T., Zhang, B., He, X., Cong, S., Liu, X., Zhang, W., Lewis, J.P., Tiedje, J.M., Wang, Q., An, Z., Wang, F., Zhang, L., Huang, T., Lu, C., Cai, Z., Wang, F., and Zhang, J. (2021) Artificial intelligence: A powerful paradigm for scientific research. *Innovation*, **2**.

8. Dutta, U., Babu, N.D., and Setlur, G.S. (2022) Artificial Intelligence in Biological Sciences: A Brief Overview, in *Information Retrieval in Bioinformatics: A Practical Approach*, pp. 19–35.

9. Bhardwaj, A., Kishore, S., and Pandey, D.K. (2022) Artificial Intelligence in Biological Sciences. *Life*, **12**.

10. Jones, D.T. (2019) Setting the standards for machine learning in biology. *Nature Reviews Molecular Cell Biology*, **20** (11), 659–660.

11. Schmidt-Erfurth, U., Waldstein, S.M., Klimscha, S., Sadeghipour, A., Hu, X., Gerendas, B.S., Osborne, A., and Bogunović, H. (2018) Prediction of individual disease conversion in early AMD using artificial intelligence. *Investigative Ophthalmology and Visual Science*, **59** (8), 3199–3208.

12. El-Sappagh, S., Alonso, J.M., Islam, S.M.R., Sultan, A.M., and Kwak, K.S. (2021) A multilayer multimodal detection and prediction model based on explainable artificial intelligence for Alzheimer’s disease. *Scientific Reports*, **11** (1), 1–26.

13. Enshaei, A., Robson, C.N., and Edmondson, R.J. (2015) Artificial Intelligence Systems as Prognostic and Predictive Tools in Ovarian Cancer. *Annals of Surgical Oncology*, **22** (12), 3970–3975.

14. Baldwin, D.R., Gustafson, J., Pickup, L., Arteta, C., Novotny, P., Declerck, J., Kadir, T., Figueiras, C., Sterba, A., Exell, A., Potesil, V., Holland, P., Spence, H., Clubley, A., O'Dowd, E., Clark, M., Ashford-Turner, V., Callister, M.E.J., and Gleeson, F. V. (2020) External validation of a convolutional neural network artificial intelligence tool to predict malignancy in pulmonary nodules. *Thorax*, **75** (4), 306–312.

15. Hu, H., Wang, H., Wang, F., Langley, D., Avram, A., and Liu, M. (2018) Prediction of influenza-like illness based on the improved artificial tree algorithm and artificial neural network. *Scientific Reports*, **8** (1), 1–8.

16. Badar, M.S. (2023) A Guide to Applied Machine Learning for Biologists. *A Guide to Applied Machine Learning for Biologists*, C1.

17. Réda, C., Kaufmann, E., and Delahaye-Duriez, A. (2020) Machine learning applications in drug development. *Computational and Structural Biotechnology Journal*, **18**, 241–252.

18. Murphy, R.F. (2011) An active role for machine learning in drug development. *Nature Chemical Biology*, **7** (6), 327–330.

19. Rauer, C., Sen, N., Waman, V.P., Abbasian, M., and Orengo, C.A. (2021) Computational approaches to predict protein functional families and functional sites. *Current Opinion in Structural Biology*, **70**, 108–122.

20. Cheng, J., Tegge, A.N., and Baldi, P. (2008) Machine Learning Methods for Protein Structure Prediction. *IEEE Reviews in Biomedical Engineering*, **1**, 41–49.

21. Liu, Q., Chen, P., Wang, B., Zhang, J., and Li, J. (2018) Hot spot prediction in protein-protein interactions by an ensemble system. *BMC Systems Biology*,

22. Elia Venanzi, N.A., Basciu, A., Vargiu, A.V., Kiparissides, A., Dalby, P.A., and Dikicioglu, D. (2023) Machine Learning Integrating Protein Structure, Sequence, and Dynamics to Predict the Enzyme Activity of Bovine Enterokinase Variants. *Journal of Chemical Information and Modeling*.

24. Ong, E., and He, Y. (2022) Vaccine Design by Reverse Vaccinology and Machine Learning. *Methods in Molecular Biology*, **2414**, 1–16.

25. Bowman, B.N., McAdam, P.R., Vivona, S., Zhang, J.X., Luong, T.,

Belew, R.K., Sahota, H., Guiney, D., Valafar, F., Fierer, J., and Woelk, C.H. (2011) Improving reverse vaccinology with a machine learning approach. *Vaccine*, **29** (45), 8156–8164.

26. Ong, E., Wang, H., Wong, M.U., Seetharaman, M., Valdez, N., and He, Y. (2020) Vaxign-ML: Supervised machine learning reverse vaccinology model for improved prediction of bacterial protective antigens. *Bioinformatics*, **36** (10), 3185–3191.

27. Lucchesi, S., Furini, S., Medaglini, D., and Ciabattini, A. (2020) From bivariate to multivariate analysis of cytometric data: Overview of computational methods and their application in vaccination studies. *Vaccines*, **8** (1).

28. Ong, E., Wong, M.U., Huffman, A., and He, Y. (2020) COVID-19 Coronavirus Vaccine Design Using Reverse Vaccinology and Machine Learning. *Frontiers in immunology*, **11**, 1581.

29. Radivojević, T., Costello, Z., Workman, K., and Garcia Martin, H. (2020) A machine learning Automated Recommendation Tool for synthetic biology. *Nature Communications*, **11** (1).

30. Helleckes, L.M., Hemmerich, J., Wiechert, W., von Lieres, E., and Grünberger, A. (2023) Machine learning in bioprocess development: from promise to practice. *Trends in Biotechnology*, **41** (6), 817–835.

31. HamediRad, M., Chao, R., Weisberg, S., Lian, J., Sinha, S., and Zhao, H. (2019) Towards a fully automated algorithm driven platform for biosystems design. *Nature Communications*, **10** (1).

32. Volk, M.J., Lourentzou, I., Mishra, S., Vo, L.T., Zhai, C., and Zhao, H. (2020) Biosystems Design by Machine Learning. *ACS Synthetic Biology*, **9** (7), 1514–1533.

33. Chicco, D. (2017) Ten quick tips for machine learning in computational biology. *BioData Mining*, **10** (1).

34. Liu, C., Ma, Y., Zhao, J., Nussinov, R., Zhang, Y.C., Cheng, F., and Zhang, Z.K. (2020) Computational network biology: Data, models, and applications. *Physics Reports*, **846**, 1–66.

35. Zhang, Y., Bao, W., Cao, Y., Cong, H., Chen, B., and Chen, Y. (2022) A survey on protein–DNA-binding sites in computational biology. *Briefings in Functional Genomics*, **21** (5), 357–375.