Turkish Journal of Biology

Volume 47 | Number 6

Article 3

12-28-2023

Deep learning in bioinformatics

Malik Yousef
malik.yousef@gmail.com

Jens ALLMER
jens.allmer@hs-ruhrwest.de

Recommended Citation
Yousef, Malik and ALLMER, Jens (2023) "Deep learning in bioinformatics," *Turkish Journal of Biology*: Vol. 47: No. 6, Article 3. https://doi.org/10.55730/1300-0152.2671
Available at: https://journals.tubitak.gov.tr/biology/vol47/iss6/3

Turkish Journal of Biology

http://journals.tubitak.gov.tr/biology/

Research Article

Turk J Biol
(2023) 47: 366-382
© TÜBİTAK
doi:10.55730/1300-0152.2671

Deep learning in bioinformatics

Malik YOUSEF[1], **Jens ALLMER**[2,*]

[1]Department of Information Systems, Zefat Academic College, Zefat, Israel
[2]Medical Informatics and Bioinformatics, Institute for Measurement Engineering and Sensor Technology, Hochschule Ruhr West, University of Applied Sciences, Mülheim an der Ruhr, Germany

Received: 20.09.2023 • **Accepted/Published Online:** 18.12.2023 • **Final Version:** 28.12.2023

Abstract: Deep learning is a powerful machine learning technique that can learn from large amounts of data using multiple layers of artificial neural networks. This paper reviews some applications of deep learning in bioinformatics, a field that deals with analyzing and interpreting biological data. We first introduce the basic concepts of deep learning and then survey the recent advances and challenges of applying deep learning to various bioinformatics problems, such as genome sequencing, gene expression analysis, protein structure prediction, drug discovery, and disease diagnosis. We also discuss future directions and opportunities for deep learning in bioinformatics. We aim to provide an overview of deep learning so that bioinformaticians applying deep learning models can consider all critical technical and ethical aspects. Thus, our target audience is biomedical informatics researchers who use deep learning models for inference. This review will inspire more bioinformatics researchers to adopt deep-learning methods for their research questions while considering fairness, potential biases, explainability, and accountability.

Key words: Deep learning, bioinformatics, neural networks, biological data analysis

1. Introduction

Biology has become a data-driven science, for example, due to next-generation sequencing and mass spectrometry (Schadt et al. 2010). Precision medicine will depend on big data because patient stratification needs to be informed, which is best done through omics analyses (Twilt, 2016; Hulsen et al., 2019). Data analysis in these fields is performed by bioinformatics and medical informatics (Chen et al., 2017). Big data, however, is not enough; it needs to be turned into information and knowledge (Mahoto et al., 2021). Therefore, the following sections explore the intersections of information science, artificial intelligence, and other fields (Figure 1).

Information science is a field of study concerned with collecting, organizing, analyzing, interpreting, and disseminating information (Seadle and Havelka, 2023). It encompasses a broad range of topics, including library science, information management, and data science, and has applications in fields such as education, business, and government (Cervone, 2016).

Artificial intelligence (AI) is the simulation of human intelligence in machines designed to think and act like humans (Korteling et al., 2021). AI systems can be trained to perform various tasks, such as image and speech recognition, decision-making, and language translation (Litjens et al., 2017). AI has become increasingly prevalent in society, with applications ranging from healthcare to autonomous vehicles (Bohr and Memarzadeh, 2020). Furthermore, AI has been used to develop novel algorithms and insights that have been used to improve many existing processes (Cordero et al., 2019; Tran et al., 2019).

Machine learning (ML) is a subfield of AI that involves the development of algorithms and statistical models that enable computers to "learn" from data (Yazdani et al., 2023). Machine learning aims to allow computers to make predictions or take actions based on input data without being explicitly programmed to do so (Baştanlar and Ozuysal, 2014). In bioinformatics, machine learning has been used to make predictions regarding the protein structure (AlQuraishi, 2021), protein-protein interactions (Sarkar and Saha 2019), gene expression (Al taweraqi and King, 2022), and disease diagnosis (Ahsan et al., 2022).

Deep learning (DL) is a subfield of machine learning based on artificial neural networks inspired by the structure and function of the human brain (Lepakshi, 2022). Deep-learning algorithms consist of multiple layers of interconnected nodes, each performing a specific computation on the data it receives (LeCun et al., 2015).

* Correspondence: jens@allmer.de

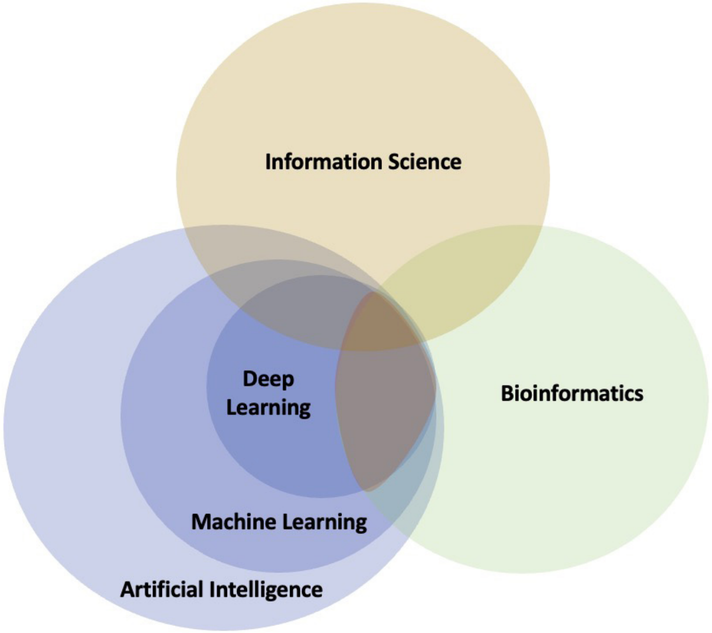

Figure 1. Relationships among bioinformatics, information science, artificial intelligence, machine learning, and deep learning. At the intersection of all circles (orange) is the application of AI, ML, and DL in other areas, such as bioinformatics.

The use of deep learning and the availability of enormous computing resources has revolutionized the field of AI, allowing computers to achieve human-like accuracy in tasks such as image and speech recognition (Hosny et al., 2018).

Deep-learning algorithms have proven highly effective in various tasks, such as image and speech recognition, natural language processing, and autonomous decision-making. For example, deep-learning algorithms have been used to classify images from the ImageNet dataset field accurately (Yang et al., 2020). Deep-learning algorithms have been used in speech recognition to achieve near-human-level accuracy (Litjens et al., 2017). Natural language processing also uses deep-learning algorithms to generate human-like text (Davenport and Kalakota, 2019). In autonomous decision-making, deep-learning algorithms have been used to control robotic agents (Sarker et al., 2021). The ability of deep-learning algorithms to automatically learn complex representations from raw data has led to remarkable breakthroughs in artificial intelligence (Alzubaidi et al., 2021).

In summary, although all these fields are related, they differ in focus and scope. Information Science deals with managing and disseminating information, whereas AI and ML focus on developing algorithms that enable computers to simulate human intelligence. Deep learning is a specific machine learning approach based on artificial neural networks.

In this study, we discuss some arbitrarily selected DL applications in bioinformatics to showcase the general impact of DL on bioinformatics. In addition, we raise the questions that users of such DL models should ask when assessing them for use in their research. We aim to provide a high-level yet comprehensive overview of the issues that may arise when applying DL models in bioinformatics. The work covers questions ranging from considering the original training data to making DL models somewhat more interpretable. We hope that equipped with these questions, bioinformaticians who employ DL models for inference can perform proper model selection.

In the following sections, ML, DL, and the relevance of DL in bioinformatics will be detailed.

1.1. Machine learning

Machine learning is a subfield of artificial intelligence that involves developing algorithms that can learn patterns in data and make predictions or decisions based on the trained model. Machine learning is used in many applications, from image recognition and natural language processing to autonomous vehicles and medical diagnosis (Acosta et al., 2022).

One of the fundamental requirements for machine learning is numeric data. Machine learning algorithms require numeric data because they use mathematical models to make predictions or decisions. Numeric data can be represented as a matrix, where each row represents a sample and each column represents a feature. Features are the characteristics of the data that the algorithm uses to make predictions or decisions.

In some cases, the available data may be in a different form than the numeric data. In these cases, the data must be transformed into numeric data before it can be used with machine learning algorithms. This process is called feature engineering and involves selecting, transforming, and combining the features in the data to create a new set of numeric features that can be used with machine learning algorithms (Roe et al., 2020).

In addition to numeric data, machine learning algorithms require high-quality data. High-quality data are accurate, complete, and representative of the problem the algorithm is trying to solve. High-quality data is essential because it ensures that the algorithm learns the correct patterns so that the resulting model can make accurate predictions or decisions (Habehh and Gohel, 2021).

Expert-crafted features can also be necessary for machine learning algorithms. Expert-crafted features are numerical representations of the data designed by domain experts who understand the problem the algorithm is trying to solve. Expert-crafted features can help improve machine learning algorithms' performance by providing more relevant information for the algorithm to learn from, especially if the amount of training data is restrictive (Lin et al., 2020).

Different machine learning algorithms can be used, depending on the problem that needs to be solved. Some popular learning algorithms include the following (Hastie et al., 2009; Jovel and Greiner, 2021):

- Linear regression is an algorithm that attempts to find the best-fit line that describes the relationship among variables.
- Logistic regression is a classification algorithm that predicts the probability of an event occurring.
- Decision trees are hierarchical algorithms that use a series of binary decisions to make predictions.
- Random forest is an ensemble algorithm that combines multiple decision trees to improve performance.
- Support vector machines are a classification algorithm that attempts to find the best hyperplane that separates the data into different classes.
- Neural networks represent an algorithm that uses layers of interconnected nodes to learn complex patterns from the data.

In conclusion, machine learning is a powerful tool that can solve many problems. It requires numeric data, high-quality data, and sometimes expert-crafted features. There are different learning algorithms to choose from depending on the problem that needs to be solved.

1.2. Deep learning

Neural networks were introduced decades ago. Back then, they typically consisted of a few layers (input, hidden, and output) of neurons, with the self-organizing maps being the smallest and containing only two layers (input and processing/presentation). Following the large increase in computational power since then, a larger number of layers can be handled today, and with more hidden layers, the networks are referred to as deep neural networks. Typically, deep learning models (deep neural networks) consist of three main layers: the input, hidden, and output layers.

The input layer (Figure 2, orange nodes) is the starting point of the neural network. It receives raw data or features used as input for the deep learning model. Each feature or data point corresponds to a node in the input layer. The values of these nodes are passed forward through the network for further processing. The number of neurons in the input layer is governed by the input. For example, if the input is a coin toss, one neuron would be enough to represent that. The more complex the input, the more neurons are needed.

Hidden layers (Figure 2, blue nodes) are intermediate between the input and output layers. They are responsible for transforming input data into more abstract and meaningful representations. Deep learning models can have multiple hidden layers, making them "deep" networks. Each node in a hidden layer can be connected to every node in the previous and subsequent layers; however, other connectivity, such as convolutional layers, is also used. The connections between nodes are represented by weights, adjusted during training to optimize the model's performance. The hidden layers can have any number of neurons, and there is no clear rule regarding how many neurons there should be. This needs to be established for each DL problem. The hidden layers also do not need the same number of neurons; they may increase, decrease, or remain the same throughout the hidden layers. In our example (Figure 2), the hidden layers contain more neurons than the input layer, but this is not always true.

The output layer is the final layer of the deep learning model. It produces the desired output or prediction based on the information processed in the hidden layers. Therefore, the number of nodes in the output layer depends on the specific task for which the model is designed. For example, in a classification task, each node in the output layer may represent a different class. In contrast, a single node may represent the predicted numeric value in a regression task. If the result is binary, one node would suffice.

Through training, deep learning models are trained by adjusting the weights in the connections between nodes to minimize errors and improve their performance on a specific task. This training is often accomplished

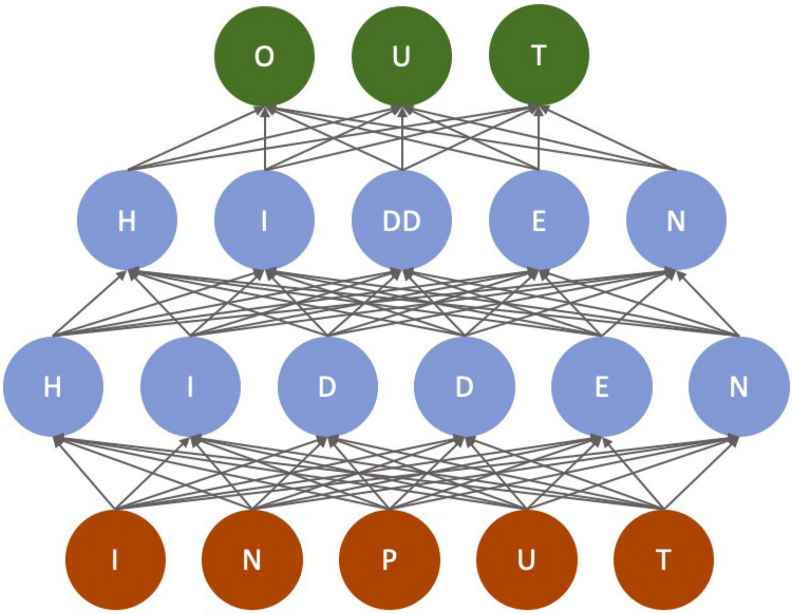

Figure 2. Fully connected deep learning network. The neurons of the input layer are in orange, those of the hidden layers are in blue, and the neurons of the output layer are in green. The connections in this example are directed from input to output and are indicated by arrows. Each arrow represents a trainable weight.

using large numbers of labeled data and optimization techniques such as backpropagation and gradient descent. When developing a DL model, it is typical that different architectures of the DL network are tried and that the most suitable architecture for the task is chosen on empirical grounds.

The major difference between deep learning and traditional machine learning is that DL detects higher-level features automatically. In contrast, ML depends on features suggested or generated by experts in the field (Sarker, 2021). Automatically discerning higher-level features can lead to a significant drawback in explainability, which is vital in healthcare, where it is crucial to understand the decision-making process (the Precise4Q consortium et al., 2020). For example, a decision tree transparently discloses all steps from input to decision. At the same time, a DL model implicitly contains all the information leading to the decision but needs to be explored or understood by humans.

We illustrate the major difference between traditional machine learning and deep learning by focusing on two aspects: feature generation and detection, as well as the interpretability of the models (Table).

On the left side, we represent traditional ML, where features are suggested or generated by experts in the field. This highlights the involvement of domain knowledge and human expertise in identifying the relevant features of the model. These features serve as inputs to the ML model, which then goes through the decision-making process. The decision-making process can be represented as a simple flow, such as a decision tree, where each step is easily understandable and traceable. This emphasizes the explainability and transparency of traditional ML models.

On the right side, we depict deep learning, which excels at automatically detecting higher-level features from raw data. Table represents this by showing an automatic feature detection process. DL models consist of complex, layered neural networks that can learn and extract abstract representations and features from data without explicit human intervention. However, this complexity also contributes to the need for more transparency and interpretability. While the DL model implicitly contains all the information leading to a decision, understanding the decision-making process or exploring the reasoning behind it is more complex for humans.

1.3. Why deep learning is relevant to bioinformatics

Bioinformatics involves the application of computational methods to analyze biological data (Bayat, 2002). Bioinformatics uses machine learning to analyze and interpret biological data. However, traditional machine learning methods require manual curation of features, which can be challenging. On the other hand, deep learning

Table. Differences between machine learning using traditional algorithms and machine learning using deep neural networks.

	ML	DL
Algorithms	Many different (SVM, DT, kNN, …)	Defined by architecture (RNN, GAN, LSTM, ...)
Data size	Can work well with smaller inputs	Requires large amount of data
Performance	Typically extremely fast	Computational complexity depends on the architecture
Features	Hand-crafted	Can be learned
Preprocessing	Significant effort	Can be trained on raw data
Fine tuning	Setting the algorithm parameters	Can be performed automatically during training
Complexity	Typical simple mathematical models	Depends on the architecture (highly flexible)
Transparency	Typically transparent	Hard to transparently show decision making
Explainability	Typically explainable	Hard to show the reasoning process

can learn higher-level features directly from the data (Lee et al., 2018). If it is possible to derive a mathematical formula that describes a problem, machine learning is no longer needed to address the issue. However, biology is a complex interplay of many factors that cannot be modeled in its entirety using mathematical formulas. Hence, the application of machine learning, specifically deep learning, is warranted. For example, in gene expression analysis, traditional machine learning methods involve the manual curation of features (Monaco et al., 2021). While this is relatively easy for gene expression, predicting whether an RNA sequence is a pre-microRNA is more challenging. Thousands of features need to be manually curated, and it needs to be clarified whether all of these features are relevant (Sacar and Allmer, 2013). In such cases, DL can learn higher-level features directly from the data (Kim et al., 2016), making it highly relevant to bioinformatics.

Despite the promising results achieved by deep learning in bioinformatics, challenges still need to be addressed. One of the major challenges is the need for large amounts of high-quality data (Sarker, 2021; Son et al., 2022). Deep-learning models require large amounts of data to learn complex patterns, and the data quality directly impacts the model's accuracy. Another challenge is the interpretability of deep-learning models (Meng et al., 2022). Deep-learning models are often described as "black boxes" since it is difficult to understand how they arrive at their predictions (Azodi et al., 2020).

In the following section, we will first discuss some applications of deep learning in bioinformatics before exploring ways to help shed light on the decision-making process of the so-called black boxes.

2. Selected applications of deep learning in bioinformatics

We have selected the following eight deep-learning tools because they represent a diverse range of applications in bioinformatics, demonstrate the power and flexibility of deep-learning approaches, and showcase the benefits of integrating deep learning into various aspects of biological research. The selected tools are DeepBind (Alipanahi et al., 2015), DeepCpG (Angermueller et al., 2017), DeepGene (Yuan et al., 2016), DeepFam (Seo et al., 2018), DeepLoc (Thumuluri et al., 2022), DeepPath (Coudray et al., 2018), ScanNet (Tubiana et al., 2022), and DeepVariant (Poplin et al., 2018). We will provide a short description of each tool in the following sections. All these tools can be considered black boxes, which limits the interpretability of their predictions (we discuss interpretability in section 3.3). Additionally, all models are trained on a subset of the chemical/biological possibilities so that they can have potential issues with the generalizability of the model. On the other hand, overfitting is a possibility for these models, which also hinders the generalizability of the model. We discuss these issues in section 3.1. Apart from these and other challenges, mentioned elsewhere in this text, we point out some limitations with each of the tools.

DeepBind (Alipanahi et al., 2015) is a deep-learning tool that predicts the binding specificity of DNA- and RNA-binding proteins and the effects of genetic mutations on these interactions. Using convolutional neural networks, DeepBind can identify binding patterns from large datasets of sequence information. This tool has significantly improved the accuracy of predicting protein-DNA interactions and has broad applications in understanding gene regulation and the impact of noncoding genomic variants in diseases. DeepBind has its limitations; for example, binding sites are heterogeneous (e.g., size, location, and sequence composition), which puts into question whether DeepBind can accurately predict all these types. Additionally, binding is a dynamic process that is further influenced by the environment, such as salt concentration. It is unlikely that all these influences are accurately modeled.

DeepCpG (Angermueller et al., 2017) is a deep-learning model that predicts methylation states in single cells using bisulfite sequencing data. By employing a combination of convolutional and recurrent neural networks, DeepCpG accurately models spatial and long-range dependencies in methylation patterns, providing valuable insights into the epigenetic landscape of individual cells. This tool is helpful for studying development, cellular differentiation, and diseases associated with epigenetic changes, such as cancer. DeepCpG is primarily designed for single-cell methylation data and, therefore, performance may degrade when using it for cell populations. Also, CpGs are not evenly distributed throughout a genome, and methylation patterns can vary across cell types. These limitations should be considered when using DeepCpG.

DeepGene (Yuan et al., 2016) is a deep-learning-based classifier for cancer subtypes using somatic point mutations. It employs a combination of restricted Boltzmann machines and deep neural networks to learn hierarchical representations of mutational patterns. This approach enables accurate classification of cancer subtypes and reveals potential driver mutations that contribute to carcinogenesis, providing valuable insights for precision oncology and personalized medicine. While many genes are well-studied across species, DeepGene was trained on a specific dataset, which may not include all possibilities. Additionally, genetic sequence features are very heterogenic, which may further limit the generalization of the model.

DeepFam (Seo et al., 2018) is a protein family classification tool that uses deep learning to predict the functional family of a given protein sequence. By employing a 1D convolutional neural network (CNN), DeepFam captures local and global sequence features, resulting in highly accurate family predictions. This tool aids in the functional annotation of proteins and supports large-scale analyses of protein sequence datasets, facilitating the discovery of novel protein families and studying protein evolution. When using the tool, one should keep in mind that protein families, for example, transcription factors, are highly heterogeneous and that many noncanonical protein sequences exist that may pose challenges to DeepFam.

DeepLoc (Thumuluri et al., 2022) is a deep-learning-based tool for predicting the subcellular localization of proteins. Using a combination of convolutional and recurrent neural networks, DeepLoc captures both sequence-based and evolutionary information, resulting in highly accurate localization predictions. This tool is essential for understanding protein function, protein-protein interactions, and cellular processes in various organisms. Similar to DeepBind, the environmental conditions exert a strong influence on prediction accuracy, which can be exemplified by membrane proteins. Especially with posttranslational modifications, this is a formidable challenge. Another limitation is that many proteins exist in multiple locales and that not always a confidence measure is attached to the predictions.

DeepPath (Coudray et al., 2018) is a deep-learning approach for inferring gene regulatory networks using gene expression data. By employing a combination of unsupervised feature learning and supervised classification, DeepPath learns the regulatory relationships between genes, providing insights into the complex regulatory mechanisms that govern cellular processes. This tool has broad applications in the study of gene regulation, disease mechanisms, and the development of therapeutic interventions. DeepPath may be sensitive to sequence variations, which may hamper the recognition of binding sites. Additionally, missing indirect evidence for binding or noncanonical sequence can limit the performance of DeepPath.

ScanNet (Tubiana et al., 2022) employs a geometric deep-learning model that directly learns features from protein structures to predict functional sites such as binding sites for small molecules, other proteins, or antibodies. ScanNet is accurate, versatile, and interpretable, making it suitable for functional site prediction tasks. It effectively detects protein-protein and protein-antibody binding sites and predicts epitopes of the SARS-CoV-2 spike protein. The same limitations that apply to DeepBind also apply to ScanNet. The applicability of ScanNet to noncanonical sequences may be especially limited.

DeepVariant (Poplin et al., 2018) is a deep-learning-based approach for variant calling in high-throughput sequencing data. Employing a deep neural network, DeepVariant identifies genomic variants with high accuracy and sensitivity while reducing false-positive calls. This tool is essential for studying genetic variation in populations, understanding the genetic basis of diseases, and advancing personalized medicine efforts. Data quality, especially base-calling accuracy, together with the depth of sequencing, affects DeepVariant's effectiveness. The same is true for contaminations in the sequencing data from, e.g., microbes.

In summary, these eight deep-learning tools showcase the versatility and power of deep-learning approaches in tackling diverse bioinformatics challenges. By harnessing the power of deep learning, these tools have significantly advanced our understanding of complex biological processes and contributed to various applications, including functional annotation, protein design, disease mechanism investigation, and personalized medicine. As deep-learning techniques continue to evolve and improve, they will undoubtedly play an increasingly important role in advancing the field of bioinformatics and our understanding of the underlying principles governing life. Clearly, all tools have general limitations that apply to all DL tools and have specific limitations, as pointed out above. However, tools that are not based on DL may have similar and/or further limitations, and we believe that with DL, predictions have become more accurate.

Bioinformatics encompasses many topics, tools, and analytical approaches. Deep learning is an approach that can be applied to many such areas. In the following, we will briefly mention some areas and indicate the selected tools. Keep in mind, that the same general limitations mentioned above and detailed later in the text apply to all the tools mentioned.

2.1. DNA sequencing

2.1.1. Sequence assembly

The genome of a species must be available for many downstream bioinformatics tasks. Therefore, sequence assembly is one of the first tasks performed in bioinformatics for any species of interest. Many tools for genome assembly exist and have been compared in an Assemblathon (Bradnam and Fass, 2013). Interestingly, we could not find an assembly tool that employs deep learning. However, for the assembly of metagenomes, MetaVelvet-DL (Liang and Sakakibara, 2021) is an extension of MetaVelvet (Namiki et al., 2012) that incorporates deep learning. The original MetaVelvet algorithm is an extension of the Velvet assembler, optimized explicitly for metagenomic data. It works by constructing de Bruijn graphs from the input sequencing reads and then identifying and partitioning these graphs into individual species or subgraphs to assemble the genomes of individual organisms. MetaVelvet-DL improves upon the original MetaVelvet algorithm by incorporating deep-learning techniques to better handle the complexity and diversity of metagenomic data. Using deep neural networks, MetaVelvet-DL can more accurately identify and partition the de Bruijn graphs, improving genome assembly and better resolution of individual species within the microbial community.

2.1.2. Genome annotation

The next step in bioinformatics analysis following sequence assembly is genome annotation.

The previously mentioned DeepVariant can be employed to annotate variants in a genome. Other approaches that are more directly targeted to genome annotation have been proposed (Yip et al., 2013; Shen et al., 2022). A more complete solution for this purpose is presented by DeepAnnotator (Amin et al., 2018), which provides a generalized computational approach for genome annotation at an F-score of 94%. DeepAnnotator is a deep-learning-based tool for functional annotation of proteins. It employs deep neural networks to predict protein function by classifying protein sequences into functional categories based on their amino acid sequences. The main goal of DeepAnnotator is to assign functional roles to proteins, which is essential for understanding the biological processes occurring within cells and for studying protein interactions, pathways, and the impact of genetic variations on protein function. DeepAnnotator uses a hierarchical deep neural network architecture to capture local and long-range dependencies within protein sequences, enabling it to learn complex sequence features and patterns associated with specific protein functions. The deep-learning model is trained on large-scale protein sequence datasets with known functional annotations, enabling it to recognize and predict the functions of novel protein sequences.

2.2. Gene expression analysis

Following genome assembly and annotation, another measure is the expression of such genes in an organism, especially the differential expression among different phenotypes. Several studies have shown that deep-learning models provide more accurate predictions of gene expression than traditional methods (Amin et al., 2018; Avsec et al., 2021). High-throughput gene expression profiling technologies, such as DNA microarrays and RNA sequencing, provide large gene expression datasets that can be analyzed using deep-learning algorithms (Zhang et al., 2021). Deep convolutional neural networks (CNNs) are currently the state-of-the-art method for predicting gene expression from DNA sequences (Avsec et al., 2021). Deep learning has also been used to discover biomarkers and identify genetic variations in human genomics (Alharbi and Rashid, 2022; Shen et al., 2022). Therefore, deep learning has been successfully applied to gene expression analysis in bioinformatics.

2.3. Gene function prediction

Gene function prediction is similar to genome annotation but focuses on the biological roles of the identified genes. Several tools are available for gene function prediction in bioinformatics.

DeepGOPlus (Kulmanov and Hoehndorf, 2021) is a deep-learning-based tool for predicting the function of proteins using their amino acid sequences. DeepGOPlus employs a neural network architecture, a precise combination of convolutional neural networks (CNNs) and long short-term memory (LSTM) networks, to predict gene ontology (GO) terms associated with proteins. DeepGOPlus captures local and global sequence features, leading to highly accurate predictions of molecular function, biological processes, and cellular component GO terms. This tool aids in understanding protein function and supports functional annotation efforts for newly sequenced genomes.

DeepGMAP (Onimaru et al., 2020) is a deep-learning-based tool for predicting the genomic location of transcription factor binding sites (TFBSs) using ChIP-seq data. DeepGMAP employs a combination of convolutional neural networks (CNNs) and recurrent neural networks (RNNs) to model the spatial and sequence-specific patterns of TF binding. DeepGMAP significantly improves the accuracy and specificity of TFBS prediction compared with traditional approaches, providing valuable insights into gene regulatory networks and the functional impact of noncoding genetic variants.

DeepNF (Gligorijević et al., 2018) is a deep-learning-based tool for predicting protein-protein interactions and functional associations. Using an unsupervised deep-learning approach, it integrates multiple types of biological data, such as protein sequence, domain composition, and protein-protein interaction networks. DeepNF employs stacked autoencoders to learn a joint representation of the input data, which can then be used to predict protein-protein interactions and functional associations accurately. This tool is essential for studying protein function, cellular processes, and the development of therapeutic interventions targeting specific protein interactions.

DeepMir (Cordero et al., 2019) is a deep-learning-based tool for identifying and classifying microRNA (miRNA) precursors, which are small noncoding RNAs that play crucial roles in gene regulation and are implicated in various biological processes and diseases. DeepMir employs a convolutional neural network to predict miRNA precursor sequences from a given genomic sequence represented as abstract images. The model achieves high accuracy and includes initial steps toward its explainability.

The use of DL can tackle many more challenges in bioinformatics, and the tools that have been mentioned so far showcase that all areas of bioinformatics currently see the application of DL. However, the application of DL in bioinformatics is challenging. Some of these will be discussed below.

2.4. Protein structure prediction

Protein structure prediction is a fundamental problem in bioinformatics because the 3D structure of a protein determines its function and interactions. However, experimentally determining protein structures is costly and time-consuming, and only a fraction of known proteins have their structures solved. Therefore, developing computational methods to accurately predict protein structures from their amino acid sequences is a major challenge and a long-standing goal of bioinformatics research. This has long been realized, and a protein structure prediction challenge was created in the 1990s (Moult et al., 1997).

One of the most successful computational methods for protein structure prediction is AlphaFold 2 (Skolnick et al., 2021), a deep-learning tool developed by Google DeepMind. AlphaFold uses a novel deep-learning architecture to learn complex patterns and relationships between amino acids in the protein sequence and predict their distances and angles in 3D space. AlphaFold also incorporates evolutionary information from multiple sequence alignments and uses a graph neural network to represent the protein as a complex system of interacting amino acids. AlphaFold outputs a confidence score for each predicted structure, indicating its reliability.

AlphaFold has demonstrated remarkable performance in the Critical Assessment of Protein Structure Prediction (CASP), a biennial community challenge for testing the accuracy of protein structure prediction methods. In CASP13 (2018), AlphaFold placed first among the participating teams, showing significant improvements over previous methods (AlQuraishi, 2019). In CASP14 (2020), AlphaFold achieved an average accuracy competitive with experimental structures, effectively solving the protein structure prediction problem in most cases (Jumper et al., 2021).

2.5. Disease diagnostics and drug discovery

Disease diagnostics is a crucial task in bioinformatics and healthcare because it involves identifying and classifying diseases based on various data types, such as clinical symptoms, laboratory tests, medical images, and genomic sequences. Disease diagnostics can benefit from applying deep learning tools, which can learn complex patterns and features from large-scale data and make accurate and robust predictions. This area is wide-ranging from multi-omic data evaluation to the application of chatbots in anamnesis. In this overview of deep learning in bioinformatics, we cannot go into details and invite the interested readers to consider the works by Park et al., Kumar et al. and Myszczynska and colleagues (Myszczynska et al., 2020; Park et al., 2021; Kumar et al., 2023). These examples illustrate the potential and diversity of deep learning tools for disease diagnostics in bioinformatics. By applying deep learning to various types of data, these tools can improve disease diagnostics' accuracy, efficiency, and reliability and contribute to a better understanding of disease mechanisms and outcomes.

After disease diagnosis, drug discovery is another task that is too large to be discussed in detail in this small overview. Drug discovery is a challenging and costly process that involves identifying and optimizing novel chemical entities that can modulate biological targets and treat diseases. Deep learning has seen considerable adoption in the field. Please consider the following three examples of the potential and diversity of deep learning tools for drug discovery in bioinformatics. These tools can improve the efficiency, accuracy, and creativity of drug discovery and development by applying deep learning to various types of data, such as chemical structures, protein sequences, biological assays, and clinical outcomes.

DeepChem (Altae Tran et al., 2017) is an open-source deep learning framework for drug discovery. It provides various modules and functionalities for data preprocessing, model building, model evaluation, and model deployment. DeepChem can be used for various drug discovery tasks, such as molecular property prediction, virtual screening, de novo drug design, and drug synthesis planning.

ODDT (Wójcikowski et al., 2015) is an open-source tool for computer-aided drug discovery (CADD). It integrates various methods and algorithms for molecular docking, pharmacophore modeling, similarity searching, machine learning, and deep learning. ODDT can be used for various CADD tasks, such as target identification, hit identification, lead optimization, and ADMET prediction.

Cyclica is a company that uses deep learning to accelerate drug discovery (Abdollahi et al., 2023). It offers various solutions for target identification, polypharmacology prediction, drug design, and drug repurposing.

3. Challenges of deep learning in bioinformatics

While information concerning data may be considered only crucial for model training, it is essential to openly disclose the data used in model training so that model users can inspect it. With this information, potential users can consider the model, compare it to others, and decide whether to use it or not.

3.1. Training data

Training data quality plays a crucial role in the success of deep-learning applications. High-quality data are essential for training accurate and robust deep-learning models that can effectively capture complex biological patterns and relationships (Fan and Shi, 2022). With an increase in data dimensions, more data is needed to train effective models; therefore, the quantity of the data becomes important and, with it, whether the data represents all diverse aspects of the biological phenomenon. Proper annotation and handling of noise and errors are also important. While it may seem of little importance to know how a model was trained when it reaches high accuracy, it may only do so for specific data. To judge this, information about the training data is needed, and several aspects must be considered.

3.1.1. Quantity of data

Deep-learning models usually require large amounts of data for training because they can learn complex patterns and representations (Lee et al., 2022). Insufficient data regarding data and model dimensionality can lead to overfitting, where the model memorizes the training data and does not generalize well to new, unseen data (Demšar and Zupan, 2021). In bioinformatics, obtaining large-scale datasets can be challenging because of various factors, such as the cost and time associated with experimental data generation, limited availability of well-annotated data, and the inherent complexity of biological systems (Faustino et al., 2008). A simple example would be creating a daily melatonin cycle model. We would need at least hourly measurements to allow an hourly resolution of the prediction. To make it more general, we would need this for many days, not one day. We suggest having ten times more data than the model parameters for DL models. Inspecting how the model was trained can help decide whether the model is suitable for the intended purpose.

3.1.2. Representation and diversity of the data

Data used for training deep-learning models should represent the studied biological system and cover various examples and scenarios (Ching et al., 2018). Bioinformatics means including data from various species, tissues, experimental conditions, and disease states. A diverse and representative dataset ensures that the model can capture the variability and complexity of biological systems and make accurate predictions on new, unseen data. When considering a DL model for inference, the amount and breadth of data used to train the model should be considered in conjunction with the purpose. If the breadth of the data that is supposed to be processed is covered in the model's training data, it can be suitable even if it may not be suitable for another closely related dataset. However, biological data can be noisy and subject to various sources of error (Tsimring, 2014). These errors can negatively impact the performance of deep-learning models, leading to inaccurate predictions and reduced generalizability (Karimi et al., 2020).

3.1.3. Noise and error

Biological data, especially those generated by high-throughput experimental techniques, can be noisy and subject to various sources of error (Li et al., 2019). These errors can arise from technical issues such as sequencing errors, experimental variability, batch effects, or other biological factors such as genetic variation or rare and uncharacterized sequences. Noisy and error-prone data can negatively impact the performance of deep-learning models, leading to inaccurate predictions and reduced generalizability (Alipanahi et al., 2015; Karimi et al., 2020). With an increasing amount of data, the impact of noise and error diminishes. Therefore, a model trained on large amounts of data in relation to the model parameters is preferable.

3.1.4. Annotation quality and consistency

In supervised deep learning, models are trained on data with known labels or annotations, such as protein functions or gene regulatory relationships. The quality and consistency of these annotations directly influence the model's performance (Chen et al., 2021). In bioinformatics, annotations can be derived from experimental data, literature curation, or computational predictions, and their quality and reliability can vary widely. Inaccurate or inconsistent annotations can lead to poor model performance and misleading predictions.

3.1.5. Data preprocessing and normalization

Appropriate data preprocessing and normalization are critical for ensuring that the input data are suitable for deep-learning models (Imran et al., 2022). Some bioinformatics analyses involve various steps, such as sequence alignment, quality control, feature extraction, and data transformation. Careful preprocessing and normalization can reduce the impact of noise and errors, ensure comparability across different datasets, and improve the performance of deep-learning models. This is even more important when considering metagenomic data. An overview of preprocessing steps for preparing microbiome sequencing data for machine learning is given by (Ibrahimi et al., 2023). Normalization of sequencing results, for example, in transcriptomics, is essential, and a recent evaluation can be found here (Ni and Qin, 2021). However, in biology, there is often no gold standard data, so normalization and preprocessing approaches cannot easily be benchmarked. Additionally, simply changing some cutoff value, e.g., for counts after RNA-seq analysis, can have a large impact on the results (Beukers and Allmer, 2023).

In conclusion, data quality is critical to the success of deep-learning applications in bioinformatics. Ensuring sufficient quantity, diversity, and representation of data, minimizing noise and errors, maintaining high-quality annotations, and employing appropriate preprocessing and normalization techniques are essential for developing accurate and robust deep-learning models that can advance our understanding of complex biological systems and contribute to various applications in molecular biology, genetics, and systems biology.

3.2. Computational requirements

Applying deep learning in bioinformatics often demands substantial computational resources because of the complexity of biological data and the inherent computational intensity of deep-learning algorithms (Moreno et al., 2022). This is especially true when training DL models; however, the computational requirements can also be prohibitive for running already trained models for inference.

3.2.1. Processing power

Deep-learning models, especially those with multiple layers, many neurons, and many edges, require significant processing power for training and inference. Inference on a laptop or PC may not be possible, depending on the DL model. High-performance processors, such as graphics processing units (GPUs) or specialized tensor processing units (TPUs), are often used to accelerate deep learning computations, as they are specifically designed for parallel processing of large-scale mathematical operations. For smaller models, it is possible to perform inference on a PC with an average GPU; however, several GPUs may be needed for larger models. Alternatively, many cloud services offer GPU access hourly so that calculations can be performed in the cloud. For example, TPUs introduced by Google are available on the Google cloud. Other large open-source projects, such as HuggingFace, also offer access to computing.

3.2.2. Memory capacity

Deep-learning models require considerable memory capacity, particularly those with billions to trillions of parameters and extensive input data. The model parameters, input data, and intermediate values, such as activations and gradients, must be stored in the memory during training. Insufficient memory capacity can limit the size and complexity of models that can be trained and the size of the input data that can be processed. Thus, having sufficient RAM and GPU memory is crucial for deep learning in bioinformatics. While training is resource-intensive, inference, the process of making predictions with a trained model, also demands a large memory capacity. It is necessary to accommodate the model's size, especially if it has been trained with several parameters. An insufficient amount of RAM during the inference phase could lead to suboptimal processing speeds, thereby affecting the usability and efficiency of the model. The number of parameters is especially prohibitive in large language models such as GPT4, but many bioinformatics applications may not be as resource-hungry. AlphaFold, for instance, has only around 100 million parameters, which is several orders of magnitude smaller than GPT4. Hence, many bioinformatics DL models may run on consumer hardware. While AlphaFold may run with 32 GB of RAM, more is better.

3.2.3. Data storage

Bioinformatics datasets, especially those generated by high-throughput experimental techniques, can be enormous, necessitating substantial storage capacity. Deep-learning models often require access to large-scale training data to learn complex patterns and representations effectively. As a result, deep-learning applications in bioinformatics may require extensive storage solutions, such as high-capacity hard drives, solid-state drives, or distributed storage systems. LLMs such as GPT may need tens of TB of data to train the model, but much less data storage capacity is needed for inference. Some bioinformatics applications, such as AlphaFold, require a relatively large storage capacity. AlphaFold, for instance, stores large sequence databases and requires 3 TB of hard disk space. Faster storage is preferable. Today, some laptops, such as the recent Macbook Pro, can come with a 4 TB hard drive capacity and 64GB of RAM, allowing the execution of AlphaFold.

3.3. Interpretability and explainability

One of the major challenges associated with deep-learning models is their black-box status. It is often difficult to understand how the model arrives at its predictions.

In bioinformatics, where the interpretation of results is critical for understanding biological processes, the need for more interpretability of deep-learning models is a significant concern (Petkovic et al., 2018). In this section, we explore the current state of interpretability in deep-learning models and discuss opportunities for overcoming the black box status of these models.

One approach is to use model-agnostic methods for interpretability. These methods involve analyzing the model's behavior using perturbations or sensitivity analysis. Examples of such methods include Local Interpretable Model-Agnostic Explanations (LIME) and Shapley Additive exPlanations (SHAP) (Sathyan et al., 2022). LIME (Ribeiro et al., 2016) is a model-agnostic method that explains individual predictions of a deep-learning model by creating a simplified, interpretable model that approximates the behavior of the original model. LIME generates perturbed versions of the input data and measures how the model's output changes for each perturbation. Then, LIME fits a linear or logistic regression model to the perturbed data and uses the regression model coefficients as weights to assign importance scores to each input feature. LIME can explain different data types, such as text, images, or tabular data. SHAP is another model-agnostic method based on the concept of Shapley values, which are derived from game theory and measure how much each player contributes to the outcome of a cooperative game. SHAP considers each input feature as a player and calculates the Shapley value for each feature by averaging all possible combinations of features and measuring how much the feature changes the model's output when added or removed from the combination. SHAP can explain different models, such as tree-based models, deep neural networks, or kernel-based models.

Another approach is to use model-specific methods for interpretability. These methods involve modifying the architecture of the deep-learning model to incorporate explicit mechanisms for interpretability. Such methods include attention mechanisms, which allow the model to focus on specific parts of the input, and saliency maps, which highlight important input regions contributing to the model's prediction (Figueroa et al., 2022). Additional layers can be incorporated into the deep-learning architecture to support the explainability (Vinuesa and Sirmacek, 2021).

Incorporating domain knowledge into the deep-learning model is another way to explain models' decisions. In bioinformatics, domain knowledge can guide learning and constrain the model to produce biologically meaningful results. For example, in gene expression analysis, prior knowledge of gene regulatory networks can be incorporated into the deep-learning model to improve its interpretability (Fortelny and Bock, 2020). Another approach is to use a multimodal data integration (Zhu et al., 2023). In bioinformatics, multimodal data integration can involve combining data from different sources, such as gene expression data, protein-protein interaction data, and pathway data, to improve the interpretability of the deep-learning model. Integrating multiple data sources makes generating more comprehensive and biologically meaningful models possible.

Several future directions can be pursued to overcome the black box status of deep-learning models in bioinformatics (Teng et al., 2022). One approach is to develop hybrid models that combine deep learning with machine learning methods, such as rule-based systems or decision trees (Ferry et al., 2023). By combining different strategies, it may be possible to generate more interpretable models that capture the complexity of biological systems. Another direction is to develop methods for evaluating the interpretability of deep-learning models (Meng et al., 2022). Currently, there is yet to be a widely accepted metric for assessing the interpretability of deep-learning models. Developing such a metric would enable researchers to compare the interpretability of different models and create new, more interpretable models.

We explored the explainability of DL models in our study on pre-microRNA prediction using DL (Cordero et al., 2019). Because we transformed the input data into images, we could build on top of large image models. This also enabled us to explore which parts of an image support the decision of whether an image represents a pre-miRNA or not, employing saliency maps (Simonyan et al., 2013).

4. Ethical and social implications

Applying deep learning in bioinformatics has great potential to advance biological knowledge and improve human health. However, it also raises ethical and social issues that must be addressed and resolved. These questions are equally valid for training and inference. When using a model for inference, it is essential to ensure that the results are unbiased. Some of these issues are as follows:

Data privacy: Deep learning requires large amounts of data to train and validate its models, which may include sensitive personal or health information. Critical ethical questions that need to be considered are how to protect the privacy and confidentiality of such data and obtain informed consent from the data providers or participants (Li et al., 2019).

Bias: Deep learning models may inherit or amplify biases in data or algorithms, leading to unfair or inaccurate outcomes or decisions. For example, deep learning models for disease diagnosis may perform differently for different populations or subgroups, depending on the data quality and representation (Ellis et al., 2022). Detecting and mitigating such biases and ensuring the fairness and transparency of deep learning models are crucial social challenges that need to be addressed (Pagano et al., 2023).

Accountability: Deep learning models may significantly impact the lives and well-being of individuals or society, especially when used for high-stakes applications such as drug discovery or precision medicine. However, deep learning models are often complex and opaque, making it difficult to explain or understand their logic or reasoning. How to ensure the accountability and responsibility of the developers and users of deep learning models and how to establish appropriate regulations and standards for their development and application are essential ethical and social issues that need to be resolved (Floridi et al., 2018).

5. Conclusion

5.1. Summary of deep learning in bioinformatics

Deep learning in bioinformatics refers to applying advanced neural network architectures and algorithms to analyze and interpret complex biological data. By leveraging the power of deep learning, researchers can uncover hidden patterns, relationships, and features within biological data, leading to new insights and discoveries in molecular biology, genetics, and systems biology.

Some critical aspects of the application of deep learning in bioinformatics areas follow:

- Handling of diverse biological data types. Deep-learning techniques can process various kinds of biological data, such as DNA sequences, protein sequences, gene expression data, and protein-protein interaction networks.
- Development of specialized deep-learning architectures. Customized deep-learning architectures, such as convolutional neural networks, recurrent neural networks, and autoencoders, are employed to tackle specific bioinformatics tasks, such as protein function prediction, gene regulatory network inference, and protein structure prediction.
- Quality of training data and its preprocessing. Ensuring high-quality data and appropriate preprocessing techniques are critical for the success of deep-learning applications in bioinformatics. This includes handling noise, errors, and diverse data representation.
- Computational requirements: Deep learning in bioinformatics requires substantial computational resources, such as processing power, memory capacity, data storage, network bandwidth, and scalability, to handle the complexity of biological data and the computational intensity of deep-learning algorithms.

Deep learning has significantly advanced the field of bioinformatics, enabling researchers to tackle complex challenges and gain a better understanding of biological processes. It has been applied to various bioinformatics tasks, such as functional annotation, protein design, disease mechanism investigation, and personalized medicine.

5.2. Future directions for deep learning in bioinformatics

Deep learning is a powerful tool for analyzing and interpreting biological data. Its ability to learn higher-level features directly from the data makes it highly relevant to bioinformatics, where traditional expert manual feature crafting approaches may be too time-consuming. Although there are challenges to be addressed, the continued application of deep learning in bioinformatics holds great promise for advancing our understanding of biological systems.

As deep learning continues to advance significantly in various fields, its application in bioinformatics is also expected to grow and evolve.

Integrating multiple omics data types, such as genomics, transcriptomics, proteomics, and metabolomics, can provide a more comprehensive understanding of biological systems. Deep-learning models can be designed to integrate and analyze multi-omics data effectively (Kang et al., 2022), leading to improved predictions, a better understanding of disease mechanisms, and the identification of novel biomarkers and therapeutic targets.

While deep-learning models have shown great success in various bioinformatics tasks, their predictions are often considered black boxes because the knowledge representation in the model is not explicit. Developing interpretable and explainable deep-learning models is essential for building trust and understanding the biological basis of their predictions, which can lead to more actionable insights and hypotheses. Developing interpretable deep-learning models will also be a key area of research, enabling us to understand these systems' underlying biology better.

In bioinformatics, obtaining large-scale, well-annotated data can be challenging. Transfer learning, as we performed for pre-miRNA prediction, and few-shot learning approaches, which involve leveraging pretrained models or learning from small amounts of data, can be employed to overcome data limitations and improve the performance of deep-learning models in tasks with limited training data.

Developing deep-learning models that can generalize well across different biological systems, species, and experimental conditions is essential for their broad applicability. Techniques to improve model generalization and robustness, such as domain adaptation and data augmentation, can enhance the utility of deep-learning models in bioinformatics. This is an important area, as seen from our work on pre-miRNA prediction (Saçar Demirci et al., 2017) and de novo sequencing (Savas Takan and Allmer, 2023).

Biological systems exhibit complex behavior across multiple scales, ranging from the molecular to cellular, tissue, and organism levels. Developing deep-learning models capable of capturing and integrating information across different scales can lead to a more comprehensive understanding of biological processes and the relationships between different levels of organization.

As deep learning advances, there will be an increasing need for interdisciplinary collaboration among computer scientists, biologists, and other domain experts. These collaborations will facilitate the development of novel deep-learning methods tailored to the unique challenges of bioinformatics and help bridge the gap between computational predictions and biological validation.

Continued advancements in hardware, such as GPUs, TPUs, and neuromorphic chips (Pastur Romay et al., 2016), will enable the training of larger and more complex deep-learning models. Furthermore, developing efficient and scalable deep-learning software frameworks will facilitate the application of deep learning to bioinformatics challenges.

In summary, the future of deep learning in bioinformatics is expected to involve the development of novel models and techniques, improved integration of multi-omics data, enhanced interpretability, better generalization and robustness, multiscale modeling, interdisciplinary collaboration, and advancements in hardware and software. These directions will help deepen our understanding of complex biological systems, drive discoveries, and contribute to various molecular biology, genetics, and systems biology applications.

Conflict of interest

The authors declare no conflict of interest.

References

Abdollahi N, Tonekaboni SAM, Huang J, Wang B, MacKinnon S (2023). NodeCoder: a graph-based machine learning platform to predict active sites of modeled protein structures. https://doi.org/10.48550/ARXIV.2302.03590

Acosta JN, Falcone GJ, Rajpurkar P, Topol EJ (2022). Multimodal biomedical AI. Nature Medicine 28 (9): 1773–1784. https://doi.org/10.1038/s41591-022-01981-2

Ahsan MM, Luna SA, Siddique Z (2022). Machine-Learning-Based Disease Diagnosis: A Comprehensive Review. Healthcare. 10 (3): 541. https://doi.org/10.3390/healthcare10030541

Al taweraqi N, King RD (2022). Improved prediction of gene expression through integrating cell signalling models with machine learning. BMC Bioinformatics. 23 (1): 323. https://doi.org/10.1186/s12859-022-04787-8

Alharbi WS, Rashid M (2022). A review of deep learning applications in human genomics using next-generation sequencing data. Human Genomics. 16 (1): 26. https://doi.org/10.1186/s40246-022-00396-x

Alipanahi B, Delong A, Weirauch MT, Frey BJ (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nature Biotechnology. 33 (8): 831–838. https://doi.org/10.1038/nbt.3300

AlQuraishi M. (2019). AlphaFold at CASP13. Bioinformatics. 35 (22): 4862–4865. https://doi.org/10.1093/bioinformatics/btz422

AlQuraishi M. (2021). Machine learning in protein structure prediction. Current Opinion in Chemical Biology. 65: 1–8. https://doi.org/10.1016/j.cbpa.2021.04.005

Altae Tran H, Ramsundar B, Pappu AS, Pande V. (2017). Low Data Drug Discovery with One-Shot Learning. ACS Central Science 3 (4): 283–293. https://doi.org/10.1021/acscentsci.6b00367

Alzubaidi L, Zhang J, Humaidi AJ, Al Dujaili A, Duan Y et al. (2021). Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. Journal of Big Data. 8 (1): 53. https://doi.org/10.1186/s40537-021-00444-8

Amin MR, Yurovsky A, Tian Y, Skiena S (2018). DeepAnnotator: Genome Annotation with Deep Learning. In: Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics. Washington DC USA: ACM. p. 254–259. https://doi.org/10.1145/3233547.3233577

Angermueller C, Lee HJ, Reik W, Stegle O (2017). DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning. Genome Biology 18 (1): 67. https://doi.org/10.1186/s13059-017-1189-z

Avsec Ž, Agarwal V, Visentin D, Ledsam JR, Grabska Barwinska A et al. (2021). Effective gene expression prediction from sequence by integrating long-range interactions. Nature Methods. 18 (10): 1196–1203. https://doi.org/10.1038/s41592-021-01252-x

Azodi CB, Tang J, Shiu SH (2020). Opening the Black Box: Interpretable Machine Learning for Geneticists. Trends in Genetics. 36 (6): 442–455. https://doi.org/10.1016/j.tig.2020.03.005

Baştanlar Y, Ozuysal M (2014). Introduction to machine learning. Methods in molecular biology (Clifton, NJ). 1107: 105–28. https://doi.org/10.1007/978-1-62703-748-8_7

Bayat A (2002). Science, medicine, and the future: Bioinformatics. British Medical Journal. 324 (7344): 1018–1022. https://doi.org/10.1136/bmj.324.7344.1018

Beukers M, Allmer J (2023). Challenges for the Development of Automated RNA-seq Analyses Pipelines. German Medical Science Medizinische Informatik, Biometrie und Epidemiologie. https://doi.org/10.3205/MIBE000245

Bohr A, Memarzadeh K (2020). The rise of artificial intelligence in healthcare applications. In: Artificial Intelligence in Healthcare. Elsevier. p. 25–60. https://doi.org/10.1016/B9780128184387000022

Bradnam K, Fass J (2013). Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. arXiv preprint.:1–31. https://doi.org/10.1186/2047-217X-2-10

Cervone HF (2016). Informatics and data science: an overview for the informationprofessional. Digital Library Perspectives. 32 (1): 7–10. https://doi.org/10.1108/DLP-10-2015-0022

Chen J, Geard N, Zobel J, Verspoor K (2021). Automatic consistency assurance for literature-based gene ontology annotation. BMC Bioinformatics. 22 (1): 565. https://doi.org/10.1186/s12859-021-04479-9

Chen J, Lin Y, Shen B (2017). Informatics for Precision Medicine and Healthcare. In: Shen B, editor. Translational Informatics in Smart Healthcare. Vol. 1005. Singapore: Springer Singapore. (Advances in Experimental Medicine and Biology). p. 1–20. https://doi.org/10.1007/978-981-10-5717-5_1

Ching T, Himmelstein DS, Beaulieu Jones BK, Kalinin AA, Do BT et al. (2018). Opportunities and obstacles for deep learning in biology and medicine. Journal of the Royal Society Interface. 15 (141): 20170387. https://doi.org/10.1098/rsif.2017.0387

Cordero J, Menkovski V, Allmer J (2019). Detection of pre-microRNA with Convolutional Neural Networks. Bioinformatics. biorXiv:preprint. https://doi.org/10.1101/840579

Coudray N, Ocampo PS, Sakellaropoulos T, Narula N, Snuderl M et al. (2018). Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. Nature Medicine. 24 (10): 1559–1567. https://doi.org/10.1038/s41591-018-0177-5

Davenport T, Kalakota R (2019). The potential for artificial intelligence in healthcare. Future Healthcare Journal 6 (2): 94–98. https://doi.org/10.7861/futurehosp.6-2-94

Demšar J, Zupan B (2021). Hands-on training about overfitting. PLoS Computational Biology 17 (3): e1008671. https://doi.org/10.1371/journal.pcbi.1008671

Ellis RJ, Sander RM, Limon A (2022). Twelve key challenges in medical machine learning and solutions. Intelligence-Based Medicine. 6: 100068. https://doi.org/10.1016/j.ibmed.2022.100068

Fan FJ, Shi Y (2022). Effects of data quality and quantity on deep learning for protein-ligand binding affinity prediction. Bioorganic & Medicinal Chemistry. 72: 117003. https://doi.org/10.1016/j.bmc.2022.117003

Faustino RS, Chiriac A, Terzic A (2008). Bioinformatic primer for clinical and translational science. Clinical and Translational Science 1 (2): 174–180. https://doi.org/10.1111/j.1752-8062.2008.00038.x

Ferry J, Laberge G, Aïvodji U (2023). Learning Hybrid Interpretable Models: Theory, Taxonomy, and Methods. https://doi.org/10.48550/ARXIV.2303.04437

Figueroa KC, Song B, Sunny S, Li S, Gurushanth K et al. (2022). Interpretable deep learning approach for oral cancer classification using guided attention inference network. Journal of Biomedical Optics 27 (1): 015001. https://doi.org/10.1117/1.JBO.27.1.015001

Floridi L, Cowls J, Beltrametti M, Chatila R, Chazerand P et al. (2018). AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. Minds & Machines. 28 (4): 689–707. https://doi.org/10.1007/s11023-018-9482-5

Fortelny N, Bock C (2020). Knowledge-primed neural networks enable biologically interpretable deep learning on single-cell sequencing data. Genome Biology 21 (1): 190. https://doi.org/10.1186/s13059-020-02100-5

Gligorijević V, Barot M, Bonneau R (2018). deepNF: deep network fusion for protein function prediction. Wren J, editor. Bioinformatics. 34 (22): 3873–3881. https://doi.org/10.1093/bioinformatics/bty440

Habehh H, Gohel S (2021). Machine Learning in Healthcare. Current Genomics 22 (4): 291–300. https://doi.org/10.2174/1389202922666210705124359

Hastie T, Tibshirani R, Friedman JH (2009). Overview of Supervised Learning. In: The Elements of Statistical Learning. Springer Series in Statistics. Springer, New York, NY. https://doi.org/10.1007/978-0-387-84858-7_2

Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL (2018). Artificial intelligence in radiology. Nature Reviews Cancer. 18 (8): 500–510. https://doi.org/10.1038/s41568-018-0016-5

Hulsen T, Jamuar SS, Moody AR, Karnes JH, Varga O et al. (2019). From Big Data to Precision Medicine. Frontiers in Medicine 6: 34. https://doi.org/10.3389/fmed.2019.00034

Ibrahimi E, Lopes MB, Dhamo X, Simeon A, Shigdel et al. (2023). Overview of data preprocessing for machine learning applications in human microbiome research. Frontiers in Microbiology 14: 1250909. https://doi.org/10.3389/fmicb.2023.1250909

Imran, Qayyum F, Kim DH, Bong SJ, Chi SY et al. (2022). A Survey of Datasets, Preprocessing, Modeling Mechanisms, and Simulation Tools Based on AI for Material Analysis and Discovery. Materials (Basel). 15 (4): 1428. https://doi.org/10.3390/ma15041428

Jovel J, Greiner R (2021). An Introduction to Machine Learning Approaches for Biomedical Research. Frontiers in Medicine 8: 771607. https://doi.org/10.3389/fmed.2021.771607

Jumper J, Evans R, Pritzel A, Green T, Figurnov M et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature. 596 (7873): 583–589. https://doi.org/10.1038/s41586-021-03819-2

Kang M, Ko E, Mersha TB (2022). A roadmap for multi-omics data integration using deep learning. Briefings in Bioinformatics. 23 (1): bbab454. https://doi.org/10.1093/bib/bbab454

Karimi D, Dou H, Warfield SK, Gholipour A (2020). Deep learning with noisy labels: Exploring techniques and remedies in medical image analysis. Medical Image Analysis 65: 101759. https://doi.org/10.1016/j.media.2020.101759

Kim J, Calhoun VD, Shim E, Lee JH (2016). Deep neural network with weight sparsity control and pre-training extracts hierarchical features and enhances classification performance: Evidence from whole-brain resting-state functional connectivity patterns of schizophrenia. NeuroImage. 124: 127–146. https://doi.org/10.1016/j.neuroimage.2015.05.018

Korteling JE (Hans), van de Boer Visschedijk GC, Blankendaal RAM, Boonekamp RC, Eikelboom AR (2021). Human- versus Artificial Intelligence. Frontiers in Artificial Intelligence 4: 622364. https://doi.org/10.3389/frai.2021.622364

Kulmanov M, Hoehndorf R (2021). DeepGOPlus: improved protein function prediction from sequence. Bioinformatics. 37 (8): 1187. https://doi.org/10.1093/bioinformatics/btaa763

Kumar Y, Koul A, Singla R, Ijaz MF (2023). Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda. Journal of Ambient Intelligence and Humanized Computing. 14 (7): 8459–8486. https://doi.org/10.1007/s12652-021-03612-z

LeCun Y, Bengio Y, Hinton G (2015). Deep learning. Nature. 521 (7553): 436–444. https://doi.org/10.1038/nature14539

Lee BD, Gitter A, Greene CS, Raschka S, Maguire F et al. (2022). Ten quick tips for deep learning in biology. Ouellette F, editor. Public Library of Sience Computational Biology. 18 (3): e1009803. https://doi.org/10.1371/journal.pcbi.1009803

Lee K, Famiglietti ML, McMahon A, Wei CH, MacArthur JAL et al. (2018). Scaling up data curation using deep learning: An application to literature triage in genomic variation resources. Public Library of Science Computational Biology. 14 (8): e1006390. https://doi.org/10.1371/journal.pcbi.1006390

Lepakshi VA (2022). Machine Learning and Deep Learning based AI Tools for Development of Diagnostic Tools. In: Computational Approaches for Novel Therapeutic and Diagnostic Designing to Mitigate SARS-CoV-2 Infection. Elsevier. p. 399–420. https://doi.org/10.1016/B978-0-323-91172-6.00011-X

Li Y, Huang C, Ding L, Li Z, Pan Y et al. (2019). Deep learning in bioinformatics: Introduction, application, and perspective in the big data era. Methods. 166: 4–21. https://doi.org/10.1016/j.ymeth.2019.04.008

Li Z, Wu Z, Jin P, Wu H (2019). Dissecting differential signals in high-throughput data from complex tissues. Bioinformatics. 35 (20): 3898–3905. https://doi.org/10.1093/bioinformatics/btz196

Liang K, Sakakibara Y (2021). MetaVelvet-DL: a MetaVelvet deep learning extension for de novo metagenome assembly. BMC Bioinformatics. 22 (S6): 427. https://doi.org/10.1186/s12859-020-03737-6

Lin W, Hasenstab K, Moura Cunha G, Schwartzman A (2020). Comparison of handcrafted features and convolutional neural networks for liver MR image adequacy assessment. Scientific Reports 10 (1): 20336. https://doi.org/10.1038/s41598-020-77264-y

Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F et al. (2017). A survey on deep learning in medical image analysis. Medical Image Analysis. 42: 60–88. https://doi.org/10.1016/j.media.2017.07.005

Mahoto NA, Shaikh A, Al Reshan MS, Memon MA, Sulaiman A (2021). Knowledge Discovery from Healthcare Electronic Records for Sustainable Environment. Sustainability. 13 (16): 8900. https://doi.org/10.3390/su13168900

Meng C, Trinh L, Xu N, Enouen J, Liu Y (2022). Interpretability and fairness evaluation of deep learning models on MIMIC-IV dataset. Scientific Reports. 12 (1): 7166. https://doi.org/10.1038/s41598-022-11012-2

Monaco A, Pantaleo E, Amoroso N, Lacalamita A, Lo Giudice C et al. (2021). A primer on machine learning techniques for genomic applications. Computational and Structural Biotechnology Journal. 19: 4345–4359. https://doi.org/10.1016/j.csbj.2021.07.021

Moreno M, Vilaça R, Ferreira PG (2022). Scalable transcriptomics analysis with Dask: applications in data science and machine learning. BMC Bioinformatics. 23 (1): 514. https://doi.org/10.1186/s12859-022-05065-3

Moult J, Hubbard T, Bryant SH, Fidelis K, Pedersen JT (1997). Critical assessment of methods of protein structure prediction (CASP): round II. Proteins. Suppl 1: 2–6.

Myszczynska MA, Ojamies PN, Lacoste AMB, Neil D, Saffari A et al. (2020). Applications of machine learning to diagnosis and treatment of neurodegenerative diseases. Nature Reviews Neurology. 16 (8): 440–456. https://doi.org/10.1038/s41582-020-0377-8

Namiki T, Hachiya T, Tanaka H, Sakakibara Y (2012). MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. Nucleic Acids Research. 40 (20): e155. https://doi.org/10.1093/nar/gks678

Ni A, Qin LX (2021). Performance evaluation of transcriptomics data normalization for survival risk prediction. Briefings in Bioinformatics. 22 (6): bbab257. https://doi.org/10.1093/bib/bbab257

Onimaru K, Nishimura O, Kuraku S (2020). Predicting gene regulatory regions with a convolutional neural network for processing double-strand genome sequence information. Makarenkov V, editor. Public Library of Science ONE. 15 (7): e0235748. https://doi.org/10.1371/journal.pone.0235748

Pagano TP, Loureiro RB, Lisboa FVN, Peixoto RM, Guimarães GAS et al. (2023). Bias and Unfairness in Machine Learning Models: A Systematic Review on Datasets, Tools, Fairness Metrics, and Identification and Mitigation Methods. Big Data and Cognitive Computing. 7 (1): 15. https://doi.org/10.3390/bdcc7010015

Park DJ, Park MW, Lee H, Kim YJ, Kim Y et al. (2021). Development of machine learning model for diagnostic disease prediction based on laboratory tests. Scientific Reports 11 (1): 7567. https://doi.org/10.1038/s41598-021-87171-5

Pastur Romay L, Cedrón F, Pazos A, Porto Pazos A (2016). Deep Artificial Neural Networks and Neuromorphic Chips for Big Data Analysis: Pharmaceutical and Bioinformatics Applications. International Journal of Molecular Sciences. 17 (8): 1313. https://doi.org/10.3390/ijms17081313

Petkovic D, Kobzik L, Re C (2018). Machine learning and deep analytics for biocomputing: call for better explainability. Proceedings of the Pacific Symposium Biocomputing. 23: 623–627. https://doi.org/10.1142/9789813235533_0058

Poplin R, Chang PC, Alexander D, Schwartz S, Colthurst T et al. (2018). A universal SNP and small-indel variant caller using deep neural networks. Nature Biotechnology. 36 (10): 983–987. https://doi.org/10.1038/nbt.4235

Ribeiro MT, Singh S, Guestrin C. 2016. “Why Should I Trust You?”: Explaining the Predictions of Any Classifier. https://doi.org/10.48550/ARXIV.1602.04938

Roe KD, Jawa V, Zhang X, Chute CG, Epstein JA (2020). Feature engineering with clinical expert knowledge: A case study assessment of machine learning model complexity and performance. Uzuner O, editor. Public Library of Sience ONE. 15 (4): e0231300. https://doi.org/10.1371/journal.pone.0231300

Saçar Demirci MD, Baumbach J, Allmer J (2017). On the performance of pre-microRNA detection algorithms. Nature Communications. https://doi.org/10.1038/s41467-017-00403-z

Sacar MD, Allmer J (2013). Data mining for microrna gene prediction: On the impact of class imbalance and feature number for microrna gene prediction. In: 2013 8th International Symposium on Health Informatics and Bioinformatics. Ankara, Turkey: IEEE Xplore. p. 1–6. https://doi.org/10.1109/HIBIT.2013.6661685

Sarkar D, Saha S (2019). Machine-learning techniques for the prediction of protein-protein interactions Journal of Biosciences. 44 (4): 104. https://doi.org/10.1007/s12038-019-9909.z

Sarker IH. 2021. Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions. Springer Nature Computer Science. 2 (6): 420. https://doi.org/10.1007/s42979-021-00815-1

Sarker S, Jamal L, Ahmed SF, Irtisam N (2021). Robotics and artificial intelligence in healthcare during COVID-19 pandemic: A systematic review. Robotics and Autonomous Systems. 146: 103902. https://doi.org/10.1016/j.robot.2021.103902

Sathyan A, Weinberg AI, Cohen K (2022). Interpretable AI for bio-medical applications. Complex Engineering Systems. 2 (4): 18. https://doi.org/10.20517/ces.2022.41

Savas Takan, Allmer J (2023). De Novo Sequencing of Peptides from Tandem Mass Spectra and Application in Proteogenomics. ResearchGate prepring. https://doi.org/10.13140/RG.2.2.29991.62882

Schadt EE, Linderman MD, Sorenson J, Lee L, Nolan GP (2010). Computational solutions to large-scale data management and analysis. Nature Reviews Genetics. 11 (9): 647–657. https://doi.org/10.1038/nrg2857

Seadle M, Havelka S (2023). Information science: Why it is not data science. Data and Information Management. 7 (1): 100027. https://doi.org/10.1016/j.dim.2023.100027

Seo S, Oh M, Park Y, Kim S (2018). DeepFam: deep learning based alignment-free method for protein family modeling and prediction. Bioinformatics. 34 (13): i254–i262. https://doi.org/10.1093/bioinformatics/bty275

Shen X, Jiang C, Wen Y, Li C, Lu Q (2022). A Brief Review on Deep Learning Applications in Genomic Studies. Frontiers in Systems Biology. 2: 877717. https://doi.org/10.3389/fsysb.2022.877717

Simonyan K, Vedaldi A, Zisserman A (2013). Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. https://doi.org/10.48550/ARXIV.1312.6034

Skolnick J, Gao M, Zhou H, Singh S (2021). AlphaFold 2: Why It Works and Its Implications for Understanding the Relationships of Protein Sequence, Structure, and Function. Journal of Chemical Information and Modeling. 61 (10): 4827–4831. https://doi.org/10.1021/acs.jcim.1c01114

Son JW, Hong JY, Kim Y, Kim WJ, Shin DY (2022). How Many Private Data Are Needed for Deep Learning in Lung Nodule Detection on CT Scans? A Retrospective Multicenter Study. Cancers (Basel). 14 (13): 3174. https://doi.org/10.3390/cancers14133174

Teng Q, Liu Z, Song Y, Han K, Lu Y (2022). A survey on the interpretability of deep learning in medical diagnosis. Multimedia Systems. 28 (6): 2335–2355. https://doi.org/10.1007/s00530-022-00960-4

The Precise4Q consortium, Amann J, Blasimme A, Vayena E, Frey D, Madai VI (2020). Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Medical Informatics and Decision Making. 20 (1): 310. https://doi.org/10.1186/s12911-020-01332-6

Thumuluri V, Almagro Armenteros JJ, Johansen AR, Nielsen H, Winther O (2022). DeepLoc 2.0: multi-label subcellular localization prediction using protein language models. Nucleic Acids Research. 50 (W1): W228–W234. https://doi.org/10.1093/nar/gkac278

Tran NH, Qiao R, Xin L, Chen X, Liu C (2019). Deep learning enables de novo peptide sequencing from data-independent-acquisition mass spectrometry. Nature Methods. 16 (1): 63–66. https://doi.org/10.1038/s41592-018-0260-3

Tsimring LS (2014). Noise in biology. Reports on Progress in Physics. 77 (2): 026601. https://doi.org/10.1088/0034-4885/77/2/026601

Tubiana J, Schneidman Duhovny D, Wolfson HJ (2022). ScanNet: an interpretable geometric deep learning model for structure-based protein binding site prediction. Nature Methods. 19 (6): 730–739. https://doi.org/10.1038/s41592-022-01490-7

Twilt M (2016). Precision Medicine: The new era in medicine. EBioMedicine. 4: 24–25. https://doi.org/10.1016/j.ebiom.2016.02.009

Vinuesa R, Sirmacek B (2021). Interpretable deep-learning models to help achieve the Sustainable Development Goals. Nature Machine Intelligence. 3 (11): 926–926. https://doi.org/10.1038/s42256-021-00414-y

Wójcikowski M, Zielenkiewicz P, Siedlecki P (2015). Open Drug Discovery Toolkit (ODDT): a new open-source player in the drug discovery field. Journal of Cheminformatics. 7: 26. https://doi.org/10.1186/s13321-015-0078-2

Yang K, Qinami K, Fei Fei L, Deng J, Russakovsky O (2020). Towards fairer datasets: filtering and balancing the distribution of the people subtree in the ImageNet hierarchy. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. Barcelona Spain: ACM. p. 547–558. https://doi.org/10.1145/3351095.3375709

Yazdani A, Costa S, Kroon B (2023). Artificial intelligence: Friend or foe? Annual Scientific Meeting of the Royal Australian and New Zealand College of Obstetricians and Gynaecologists. 63 (2): 127–130. https://doi.org/10.1111/ajo.13661

Yip KY, Cheng C, Gerstein M (2013). Machine learning and genome annotation: a match meant to be? Genome Biology. 14 (5): 205. https://doi.org/10.1186/gb-2013-14-5-205

Yuan Y, Shi Y, Li C, Kim J, Cai W (2016). DeepGene: an advanced cancer type classifier based on deep learning and somatic point mutations. BMC Bioinformatics. 17 (S17): 476. https://doi.org/10.1186/s12859-016-1334-9

Zhang X, Jonassen I, Goksøyr A (2021). Machine Learning Approaches for Biomarker Discovery Using Gene Expression Data. In: Department of Clinical and Toxicological Analyses, School of Pharmaceutical Sciences, University of Sao Paulo, Sao Paulo, Brazil, Nakaya HI, editors. Bioinformatics. Exon Publications. p. 53–64. https://doi.org/10.36255/exonpublications.bioinformatics.2021.ch4

Zhu Q, Xu B, Huang J, Wang H, Xu R (2023). Deep Multi-Modal Discriminative and Interpretability Network for Alzheimer's Disease Diagnosis. IEEE Transactions on Medical Imaging. 42 (5): 1472–1483. https://doi.org/10.1109/TMI.2022.3230750