Artificial Intelligence and Machine Learning in Bioinformatics 16

Shabroz Alam, Juveriya Israr, and Ajay Kumar

Abstract

Artificial intelligence (AI) and machine learning (ML) have emerged over the past decade as the cutting-edge technologies most expected to revolutionize the research and development sector. This is fueled in part by game-changing developments in computer technology and the concomitant evaporation of barriers to collecting massive amounts of data. Meanwhile, the cost of researching, testing, manufacturing, and distributing new pharmaceuticals has risen. In light of these challenges, the pharmaceutical industry is interested in AI/ML methods because to their automation, predictability, and the ensuing anticipated boost in efficiency. The use of ML techniques in the pharmaceutical industry has matured during the past 15 years. Clinical trial design, management, and analysis are the most recent drug development process steps to benefit from AI and ML. As we move toward a world in which AI/ML is increasingly integrated into R&D, it is essential to sort through the corresponding jargon and hype. Equally crucial is the understanding that the scientific method is still relevant for drawing conclusions from evidence. By doing so, we can better evaluate the potential benefits of AI/ML in the pharmaceutical industry and make well-informed decisions on their best application. The purpose of this paper is to clarify certain fundamental ideas, provide some examples of their

S. Alam
Department of Biotechnology, Era University, Lucknow, Uttar Pradesh, India

J. Israr
Department of Biotechnology, Era University, Lucknow, Uttar Pradesh, India

Institute of Biosciences and Technology, Shri Ramswaroop Memorial University, Lucknow- Deva Road, Barabanki, Uttar Pradesh, India

A. Kumar (✉)
Department of Biotechnology, Rama University, Kanpur, Uttar Pradesh, India

V. Singh, A. Kumar (eds.), *Advances in Bioinformatics*,
https://doi.org/10.1007/978-981-99-8401-5_16

application, and then provide some helpful perspective on how to best apply AI/ML techniques to research and development.

Keywords

Artificial intelligence · Machine learning · Deep learning · Drug discovery

16.1 Introduction

Machine Learning, a subset of Artificial Intelligence, develops algorithms and models to help robots learn and behave like people. Knowledge, comprehension, and competence are the focus of the study, teaching, and experience that make up the area of machine learning, which integrates computer science and statistics. Assimilation of new information leads to a dynamic shift in behavior (Alpaydin 2020).

Machine learning, a bioinformatics field, transforms computing systems to do complex AI-like processes. The above bioinformatics activities include pattern recognition, disease diagnostics, computational planning, robotic control systems, and predictive modeling. The "alterations" may include system enhancements or new system building (Chetty et al. 2022).

In recent years, medical oncology has gained a remarkable understanding of cancer biology and pathogenesis. Bioinformatics has improved our ability to study and model complex biological processes thanks to next-generation sequencing technologies, particularly single-cell RNA sequencing. This includes incredibly deep and exact research and characterization of complicated issues like cancer heterogeneity, resistance mechanisms, and illness causation. In addition, collaborative efforts and extensive projects in bio specimen collection and bioinformatics, such as The Cancer Genome Atlas (TCGA), have helped consolidate, organize, and examine an unprecedented volume of patient data. This has led to the identification of novel therapeutic targets and the examination of established targets in previously unexplored illness contexts (Alpaydin 2020).

Despite the growth of cancer biology, drug discovery still faces several hurdles. Despite high-throughput screening technology, development timetables and expenses are long and expensive. Bringing a pharmaceutical molecule to market takes years, usually a decade. This complex procedure requires enormous R&D and financial investments of over $2.8 billion. Suboptimal pharmacokinetics, toxicity, and clinical efficacy can cause candidate medication failure in the drug development pipeline (Gupta et al. 2021).

In bioinformatics, using pre-existing medications to treat new diseases is a promising way to overcome the challenges of drug development for novel chemicals. To enter the market, approved pharmaceuticals have passed rigorous clinical trials, including preclinical studies, human testing, and careful evaluation. Therefore, these medications have a well-known safety profile. Bioinformatics can greatly benefit from discovering a new clinical indication for an approved medicine. This fascinating idea allows the medicine to re-enter Phase II clinical trials. This strategy reduces research and development risks and time and money expenses (Vamathevan et al. 2019).

The extensive use of computational algorithms spanning a variety of methodologies and approaches has advanced medication repurposing research in recent years. The structural biology of therapeutic protein targets can be fully explored using molecular modeling. It also enables high-throughput virtual screenings, which identify interesting drug candidates with therapeutic potential. Bioinformatics has advanced rapidly because of advances in machine learning and artificial intelligence, particularly in deep learning (Nosi et al. 2021). These cutting-edge technologies have transformed our understanding of drug-target interactions and the complex link between drug physicochemical features and phenotypic changes. These methods also help find new cancer targets in the vast cancer data repositories accumulated via many joint efforts. Due to the extensive use of high-throughput and multi-omics drug profiling experiments, chemical and bioactivity data is growing, making bioinformatics crucial to cancer treatment discovery. Additionally, the increased accessibility of these publicly available dataset collections considerably improves computational techniques (Min et al. 2017). These methods can be used for more than only experimental and biological data. Bioinformatics benefits from clinical dataset integration, notably electronic health records. This in-depth chapter discusses state-of-the-art computational techniques for oncology drug repurposing. Machine learning and deep neural networks are highlighted.

16.2 Artificial Intelligence

Machine learning is a subfield of artificial intelligence (AI). Academic interest in machine learning from data dates back to the earliest days of artificial intelligence. They tried to solve it using a wide range of symbolic techniques, including “neural networks” (primarily perceptron’s and related models, which were later shown to be statistical generalized linear re-imaginings). Automated medical diagnosis, in particular, made extensive use of probabilistic reasoning (Sarle Warren 1994).

However, a divide between AI and machine learning was produced by an increased focus on the logical, knowledge-based approach. Issues with data collection and representation, both theoretical and practical, afflicted probabilistic systems. By 1980, expert systems had supplanted statistics as the dominant approach to artificial intelligence. While research into symbolic/knowledge-based learning and its offshoot, inductive logic programming, continued inside AI, work along a more statistical line of inquiry moved out of AI and into pattern recognition and information retrieval. Around the same time, artificial intelligence and computer science ceased their investigation into neural networks. Hopfield, Rumelhart, and Hinton, who had previously worked in artificial intelligence and computer science, went on to develop this line of thought as “connectionism” in their new fields of study. In the mid-1980s, when they rediscovered backpropagation, they saw their greatest success (Stuart and Peter 2003).

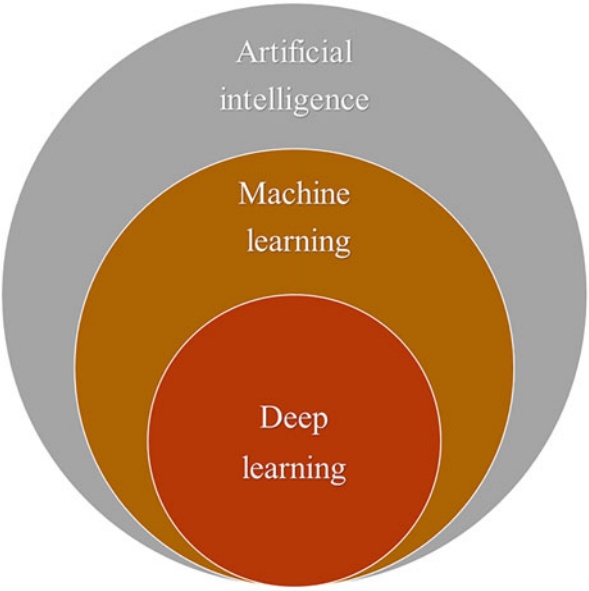


Fig. 16.1 The subfield of artificial intelligence that is known as machine learning

In the 1990s, machine learning (ML) began to flourish as a distinct discipline. The field shifted its focus from developing artificial intelligence to solving real-world issues. It abandoned the symbolic methodologies it had received from AI in favor of statistical, fuzzy logic, and probability theory-based procedures and models in Fig. 16.1 (Langley 2011).

16.3 Importance of Machine Learning

In the realm of bioinformatics, certain computational challenges elude precise definition, save for the provision of illustrative instances. These instances may consist of well-defined input/output pairs, while the connection between what is put in and what comes out remains elusive to articulate succinctly. The objective is to enable machines to dynamically adapt their internal configuration, allowing them to generate accurate outputs for a vast array of sample inputs. This process aims to effectively restrict their input/output mechanism, thereby approximating the underlying relationship inherent in the provided examples.

In the vast expanse of data, lies the potential for unearthing concealed connections and intricate correlations. Machine learning techniques, commonly employed in the field of bioinformatics, have proven to be highly effective in extracting intricate relationships from complex datasets, a process commonly referred to as data mining (Ngiam and Khor 2019).

The individual in question possesses a keen interest in the field of bioinformatics, a discipline that combines the phenomenon of human designers frequently encountering challenges in achieving optimal performance of machines within their designated environments is a well-documented observation. In reality, the comprehensive understanding of all aspects of the working environment may not be fully ascertainable during the initial design phase. When it comes to bioinformatics, machine learning applications are becoming increasingly popular. It has demonstrated its potential for enhancing the performance and optimization of existing machine designs (Mohsen et al. 2021).

The user has provided a brief statement. In the area of bioinformatics, the vast expanse of knowledge pertaining to specific tasks often exceeds the capacity for direct human encoding, the potential for machines to acquire knowledge incrementally holds great promise in surpassing the limitations of human documentation. These intelligent systems have the capacity to assimilate a wealth of information that may surpass the extent to which humans are inclined to transcribe (Erickson 2021).

The individual in question has a keen interest in the field of bioinformatics. They possess a deep understanding Environmental conditions undergo dynamic transformations throughout the course of temporal progression. The development of adaptable machines capable of dynamically responding to environmental changes holds great potential in mitigating the necessity for recurrent redesign efforts.

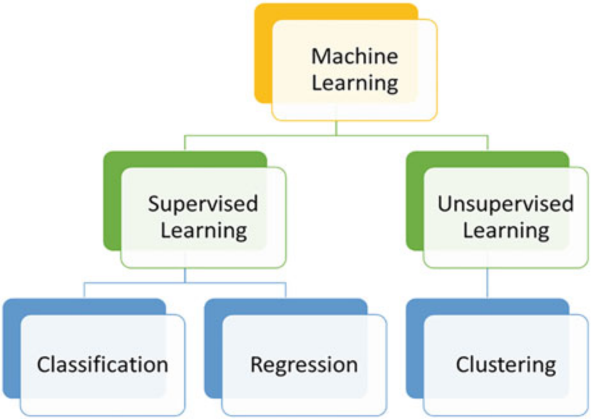
Humans are perpetually unearthing novel insights pertaining to various tasks. The user's text will be transformed to incorporate bioinformatics terminology and vocabulary. The ever-evolving landscape of global affairs presents a perpetual influx of novel occurrences. The ongoing endeavor to reengineer artificial intelligence (AI) systems in accordance with emerging insights presents inherent challenges. However, leveraging the potential of machine learning techniques holds promise in effectively monitoring and assimilating a substantial portion of this evolving knowledge landscape (Munjal et al. 2023).

16.4 Types of Machine Learning

Machine learning, a subfield of bioinformatics, encompasses a wide range of computational techniques that enable the analysis and interpretation of complex biological data. While classification is indeed a fundamental aspect of machine learning, it is important to recognize that this field extends far beyond this single task. By leveraging advanced algorithms and statistical models, machine learning enables researchers to uncover in the field of bioinformatics, a diverse range of problem classes can be identified (Fig. 16.2). These problem classes serve as the foundation for addressing various biological and computational challenges.

1. **Classification learning**: an essential task in bioinformatics, where the goal is to acquire the ability to accurately assign instances to predetermined classes. This process involves the utilization of various computational algorithms and statistical techniques to train models that can effectively distinguish between different classes based on specific features or attributes. Classification learning's ability to harness the power of machine learning is essential in many bioinformatics

Fig. 16.2 Types of machine learning



applications, including those for analyzing gene expression, predicting protein function, and diagnosing disease (Medin and Schaffer 1978).

2. Through in the realm of bioinformatics, **association learning** is a fundamental concept that involves the acquisition of knowledge regarding the intricate relationships that exist between various attributes. Through meticulous analysis and exploration, researchers strive to uncover and comprehend the intricate connections and dependencies that may exist within biological datasets. By employing sophisticated algorithms and statistical techniques, association learning enables the identification of significant associations and patterns.
3. Thereby shedding light on **Clustering:** Uncovering cohesive groups of instances that exhibit similar characteristics (Karim et al. 2021).
4. In the realm of bioinformatics, one fascinating area of study involves the task of **numeric prediction**. Rather than focusing on classifying data into distinct categories, this branch of research delves into the realm of forecasting numeric quantities. By employing sophisticated algorithms and machine learning techniques, scientists and researchers strive to develop models that can accurately predict numerical values associated with various biological phenomena. This research has far-reaching ramifications in areas like genetics, proteomics, and drug development, and holds tremendous promise for enhancing our knowledge of complex biological systems through the utilization of vast datasets and cutting-edge computational methodologies.

16.5 Supervised and Unsupervised Learning

Supervised learning, a fundamental concept in bioinformatics, refers to the learning process wherein training instances are meticulously annotated with the correct outcomes. This meticulous labelling enables the system to receive valuable feedback, facilitating an understanding of the progress made in the learning journey. In the area of unsupervised learning, the objective becomes more challenging as it necessitates the absence of predetermined categorizations (Goudbeek et al. 2009).

16.5.1 Supervised Learning

In bioinformatics, supervised learning is often used in categorization issues. The main goal is to teach computational systems categorization systems. Training neural networks and decision trees relies heavily on supervised learning, a common bioinformatics technique. Both computational approaches use predetermined classification data substantially. Classification helps neural networks measure inaccuracy and fine-tune parameters to reduce discrepancies. In decision trees, classifications help identify attributes with the most informational value, solving complex classification problems (Le et al. 2020).

Supervised learning is essential in bioinformatics. These methods can be used to create prediction models that can identify patterns and relationships in input and output data. These models can learn from the dataset and accurately anticipate unseen variables by carefully analyzing the available data. This method has great potential in bioinformatics applications, helping researchers understand complicated biological events and living organisms (Chen and Gao 2016).

16.5.2 Unsupervised Learning

Unsupervised learning is a difficult bioinformatics activity that trains computers to learn and accomplish tasks without explicit instructions. The goal is to let machines learn and do tasks without human involvement. In computational biology and bioinformatics, unsupervised learning has two ways. The initial technique instructs the agent via rewards rather than explicit categorizations to indicate achievement. Clustering is a popular unsupervised learning paradigm and a second bioinformatics computational method. In computational biology, this learning paradigm seeks to identify patterns and resemblances in the training dataset rather than optimize a utility function. The clusters identified are expected to match an intuitive categorization. Demographic clustering can divide people into two groups: affluent and impoverished (Goudbeek et al. 2009).

Unsupervised learning in bioinformatics groups and interprets data based on input data. This approach explores underlying data patterns and structures without labels or annotations. Unsupervised learning algorithms use algorithms and statistics to get insights from unannotated datasets. This method is essential for clustering analysis, dimensionality analysis, and other bioinformatics applications (Chen and Gao 2016).

16.5.3 Semi Supervised Learning

In the realm of bioinformatics, semi-supervised learning is a computational approach that combines features of supervised and unsupervised learning. The dataset under examination is a hybrid of unannotated and annotated data, including a wide variety of sources. The fundamental objective of this research is to create a computational technique that can reliably predict output values for inputs that are either poorly described or for which no outputs are available. There is a little amount of labeled data and a huge amount of unlabeled data in the given database. In addition to the well-established paradigms of supervised and unsupervised learning, the field of bioinformatics encompasses a diverse array of learning algorithms, including reinforcement learning, among others. Both supervised and unsupervised learning methods have gained significant popularity and are extensively utilized in various domains, including computational biology and pattern recognition. These approaches play a crucial role in real-world applications, facilitating advancements in diverse fields (Yan and Wang 2022).

16.5.4 Reinforcement Learning

The field of reinforcement learning is a subset of machine learning that seeks to create intelligent decision-making algorithms and models via trial and error. The algorithms employed in this context are specifically designed to identify an optimal policy that effectively maps various states of the world to corresponding actions. The selection of actions is determined from a set of available options that an agent is expected to undertake based on the prevailing states, with the ultimate objective of optimizing a measure of cumulative reward over an extended period. Bioinformatics has revolutionized the field of machine learning by introducing a novel approach that sets it apart from traditional methods. One of its key differentiating factors lies in its ability to leverage biological data to drive predictive models and uncover hidden patterns. This distinctive characteristic has propelled bioinformatics to the forefront of cutting-edge research, enabling scientists to tackle complex problems in diverse domains such as genomics, proteomics, and drug discovery. By harnessing the power of biological information, bioinformatics has opened up new avenues for understanding and manipulating biological systems, paving the way for ground breaking advancements in the field of the absence of input-output pairs within a database characterizes this system, which is primarily designed to optimize online performance (Weltz et al. 2022; Liu et al. 2021).

16.5.5 Optimization

Optimization, a fundamental concept in bioinformatics, plays a crucial role in the field's pursuit of identifying the most optimal solution within a vast array of potential solutions. In the realm of bioinformatics, the pursuit of knowledge through data analysis is akin to a quest for the most suitable model that accurately captures the intricacies of the data. Consequently, the utilization of optimization techniques becomes an integral component in the process of constructing these models. In the past decade, there has been a significant proliferation of both exact and heuristic optimization algorithms across various domains.

16.5.6 Machine Learning and Statistics

In the field of bioinformatics, statistical analysis plays a crucial role in hypothesis testing, allowing researchers to assess the significance of their findings. Conversely, machine learning approaches in bioinformatics focus on the development of algorithms that facilitate the process of generalization by exploring and evaluating various hypotheses. By leveraging computational power, machine learning techniques aid in the discovery of patterns and relationships within complex biological datasets, enabling researchers to make informed predictions and decisions. Statistics is a multifaceted discipline that extends beyond the realm of hypothesis testing. In the realm of bioinformatics, it plays a crucial role in analyzing and interpreting complex biological data. Moreover, it is worth noting that numerous machine learning methodologies exist that do not rely on traditional search algorithms. These techniques leverage sophisticated computational models to uncover patterns and make predictions, thereby enhancing our understanding of biological systems. Machine learning algorithms commonly employ statistical tests during the construction of rules or trees, as well as for the purpose of rectifying models that exhibit "overfitting" tendencies. Overfitting occurs when models excessively rely on specific examples utilized during their creation, leading to a lack of generalizability. Statistical tests play a crucial role in the realm of bioinformatics by serving as a means to validate and evaluate machine learning models and algorithms. These tests enable researchers to assess the performance and reliability of such computational tools, ensuring their efficacy in addressing complex biological problems. Through rigorous statistical analysis, bioinformaticians can confidently determine the accuracy, precision, and generalizability of machine learning approaches, thereby facilitating their integration into various biological research domains (Venkatesh et al. 2020).

16.6 Selecting the Right Algorithm

In the field of bioinformatics, the task of algorithm selection can be a daunting endeavor. With a multitude of both supervised and unsupervised machine learning algorithms at one's disposal, each algorithm exhibits a unique methodology for acquiring knowledge. In the field of bioinformatics, it is widely acknowledged that the absence of a universally optimal approach or a one-size-fits-all solution is a prevailing reality. The process of identifying the optimal algorithm involves a combination of empirical exploration and meticulous analysis. Even seasoned bioinformaticians acknowledge that the efficacy of an algorithm cannot be ascertained a priori, necessitating iterative experimentation. In the field of bioinformatics, it is widely acknowledged that models exhibiting a high degree of flexibility possess the inherent risk of succumbing to overfitting. This phenomenon occurs when such models, in their quest to capture intricate patterns and nuances within the data, inadvertently incorporate even the minutest variations that may potentially be attributed to mere noise. In the field of bioinformatics, it is widely acknowledged that the interpretability of models is inversely proportional to their complexity. Consequently, simpler models tend to offer a more straightforward understanding of the underlying biological phenomena. However, it is important to note that this simplicity often comes at the cost of reduced accuracy. The selection of an appropriate algorithm necessitates a careful consideration of various factors, wherein the trade-offs between different advantages come into play. These considerations encompass crucial aspects such as the computational efficiency, precision, and intricacy of the model at hand. The iterative process of experimentation and algorithmic exploration lies at the heart of machine learning, wherein the pursuit of optimal solutions necessitates the continuous evaluation and refinement of various approaches.

16.6.1 Machine Algorithms in Omics Field

In the ever-expanding state of bioinformatics, the imperative to remain at the forefront is twofold: to seamlessly assimilate burgeoning data and to continuously advance algorithmic methodologies. In the field of bioinformatics, the integration of machine learning (ML) algorithms has become indispensable for conducting predictive analytics and unravelling the intricate biological mechanisms inherent in the human body. The adoption of machine learning techniques has improved some difficult areas of bioinformatics. Genomics, proteomics, microarrays, systems biology, evolutionary biology, and text mining are all examples of these disciplines (Li et al. 2022; Perakakis et al. 2018).

16.6.2 Genomics

The burgeoning demand for the advancement of machine learning algorithms designed to autonomously identify the precise genomic coordinates of protein-coding genes within a provided DNA sequence has become increasingly evident. The issue at hand pertains to the field of computational biology, specifically gene prediction. Machine learning techniques have been effectively employed in the realm of bioinformatics to address the intricate task of multiple sequence alignment. This intricate process entails the alignment of numerous DNA or amino acid sequences, with the aim of identifying regions of similarity that may signify a common evolutionary lineage. Bioinformatics is a powerful tool that finds utility not only in the identification and visualization of genome rearrangements, but also in a myriad of other applications (Libbrecht and Noble 2015; Esposito et al. 2019).

16.6.3 Proteomics

A novel bioinformatics method classifies amino acids in a protein sequence into their structural classes using machine learning methods. Helix, sheet, and coil structural motifs can be accurately identified using this novel method. This ground breaking technology revolutionizes protein analysis by using machine learning to reveal the complex link between amino acid content and protein structure. For secondary structure prediction in bioinformatics, Deep CNF is the latest method. This advanced method uses artificial neural networks, a machine learning model, to achieve 84% accuracy. Theoretical studies estimate that three-state protein secondary structure occurs around 88–90%. Machine learning has solved complex proteomics problems. These include protein side-chain prediction, loop modeling, and contact map estimate (Mou et al. 2022; Kelchtermans et al. 2014).

16.6.4 Microarrays

One of the primary challenges encountered in the area of bioinformatics revolves around the discernment of gene expression patterns through the analysis of gathered data. Moreover, owing to the vast multitude of genes encompassed in the microarray dataset, a substantial volume of extraneous data is present, thereby exacerbating the intricacy of the expressed gene identification task. Machine learning, a cutting-edge field at the intersection of computer science and biology, offers a promising avenue to address this challenge. Leveraging a diverse range of classification techniques, machine learning algorithms can be harnessed to effectively carry out the task of identification in question. In the realm of bioinformatics, a plethora of methodologies has emerged as prominent tools for data analysis and pattern recognition. Radial basis function networks, deep learning methods, Bayesian classification, decision trees, and random forest models are popular. These methods, renowned for their versatility and efficacy, have proven instrumental in unravelling complex biological phenomena and extracting meaningful insights from vast datasets. By leveraging the power of these computational approaches, researchers in the field of bioinformatics are able to navigate the intricacies of biological systems and make significant strides towards advancing our understanding of life’s fundamental processes (Ekins and Chu 1999; Pirooznia et al. 2008).

16.6.5 Systems Biology

Machine learning has made computational modeling complex biological system interactions easier. This is notably the case in the context of metabolic pathways, signal transduction pathways, and genetic networks. Probabilistic graphical models, a popular bioinformatics computational framework, can reveal complex variable interactions. These methods use machine learning to untangle genomic networks’ complicated structure. Probabilistic graphical models have become a standard tool for modeling genetic networks, enabling extensive studies of biological systems' mechanisms. Complex systems biology issues have also been addressed by machine learning in the bioinformatics community. Locating binding sites for transcription factors is crucial for controlling gene expression. The intricate patterns of these binding sites can be revealed by using machine learning methods in conjunction with Markov chain optimization. Natural selection-based genetic algorithms have found widespread usage in simulating biological regulation and control networks. These methods employ machine learning to recreate the interactions between genetic elements, illuminating the complex dynamics of biological systems (Muggleton 2005).

Machine learning in systems biology is one of several bioinformatics applications. Machine learning methods are used to predict enzyme function based on molecular characteristics. Machine learning is also used to analyze high-throughput microarray data, allowing researchers to gain insights from massive genetic data. Genome-wide association studies use machine learning methods to reveal complex genetic marker-disease susceptibility correlations. Last but not least, machine learning helps identify and characterize proteins based on their structural and functional properties. These applications demonstrate how machine learning improves our understanding of complicated biological processes (Liu et al. 2013).

16.6.6 Text Mining

The utilization of machine learning in the field of bioinformatics has paved the way for efficient knowledge extraction methodologies. By employing modern methods like natural language processing, valuable insights can be extracted from vast repositories of human-generated reports stored within databases. The utilization of this methodology has been extensively employed in the pursuit of discovering innovative pharmaceutical targets. This endeavor necessitates the meticulous scrutiny of data repositories encompassing biological databases and scholarly publications. Protein databases frequently lack comprehensive annotations that encompass the entirety of available knowledge for each protein. Consequently, it becomes necessary to extract supplementary information from the vast pool of biomedical literature. The application of machine learning techniques has revolutionized the field of bioinformatics by enabling automated annotation of gene and protein functions, prediction of subcellular localization of proteins, analysis of DNA-expression arrays, exploration of large-scale protein interaction networks, and investigation of molecular interactions. Text mining has emerged as a valuable tool in the realm of bioinformatics, with diverse applications including the identification and graphical representation of unique DNA regions, provided an ample amount of reference data is available (Mohsen et al. 2021).

16.7 Commonly Used Machine Learning Algorithms in Bioinformatics

In the field of bioinformatics, some of the most commonly used learning algorithms are Support Vector Machines, Linear Regression, Logistic Regression, Naive Bayes, Linear Discriminant Analysis, Decision Trees, K-Nearest Neighbor Algorithm, and Neural Networks (especially Multilayer Perception).

16.7.1 Decision Tree Classifier

Decision tree classifiers are extensively employed in the field of bioinformatics due to their numerous advantageous features. These classifiers are highly favored for their simplicity, efficiency, and effectiveness in analyzing complex biological data. Moreover, their ability to provide visually intuitive graphical representations further enhances their utility in bioinformatics research and analysis. The decision tree model is constructed using a recursive top-down approach, a widely employed methodology in bioinformatics. This approach facilitates the creation of a model that is both comprehensible and verifiable, making it highly suitable for analysis and interpretation. In the field of bioinformatics, a decision tree is a widely used computational model for classification and regression analysis. It consists of nodes that represent various features or attributes, with the topmost node referred to as the root. The remaining nodes within the tree structure are known as internal nodes, which aid in the decision-making process by evaluating different criteria and branching out accordingly. The construction of the tree follows a recursive approach, starting from the root node and considering each feature individually. Each node in the tree represents an input parameter, allowing for a systematic evaluation of the data. The sample is partitioned through the iterative process of posing recursive inquiries. The terminal node, also known as the leaf node, serves as the final prediction node in the bioinformatics analysis (Charbuty and Abdulazeez 2021; Navada et al. 2011).

16.7.2 Naïve Bayes Classifier

In bioinformatics, classification tasks are often handled using the Naive Bayes classifier, a machine learning method. It functions on the premise that the parameters employed in classification are not reliant on one another, which is to say that it operates on the assumption of feature independence. Since this assumption simplifies the computation of probabilities and reduces the computational cost of the algorithm, it permits efficient and successful categorization. The Naive Bayes classifier is useful in a wide variety of bioinformatics applications due to its ability to reliably categorize data points based on their feature values by exploiting the independence assumption. A common probabilistic machine learning approach in bioinformatics is the Naive Bayes classifier. Assuming that the features are conditionally independent given the class label, Bayes' theorem provides a method for classifying data. Equation 16.1 is a mathematical representation of the classifier that captures its core functional principles.

$$\mathbf{P}\ (\mathbf{C1}|\mathbf{P1},\mathbf{P2}) = \mathbf{P}(\mathbf{P1}|\mathbf{C1})\ \mathbf{P}(\mathbf{P2}|\mathbf{C2}).\mathbf{P}(\mathbf{C1})/\mathbf{P}(\mathbf{P1})\ \mathbf{P}(\mathbf{P2}) \tag{16.1}$$

The probability that the input will fall into class C1 can be calculated using Eq. (16.1), using the parameters P1 and P2.

The conditional probability of observing event C1 given events P1 and P2 can be expressed as the expression (16.1) represents the conditional probability of event C1 given events P1 and P2, divided by the joint probability of events. Equation (16.1) provides the probabilistic assessment of the input's membership in class C1, utilizing the parameters P1 and P2. The probability of obtaining class C1, given parameters P1 and P2, can be expressed as the ratio between the product of the probabilities of P1 occurring with class C1 and P2 occurring with class C2, and the product of the probabilities of P1 and P2 occurring. The utilization of the Bayes formula is evident in this context (Berrar 2018; Saritas and Yasar 2019).

16.7.3 Support Vector Machines

One of the most popular classification approaches in modern bioinformatics is practiced by this person, who is considered an authority in the field. It has risen to the top as a favorite amongst industry professionals thanks to its solid computational base and outstanding accuracy in a wide range of practical applications. Classification of data points is made possible in bioinformatics with the use of Support Vector Machines (SVMs), which work by projecting them into a higher dimensional space. By using this transformation, we may generate a hyperplane that cleanly demarcates between several types of situations. SVMs reliably identify new instances by finding the hyperplane that minimizes the distance to the nearest data points of each class. Building two extra parallel hyperplanes, one on each side of the initial hyperplane, is what is meant by the proposed method. Finding the hyperplane that optimizes the gap between two parallel hyperplanes is the goal of the support vector machine (SVM) method. It is hypothesized that increasing the distance between these hyperplanes will improve the classifier's ability to forecast. Large portions of this domain's division appear to be controlled by two tests that are almost coincident with parallel hyperplanes. Support vectors is a term that is frequently used to describe these cases in the field of bioinformatics. Because of the difficulty in correctly categorizing them, these samples are notoriously difficult to study in the field of bioinformatics. In bioinformatics, it might be difficult to accurately and completely separate training points into their respective classes. These incorrectly classified locations cannot be located too far from the partition zone's outermost boundary. Since support vector machines (SVMs) are so effective at classifying data and addressing a wide range of computational problems, they have become increasingly prominent in the field of bioinformatics. However, they have been criticized for not being sufficiently expressive and understandable in terms of the mathematics they employ (Meyer and Wien 2001; Burbidge et al. 2001).

16.8 Commonly Used Unsupervised Machine Learning Algorithms

16.8.1 Partitional Clustering

This family of clustering algorithms uses a strategy in which each sample is placed into a unique cluster, creating a division in the data set. The user must decide ahead of time how many groups should be created in the dataset before applying a partitional clustering technique. Despite the availability of a number of heuristic approaches in bioinformatics, determining the appropriate cluster size remains a persistent problem. In bioinformatics, the k-implies computation is a standard, go-to method for partitional cluster analysis. This computational method seeks to reduce the sum of squares for each cluster of tests by grouping them into K distinct clusters. At its core, the algorithm relies on the transformational interplay of two fundamental and expedient processes in the realm of bioinformatics. Before the initiation of the sequential progression of these two distinct phases, a preliminary assessment involving a series of examinations is conducted on K initial clusters. During the initial phase, the provided examples are assigned to specific groups based on their proximity to the centroid, typically determined by the Euclidean distance. During the subsequent iteration, the recalibration of group centroids is performed as part of the algorithmic process. The culmination of the dual phases is terminated upon the cessation of protest development, as an alternative assemblage shall diminish the aggregate count of internal blocks. The author explores various computational approaches in the field of bioinformatics, with a specific focus on optimizing the efficiency of K implies calculation. The study aims to enhance processing times, thereby improving the overall performance of high jumper sity. The main limitation of this approach lies in its inability to consistently produce identical results across different runs, as the final configuration of clusters is contingent upon the initial random assignment of points to K initial clusters. In the context of bioinformatics, fluffy and probabilistic clustering methods are employed to analyze and classify biological data sets. These methods allow for a more nuanced approach to clustering, as they do not enforce strict membership of examples to a single cluster. Instead, they consider the likelihood or probability of an example belonging to each cluster, allowing for a more flexible and probabilistic assignment. This approach recognizes the inherent complexity and uncertainty in biological data, enabling a more comprehensive understanding of the underlying patterns and relationships within the data set. Through the utilization of these bioinformatics methodologies, each data point possesses a distinct degree of membership within the various clusters. Driven by the principle of reducing intracluster variation, the aforementioned composition showcases captivating methodologies in the realm of fluffy and probabilistic clustering. The domain remains ripe with untapped prospects for further dissemination endeavors (Celebi 2014; Sonagara and Badheka 2014).

16.8.2 Hierarchical Clustering

The idea of clustering presented here is widely employed in the field of bioinformatics. The output of hierarchical clustering algorithms is a dendrogram, or stable and progressive tree structure, in which the lowest level represents individual samples and the highest level represents a cluster containing all elements. Agglomerative approaches typically used in bioinformatics start at the root of the tree and work their way up. Though also used in this context, disruptive algorithms tend to cluster around the optimal starting point. Agglomerative methods are used to construct dendrograms in bioinformatics by combining clusters based on individual occurrences. Difficult techniques typically don't have a lot of ties between them because of their inefficiency. The expert can strategically cut the dendrogram at a particular level to partition a segment into a desired number of disjoint groups due to its simplicity and intuitiveness. Hierarchical clustering in bioinformatics has been made easier by the ability to choose which clusters to consider. In bioinformatics, a difference grid controls the complex agglomerative combining process. This procedure of merging bunches uses the difference grid to guide each step. The difference grid helps this sophisticated bioinformatics technique run smoothly by separating these sets. Scientific literature offers many clustering analysis separation metrics. Several bioinformatics clustering analysis methods are well-known. Single-linkage measures the distance between two groups' closest people. Complete linkage, which defines distance between two groups as the maximum distance between any two points inside each group, is another popular metric. However, Ward's progressive clustering technique merges the two groups with the lowest increase in the total within-group sum of squares at each algorithm stage. Commonly used centroid distance measures the distance between cluster centroids. Bioinformatics clustering techniques also use the median distance and group average linkage, which calculate the average dissimilarity between all pairs of individuals, one from each group (Nunez-Iglesias et al. 2013; Contreras and Murtagh 2015).

16.9 Open Source Machine Learning Software Tools

16.9.1 Weka 3: Machine Learning Software in Java

Weka uses advanced machine learning methods to solve complicated data mining problems. The bioinformatics toolset includes data preparation, predictive modeling, pattern identification, data grouping, knowledge finding, and data representation.

Open-source The Weka software is available for use under the GNU Public License. A popular bioinformatics application, it offers machine learning algorithms and data mining methods. Weka is famous among bioinformatics researchers and practitioners because to its user-friendly interface and vast capability. Weka's adaptable and customized platform lets users study and interpret complicated biological data, advancing bioinformatics research (Bouckaert et al. 2010).

A carefully designed set of free online courses in machine learning and data mining uses the powerful Weka software suite as the main teaching tool. The classes' multimedia content is available on YouTube.

Popular open-source machine learning program Weka supports deep learning. This feature lets Weka customers employ neural networks and other deep learning algorithms. Integrating deep learning (Frank et al. 2010).

16.9.2 The R Project for Statistical Computing

The R Core Team and the Foundation for Statistical Computing advocate for the use of R, a high-level programming language for statistical computing and graphical representation. Legends in the fields of bioinformatics and computational biology include Ross Ihaka and Robert Gentleman. They are famous for their ground breaking contributions to R, a high-level language and software environment for data processing and statistical modeling in bioinformatics. By revolutionizing data analysis and the development of statistical software, Ihaka and Gentleman have pushed bioinformatics forward. The fields of bioinformatics, data mining, and statistics all benefit from this potent resource. R, a sophisticated programming language and software environment, has many extension packages with reusable code and extensive documentation. Bioinformaticians and researchers use these tools to rapidly analyze and interpret complicated biological data. These extensions enable data manipulation, statistical analysis, visualization, and machine learning. R uses bioinformatics community expertise (Persson Hoden et al. 2021).

User polls and scholarly literature database analysis show that R, a popular programming language, dominates data mining. R, a bioinformatics programming language, ranks 16th in the TIOBE index as of April 2023. It dropped somewhat from 8th in August 2020. Bioinformaticians like R for its versatility and wide selection of biological data analysis tools, as well as its statistical computation and graphical capabilities (Ripley 2001).

R, developed by the GNU Project, is open-source and free under the GNU General Public License. The software framework uses C, FORTRAN, and R, with partial self-hosting. Many bioinformatics operating systems offer precompiled executables. These expert-crafted executables are essential for biological data computational analyses and simulations. By harness R, a strong and adaptable programming language, has a command line interface (CLI) for easy software interaction. This CLI lets users perform R scripts and instructions from the terminal, making data analysis, statistical modeling, and visualization easy and efficient. The bioinformatics community values third-party GUIs like RStudio, an IDE, and Jupyter, a notebook interface (Tierney 2012).

16.9.3 Bioconductor

Bioconductor is an esteemed and revolutionary software project that operates under the principles of freedom, openness, and collaborative development. It is specifically designed to facilitate the intricate analysis and comprehensive understanding of genomic data derived from wet lab experiments in the field of molecular biology. Bioconductor, a prominent bioinformatics platform, is predominantly built upon the robust statistical capabilities of the R programming language. However, it also encompasses valuable contributions from various other programming languages, augmenting its versatility and functionality. The software exhibits a biannual release pattern, synchronizing with the semi-annual updates of the R programming language. In the realm of bioinformatics, a dynamic ecosystem exists where two distinct versions coexist harmoniously. The first is the release version, meticulously aligned with the currently unleashed iteration of the esteemed R programming language. The second is the development version, intricately intertwined with the ongoing evolution of R, as it progresses towards its forthcoming manifestation. The majority of users will discover that the release version is well-suited to fulfil their requirements in the realm of bioinformatics. Furthermore, a plethora of genome annotation packages exists, primarily designed for various microarray applications, although not exclusively limited to such (Gentleman et al. 2004; Reimers and Carey 2006).

16.9.4 RapidMiner

RapidMiner, an innovative bioinformatics tool, uses a client/server design for data analysis and processing. Users can access RapidMiner's sophisticated features and capabilities through a server infrastructure housed on-premises or in public or private clouds. This flexible deployment option lets academics and scientists easily use RapidMiner's broad set of tools and resources for bioinformatics study (Kotu and Deshpande 2014).

RapidMiner is state-of-the-art bioinformatics software that provides an extensive suite of data mining and machine learning techniques. Data loading and transformation (ETL), data pre-treatment and visualization, predictive analytics and statistical modeling, comprehensive review, and rapid deployment are just some of the areas in which it shines. Using bioinformatics, scientists are able to gain new insights with the help of RapidMiner. RapidMiner, a popular data mining and machine learning package, uses Java. One of the most sophisticated bioinformatics tools, RapidMiner, has a simple graphical interface for designing and running complex analytical workflows. RapidMiner "Processes" are collections of "Operators" that perform computational tasks. Bioinformatics operators are carefully built to do a certain duty in the complex process. Each operator's result feeds the next, accelerating workflow. External software applications or APIs can call the engine. The command line interface supports individual function execution. The comprehensive bioinformatics program RapidMiner includes a variety of learning techniques, models, and algorithms for data analysis and interpretation. It integrates well with R and Python, allowing users to add own scripts. RapidMiner, a comprehensive data science platform, can integrate several plugins from the RapidMiner Marketplace to expand functionality. The RapidMiner Marketplace allows developers to carefully create and share powerful data analysis algorithms with the dynamic and collaborative data enthusiast community.

The RapidMiner Studio Free Edition bioinformatics software helps computational biologists analyze and interpret data. Following open-source development principles, this edition is licensed under AGPL. One logical processor may handle up to 10,000 data rows, making bioinformatics data manipulation and exploration efficient (Hofmann and Klinkenberg 2016).

16.9.5 Orange

Bioinformatics-specific Orange is cutting-edge, modular software. Using data visualization, machine learning, data mining, and analysis, Orange aids researchers and scientists in gaining insights from large biological datasets. Users may quickly and effectively integrate several data sources and algorithms into complex processes and pipelines because to its straightforward visual programming interface. Through the analysis of molecular networks, the prediction of protein 3D structures, and the identification of genetic relationships, Orange contributes to the unraveling of life’s secrets.

“Orange components” are like widgets in the world of bioinformatics. Data visualization, subset selection, preprocessing, experimental evaluation of learning methods, and predictive modeling are all examples of what fall under the umbrella of bioinformatics.

In bioinformatics, “visual programming” refers to the use of an interface for the connection of pre-existing or user-created widgets in order to design workflows. Python experts can use Orange as a library to modify data and interface components (Demšar et al. 2013).

16.10 Applications of Machine Learning in Bioinformatics

16.10.1 Facilitating Gene Editing Experiments

Gene editing, a revolutionary bioinformatics approach, involves complex genomic changes. Specific DNA segments are deleted, inserted, and replaced during these alterations. Gene editing allows scientists to comprehend and manipulate life’s fundamental building elements in new ways. Bioinformatics analysis relies on CRISPR, a highly effective approach. The search for optimal DNA sequence selection for manipulation in bioinformatics continues, with space for improvement. However, the promising field of machine learning (ML) aids this effort. Scientists can optimize gene editing studies and reliably predict their results using machine learning in bioinformatics. The team used machine learning methods to find the best amino acid residue combinations for Cas9 binding to target DNA. Due to the massive number of genetic differences, a large-scale experiment would have been impracticable. By using machine learning-driven engineering, screening was greatly simplified, reducing it by 95% (Krohannon et al. 2022).

16.10.2 Identifying Protein Structure

Proteomics, a bioinformatics area, studies proteins' complicated nature, interactions, composition, and vital role in the body's complex machinery. Bioinformatics analyzes and interprets large biological databases, which demand a lot of processing power. Bioinformatics jobs are computationally complex and require advanced algorithms and high-performance computing to handle and analyze data. Innovative technologies like machine learning are crucial in bioinformatics. A major bioinformatics success is the use of convolutional neural networks (CNNs) to classify protein amino acids into sheet, helix, and coil categories. Neural networks have achieved 84% accuracy, reaching the theoretical top bounds of 88–90%.

Machine learning (ML) has been used in proteomics, a topic that combines biology and computer science. Protein model score, essential for protein structure prediction, is one use. Researchers use ML algorithms to improve protein structure prediction, improving protein function and drug development. ML in proteomics has helped resolve the intricate link between protein structure and function, advancing bioinformatics. Fayetteville State University bioinformatics researchers used machine learning. ML was used to improve protein model scoring accuracy. The protein models were grouped and analyzed using a machine learning method. This approach determined the most important features for evaluating models in each group. The data feature vectors were used to improve machine learning algorithms during training, with each group trained separately.

16.10.3 Spotting Genes Associated with Diseases

Bioinformatics researchers increasingly use machine learning to uncover disease-related genes. The process uses RNA sequencing and gene expression microarray analysis. In cancer research, gene identification helps locate cancer-causing genes and classify tumors molecularly. Cancer prediction and classification were evaluated using decision tree, support vector machine, and neural network bioinformatics at the University of Washington. RNA sequencing data from The Cancer Genome Atlas project showed that linear support vector machine identified cancer best with 95.8% accuracy. Using gene expression data using ML, another study categorized breast cancer types. This team used Cancer Genome Atlas data. Researchers categorized breast cancer samples into triple negative and non-triple negative. Support vector machine classifiers excelled again (Athreya et al. 2018). Penn researchers employed machine learning to uncover CAD drug targets in non-cancerous illnesses. The researchers uncovered CAD-related SNPs using

ML-powered Tree-based Pipeline Optimization Tool. They detected 28 relevant SNPs in UK Biobank genomic data. This study confirmed that the top SNPs on this list were connected to CAD in the literature (Liu et al. 2022).

16.10.4 Traversing the Knowledge Base in Search of Meaningful Patterns

Researchers are trying to gain insights from genomic databases that double every 2.5 years thanks to advanced sequencing technologies. Biomedical articles and studies can be analyzed using machine learning to find genes and proteins and their functions. It can also annotate protein databases and provide literature information. A group of researchers used bioinformatics and machine learning in literature mining to score protein models. Multiple protein-protein docking models are usually produced and scored based on structural constraints. The team utilized ML techniques to search PubMed papers on protein-protein interactions for residues to establish model score constraints. To ensure the limitations are meaningful, scientists tested machine learning techniques to examine all residues for relevance.

This study found that computationally expensive neural networks and less resource-intensive support vector machines performed similarly (Zhou et al. 2022).

16.10.5 Repurposing Drugs

In the area of bioinformatics, researchers adeptly leverage the strategy of drug repurposing, also known as reprofiling, to explore novel applications for existing pharmaceutical agents. The utilization of artificial intelligence (AI) methodologies by bioinformatics researchers enables the comprehensive analysis of vast datasets from Binding DB and DrugBank. Drug repurposing, also known as drug repositioning, encompasses a multifaceted strategy that involves the exploration of existing drugs for novel therapeutic applications. This innovative field of research employs three primary approaches to identify potential drug candidates for repurposing (Pushpakom et al. 2019). These approaches include:

Target-based approach field of drug-target interaction encompasses the investigation of the direct binding between drugs and their target proteins.

Drug-drug interaction studies elucidate the intricate interplay between pharmaceutical agents, shedding light on the multifaceted mechanisms by which these compounds interact within biological systems.

The exploration of intracellular protein surfaces for hotspots and allosteric regions is a fundamental aspect of protein-protein interaction searches in the field of bioinformatics.

Researchers from China University of Petroleum and Shandong University employed a cutting-edge deep neural network methodology to analyze and extract valuable insights from the extensive DrugBank database. The primary focus of their research revolved around investigating the drug-target interactions involving mitochondrial fusion protein 2 (MFN2), a protein that has been implicated as a potential etiological factor in Alzheimer's disease. A recent investigation has successfully identified a collection of 15 distinct medicinal compounds exhibiting promising binding potential. Subsequent investigations have revealed that the protein 11 exhibits the capability to engage in docking interactions with the mitochondrial fusion protein MFN2. The quintet exhibits a range of medium-to-strong binding affinities (Wang et al. 2021).

16.11 Conclusion

The integration of Artificial Intelligence (AI) and Machine Learning (ML) methodologies has exhibited remarkable promise within the realm of bioinformatics. AI, an expansive domain encompassing machine learning (ML), empowers systems to acquire knowledge from data and subsequently generate predictions or make informed decisions. Bioinformatics, a burgeoning field at the intersection of biology and computer science, has witnessed the utilization of cutting-edge artificial intelligence (AI) algorithms to meticulously scrutinize vast and intricate datasets. These datasets encompass a wide array of genetic variations, harboring invaluable information that can be harnessed to unravel patterns and glean profound insights. By leveraging the power of AI, bioinformaticians strive to unlock novel avenues for drug discovery and treatment development, thus revolutionizing the landscape of modern medicine. In conclusion, the integration of artificial intelligence (AI) and machine learning (ML) methodologies has emerged as indispensable assets within the realm of bioinformatics. These cutting-edge technologies empower scientific investigators to scrutinize vast and intricate datasets, thereby facilitating the identification of intricate patterns and invaluable insights that would otherwise prove arduous or unattainable through conventional approaches. The burgeoning field of bioinformatics is witnessing a remarkable surge in the utilization of Artificial Intelligence (AI) and Machine Learning (ML) methodologies. This trend is anticipated to persist in the foreseeable future, driven by the scientific community's pursuit of novel therapeutic interventions and pharmaceutical advancements targeting diverse ailments and medical conditions.

References

Alpaydin E (2020) Introduction to machine learning, 4th ed. p 1–3, 13–18

Athreya AP, Gaglio AJ, Cairns J, Kalari KR, Weinshilboum RM, Wang L, Kalbarczyk ZT, Iyer RK (2018) Machine learning helps identify new drug mechanisms in triple-negative breast cancer. IEEE Trans Nanobioscience 17(3):251–259. https://doi.org/10.1109/TNB.2018.2851997

Berrar D (2018) Bayes' theorem and naive Bayes classifier. In: Encyclopedia of bioinformatics and computational biology: ABC of bioinformatics. Elsevier, pp 403–412

Bouckaert RR, Frank E, Hall MA, Holmes G, Pfahringer B, Reutemann P, Witten IH (2010) WEKA—experiences with a Java open-source project. J Mach Learn Res 11:2533–2541

Burbidge R, Trotter M, Buxton B, Holden S (2001) Drug design by machine learning: support vector machines for pharmaceutical data analysis. Comput Chem 26(1):5–14

Celebi ME (ed) (2014) Partitional clustering algorithms. Springer

Charbuty B, Abdulazeez A (2021) Classification based on decision tree algorithm for machine learning. J Appl Sci Technol Trends 2(01):20–28

Chen XW, Gao JX (2016) Big data bioinformatics. Methods 111:1–2. https://doi.org/10.1016/j.ymeth

Chetty M, Hallinan J, Ruz GA, Wipat A (2022) Computational intelligence and machine learning in bioinformatics and computational biology. Biosystems 222:104792. https://doi.org/10.1016/j.biosystems.2022.104792

Contreras P, Murtagh F (2015) Hierarchical clustering. In: Handbook of cluster analysis, pp 103–123

Demšar J, Curk T, Erjavec A, Gorup Č, Hočevar T, Milutinovič M et al (2013) Orange: data mining toolbox in python. J Mach Learn Res 14(1):2349–2353

Ekins R, Chu FW (1999) Microarrays: their origins and applications. Trends Biotechnol 17(6): 217–218

Erickson BJ (2021) Basic artificial intelligence techniques: machine learning and deep learning. Radiol Clin N Am 59(6):933–940. https://doi.org/10.1016/j.rcl.2021.06.004

Esposito S, Carputo D, Cardi T, Tripodi P (2019) Applications and trends of machine learning in genomics and phenomics for next-generation breeding. Plan Theory 9(1):34

Frank E, Hall M, Holmes G, Kirkby R, Pfahringer B, Witten IH, Trigg L (2010) Weka-A machine learning workbench for data mining. In: Data mining and knowledge discovery handbook, pp 1269–1277

Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S et al (2004) Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 5(10): 1–16

Goudbeek M, Swingley D, Smits R (2009) Supervised and unsupervised learning of multidimensional acoustic categories. J Exp Psychol Hum Percept Perform 35(6):1913–1933. https://doi.org/10.1037/a0015781

Gupta R, Srivastava D, Sahu M, Tiwari S, Ambasta RK, Kumar P (2021) Artificial intelligence to deep learning: machine intelligence approach for drug discovery. Mol Divers 25(3):1315–1360. https://doi.org/10.1007/s11030-021-10217-3

Hofmann M, Klinkenberg R (eds) (2016) RapidMiner: data mining use cases and business analytics applications. CRC Press

Karim MR, Beyan O, Zappa A, Costa IG, Rebholz-Schuhmann D, Cochez M, Decker S (2021) Deep learning-based clustering approaches for bioinformatics. Brief Bioinform 22(1):393–415. https://doi.org/10.1093/bib/bbz170

Kelchtermans P, Bittremieux W, De Grave K, Degroeve S, Ramon J, Laukens K et al (2014) Machine learning applications in proteomics research: how the past can boost the future. Proteomics 14(4–5):353–366

Kotu V, Deshpande B (2014) Predictive analytics and data mining: concepts and practice with rapidminer. Morgan Kaufmann

Krohannon A, Srivastava M, Rauch S, Srivastava R, Dickinson BC, Janga SC, Sowary CA (2022) CRISPR-Cas13 guide RNA predictor for transcript depletion. BMC Genomics 23(1):172. https://doi.org/10.1186/s12864-022-08366-2

Langley P (2011) The changing science of machine learning. Mach Learn 82(3):275–279. https://doi.org/10.1007/s10994-011-5242-y

Le NQK, Do DT, Hung TNK, Lam LHT, Huynh TT, Nguyen NTK (2020) A computational framework based on ensemble deep neural networks for essential genes identification. Int J Mol Sci 21:9070. https://doi.org/10.3390/ijms21239070

Li R, Li L, Xu Y, Yang J (2022) Machine learning meets omics: applications and perspectives. Brief Bioinform 23(1):bbab460

Libbrecht MW, Noble WS (2015) Machine learning applications in genetics and genomics. Nat Rev Genet 16(6):321–332

Liu C, Che D, Liu X, Song Y (2013) Applications of machine learning in genomics and systems biology. Comput Math Methods Med 2013:587492

Liu Y, Qiao N, Altinel Y (2021) Reinforcement learning in Neurocritical and neurosurgical care: principles and possible applications. Comput Math Methods Med 6657119:1. https://doi.org/10.1155/2021/6657119

Liu L, Zhai W, Wang F, Yu L, Zhou F, Xiang Y, Huang S, Zheng C, Yuan Z, He Y, Yu Z, Ji J (2022) Using machine learning to identify gene interaction networks associated with breast cancer. BMC Cancer 22(1):1070. https://doi.org/10.1186/s12885-022-10170-w

Medin DL, Schaffer MM (1978) Context theory of classification learning. Psychol Rev 85(3): 207–238. https://doi.org/10.1037/0033-295X.85.3.207

Meyer D, Wien FT (2001) Support vector machines. R News 1(3):23–26

Min S, Lee B, Yoon S (2017) Deep learning in bioinformatics. Brief Bioinform 18:851–869. https://doi.org/10.1093/bib/bbw068

Mohsen Y-N, Earl H, Dan T, John S, Milad E (2021) Application of machine learning algorithms in plant breeding: predicting yield from hyperspectral reflectance in soybean? Front Plant Sci 11: 624273. https://doi.org/10.3389/fpls.2020.624273

Mou M, Pan Z, Lu M, Sun H, Wang Y, Luo Y, Zhu F (2022) Application of machine learning in spatial proteomics. J Chem Inf Model 62(23):5875–5895

Muggleton SH (2005) Machine learning for systems biology. In: International conference on inductive logic programming. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 416–423

Munjal NK, Fleischman AD, Coller RJ (2023) Machine learning, predicting future hospitalizations, and the importance of perception. Hosp Pediatr 13(5):e114–e116. https://doi.org/10.1542/hpeds.2023-007224

Navada A, Ansari AN, Patil S, Sonkamble BA (2011) Overview of use of decision tree algorithms in machine learning. In IEEE control and system graduate research colloquium. IEEE. p 37–42

Ngiam KY, Khor IW (2019) Big data and machine learning algorithms for health-care delivery. Lancet Oncol 20(5):e262–e273. https://doi.org/10.1016/S1470-2045(19)30149-4. Erratum in: Lancet Oncol. 20(6):293

Nosi V, Luca A, Milan M, Arigoni M, Benvenuti S, Cacchiarelli D, Cesana M, Riccardo S, Di Filippo L, Cordero F et al (2021) MET exon 14 skipping: a case study for the detection of genetic variants in cancer driver genes by deep learning. Int J Mol Sci 22:4217. https://doi.org/10.3390/ijms22084217

Nunez-Iglesias J, Kennedy R, Parag T, Shi J, Chklovskii DB (2013) Machine learning of hierarchical clustering to segment 2D and 3D images. PLoS One 8(8):e71715

Perakakis N, Yazdani A, Karniadakis GE, Mantzoros C (2018) Omics, big data and machine learning as tools to propel understanding of biological mechanisms and to discover novel diagnostics and therapeutics. Metabolism 87:A1–A9

Persson Hoden K, Hu X, Martinez G, Dixelius C (2021) Smart PARE: an R package for efficient identification of true mRNA cleavage sites. Int J Mol Sci 22:4267. https://doi.org/10.3390/ijms22084267

Pirooznia M, Yang JY, Yang MQ, Deng Y (2008) A comparative study of different machine learning methods on microarray gene expression data. BMC Genomics 9:1–13

Pushpakom S, Iorio F, Eyers PA, Escott KJ, Hopper S, Wells A, Doig A, Guilliams T, Latimer J, McNamee C, Norris A, Sanseau P, Cavalla D, Pirmohamed M (2019) Drug repurposing: progress, challenges and recommendations. Nat Rev Drug Discov 18(1):41–58. https://doi.org/10.1038/nrd.2018.168

Reimers M, Carey VJ (2006) Bioconductor: an open source framework for bioinformatics and computational biology. Methods Enzymol 411:119–134

Ripley BD (2001) The R project in statistical computing. MSOR Connections. The Newsletter of the LTSN Maths, Stats & OR Network 1(1):23–25

Saritas MM, Yasar A (2019) Performance analysis of ANN and Naive Bayes classification algorithm for data classification. Int J Intell Syst Appl Eng 7(2):88–91

Sarle Warren S (1994) Neural networks and statistical models. In SUGI 19: proceedings of the nineteenth annual SAS users group international conference. SAS Institute, p 1538–1550. ISBN 9781555446116. OCLC 35546178

Sonagara D, Badheka S (2014) Comparison of basic clustering algorithms. Int J Comput Sci Mob Comput 3(10):58–61

Stuart R, Peter N (2003) Artificial intelligence: a modern approach, 2nd edn. Prentice Hall. ISBN 978-0137903955

Tierney L (2012) The R statistical computing environment. In: Statistical challenges in modern astronomy V. Springer New York, New York, NY, pp 435–447

Vamathevan J, Clark D, Czodrowski P, Dunham I, Ferran E, Lee G, Li B, Madabhushi A, Shah P, Spitzer M, Zhao S (2019) Applications of machine learning in drug discovery and development. Nat Rev Drug Discov 18(6):463–477. https://doi.org/10.1038/s41573-019-0024-5

Venkatesh KK, Strauss RA, Grotegut CA, Heine RP, Chescheir NC, Stringer JSA, Stamilio DM, Menard KM, Jelovsek JE (2020) Machine learning and statistical models to predict postpartum hemorrhage. Obstet Gynecol 135(4):935–944. https://doi.org/10.1097/AOG.0000000000003759

Wang S, Liu D, Ding M, Du Z, Zhong Y, Song T, Zhu J, Zhao R (2021) SE-onion net: a convolution neural network for protein-ligand binding affinity prediction. Front Genet 11: 607824. https://doi.org/10.3389/fgene.2020.607824

Weltz J, Volfovsky A, Laber EB (2022) Reinforcement learning methods in public health. Clin Ther 44(1):139–154. https://doi.org/10.1016/j.clinthera.2021.11.002

Yan J, Wang X (2022) Unsupervised and semi-supervised learning: the next frontier in machine learning for plant systems biology. Plant J 111(6):1527–1538. https://doi.org/10.1111/tpj.15905. Epub 2022 Jul 27

Zhou Y, Shi W, Zhao D, Xiao S, Wang K, Wang J (2022) Identification of immune-associated genes in diagnosing aortic valve calcification with metabolic syndrome by integrated bioinformatics analysis and machine learning. Front Immunol 13:937886. https://doi.org/10.3389/fimmu.2022.937886