Gastroenterology and Hepatology From Bed to Bench.
2024 **RIGLD**, Research Institute for Gastroenterology and Liver Diseases

REVIEW ARTICLE

p-ISSN: 2008-2258 e-ISSN: 2008-4234

Artificial intelligence and bioinformatics: a journey from traditional techniques to smart approaches

Hamid Jamialahmadi[1,*], Ghazaleh Khalili-Tanha[1,*], Elham Nazari[2,#], Mostafa Rezaei-Tavirani[2,#]

[1]Department of Medical Genetics and Molecular Medicine, School of Medicine, Mashhad University of Medical Sciences, Mashhad, Iran
[2]Proteomics Research Center, Faculty of Paramedical Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran
[#] Corresponding authors
** These authors equally contributed to this study as the first authors.*

ABSTRACT

The incorporation of AI models into bioinformatics has brought about a revolutionary era in the analysis and interpretation of biological data. This mini-review offers a succinct overview of the indispensable role AI plays in the convergence of computational techniques and biological research. The search strategy followed PRISMA guidelines, encompassing databases such as PubMed, Embase, and Google Scholar to include studies published between 2018 and 2024, utilizing specific keywords. We explored the diverse applications of AI methodologies, including machine learning (ML), deep learning (DL), and natural language processing (NLP), across various domains of bioinformatics. These domains encompass genome sequencing, protein structure prediction, drug discovery, systems biology, personalized medicine, imaging, signal processing, and text mining. AI algorithms have exhibited remarkable efficacy in tackling intricate biological challenges, spanning from genome sequencing to protein structure prediction, and from drug discovery to personalized medicine. In conclusion, this study scrutinizes the evolving landscape of AI-driven tools and algorithms, emphasizing their pivotal role in expediting research, facilitating data interpretation, and catalyzing innovations in biomedical sciences.

Keywords: Artificial intelligence, Bioinformatics, Drug discovery, Personalize medicine, Visual data processing.

(Please cite as: **Jamialahmadi H, Khalili-Tanha G, Nazari E, Rezaei-Tavirani M. Artificial intelligence and bioinformatics: a journey from traditional techniques to smart approaches. Gastroenterol Hepatol Bed Bench 2024;17(3):241-252. https://doi.org/10.22037/ghfbb.v17i3.2977).**

Introduction

Bioinformatics serves as a crucial bridge between different aspects of biology and computational analysis, unlocking hidden patterns and insights from complex datasets, and providing the essential tools and methods to analyze, interpret, and extract meaningful knowledge from these massive data (1). The past decade has witnessed a transformative shift in the field of bioinformatics, traditionally characterized by the application of computational tools to analyze biological data. The emergence of artificial intelligence (AI) has created a powerful engine for revolutionizing biological research approaches and the onset of a new era of innovation (2, 3). The exponential growth of biological data (Big Data), powered by high-throughput sequencing and other cutting-edge technologies, has faced us with a significant challenge due to its time-consuming and complex nature (3). Machine algorithms, such as support vector machines (SVM), random forests, and neural networks, are becoming widely used indispensable tools in bioinformatics. These powerful techniques are empowering researchers

Received: 07 April 2024 *Accepted*: 11 May 2024
Reprint or Correspondence: **Elham Nazari, Mostafa Rezaei-Tavirani,** *Department of Health Information, Technology and Management, School of Allied Medical Sciences, Shahid Beheshti University of Medical Science, Tehran, Iran, Proteomics Research Center, Faculty of Paramedical Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran.*
E-mail: Elham.Nazari@sbmu.ac.ir, Tavirani@sbmu.ac.ir
ORCID ID: 0009-0000-8452-2946, 0000-0003-1767-7475

to tackle complex challenges across various fields, including deciphering the genetic codes (4), predicting protein structures and drug development with remarkable accuracy, together with disease diagnosis, and identifying novel biomarkers (5, 6).

These are just a few examples of the transformative power of AI in bioinformatics. In essence, AI is rapidly transforming bioinformatics, paving the way for a deeper understanding of life processes, revolutionizing disease diagnosis, and ultimately leading to the development of personalized medical strategies (7-10). However, it is crucial to acknowledge that AI is not a panacea. Ethical considerations surrounding data privacy, algorithmic bias, and the potential for misuse necessitate careful attention. By embracing this powerful technology responsibly, ethically, and collaboratively, we can unlock a new era of personalized medicine, improved healthcare, and a deeper appreciation for the intricate complexities of life. In the current review, we have summarized key categories of emerging and established bioinformatics applications that have been significantly impacted by AI. Of note is also the importance of AI-driven tools and algorithms in speeding up research, improving data interpretation, and driving innovation in biomedical sciences (Table 1).

Table 1. Summary of examples illustrating the usage of AI in bioinformatics

Field of bioinformatics	Input data	AI algorithms examples	References
1. Prediction of molecular interactions, modeling, and drug discovery	Protein sequences, protein structures	1- Support Vector Machines (SVMs): Classify protein-protein interactions and predict drug-binding sites.	Cai et al. (65)
		2- Deep Learning (Convolutional Neural Networks - CNNs): Predict protein structures and model protein-protein interactions.	Strokach et al. (66)
2. Omics: Genomics, Transcriptomics, Epigenomics, Proteomics, Metagenomics	DNA sequences, RNA sequences, epigenetic data, protein sequences, metagenomics data	1- Clustering algorithms (K-means, Hierarchical): Identify co-expressed genes/proteins with similar expression patterns.	Wu (67), Wei et al. (68)
		2- Random Forest (RF,) Support Vector Machine (SVM), and Extreme Gradient Boosting (XGBoost): Identify differentially expressed genes/proteins	Pragya et al. (69), Hoque et al. (70), Abbas, et al. (71)
3. Phylogenetic assessments	DNA sequences, protein sequences	1-Nearest Neighbors: Identify the closest evolutionary neighbors of a sequence.	Collienne (72)
		2-Maximum Likelihood Estimation: Reconstruct phylogenetic trees based on sequence data.	Lin et al. (73)
4. System Biology	Omics data, protein-protein interaction data, metabolic pathway data	1-- Bayesian Networks: Model relationships between genes/proteins in a biological system.	Largo et al. (74)
		2- Ordinary Differential Equations (ODEs): Model the dynamics of biological processes.	Liu et al. (75)
5. Personalize Medicine: Diagnostic, Prognostic, and Predictive biomarkers	Genomic data, clinical data, biomarker levels	1- Machine Learning models (Logistic Regression, Random Forests): Predict disease risk based on an individual's genetic data.	Peng et al. (76)
		2- Deep Learning (ANN, CNN): Analyze biomarkers and data from medical images for disease diagnosis and prognosis.	Guadiana-Alvarez et al. (77)
6. Medical Visual Data: Biomedical imaging, Signal processing	Medical images (X-rays, CT scans, MRIs), biomedical signals (ECG, EEG, etc.)	1- Deep Learning (CNNs): Segment medical images, identify abnormalities, and classify diseases.	Sarvamangala (78)
		2- Computer vision techniques: Extract features from medical images for analysis.	Litjens (79)
7. Biomedical text mining	Biomedical texts (scientific articles, clinical records)	1-- Natural Language Processing (NLP) techniques: Extract information from text, and identify entities and relationships.	Zeng (80)
		2- Machine Learning models (topic modeling, sentiment analysis): Identify relevant information and classify documents.	Liu (81)

Methods

Relevant articles were identified through a comprehensive search strategy conducted in PubMed, Embase, and Google Scholar. The search included studies published between 2018 and 2024. The process of study screening is delineated in the PRISMA flowchart displayed in Figure 1. For this review, we initially retrieved 524 pieces of literature, subsequently excluding irrelevant studies by assessing their titles and abstracts. Furthermore, after conducting a full-text review, certain studies were excluded due to inconsistencies in their findings, being conference or unpublished papers, focusing solely on AI or bioinformatics, or falling outside the timeframe of our study. Finally, 52 original research, plus some relevant review and survey articles were included for further assessments. The literature search was independently conducted by two individuals, and any uncertainties were resolved through consultation with another author. The keywords used were: ("Artificial Intelligence" OR "AI" OR "Machine Intelligence") AND ("Bioinformatics" OR "Biological Data Analysis") AND ("Machine Learning" OR "ML") AND ("Deep Learning" OR "DL") AND ("Natural Language Processing" OR "NLP") AND ("Genome Sequencing" OR "Genomic Analysis") AND ("Protein Structure Prediction") AND ("Drug Discovery") AND ("System Biology") AND ("Personalized Medicine") AND ("Imaging") AND ("Signal Processing") AND ("Text Mining"). Finally, the reference lists of the retrieved publications were also manually searched to find any possibly related research.

Exploring the Transition from Classical to Smart Bioinformatics

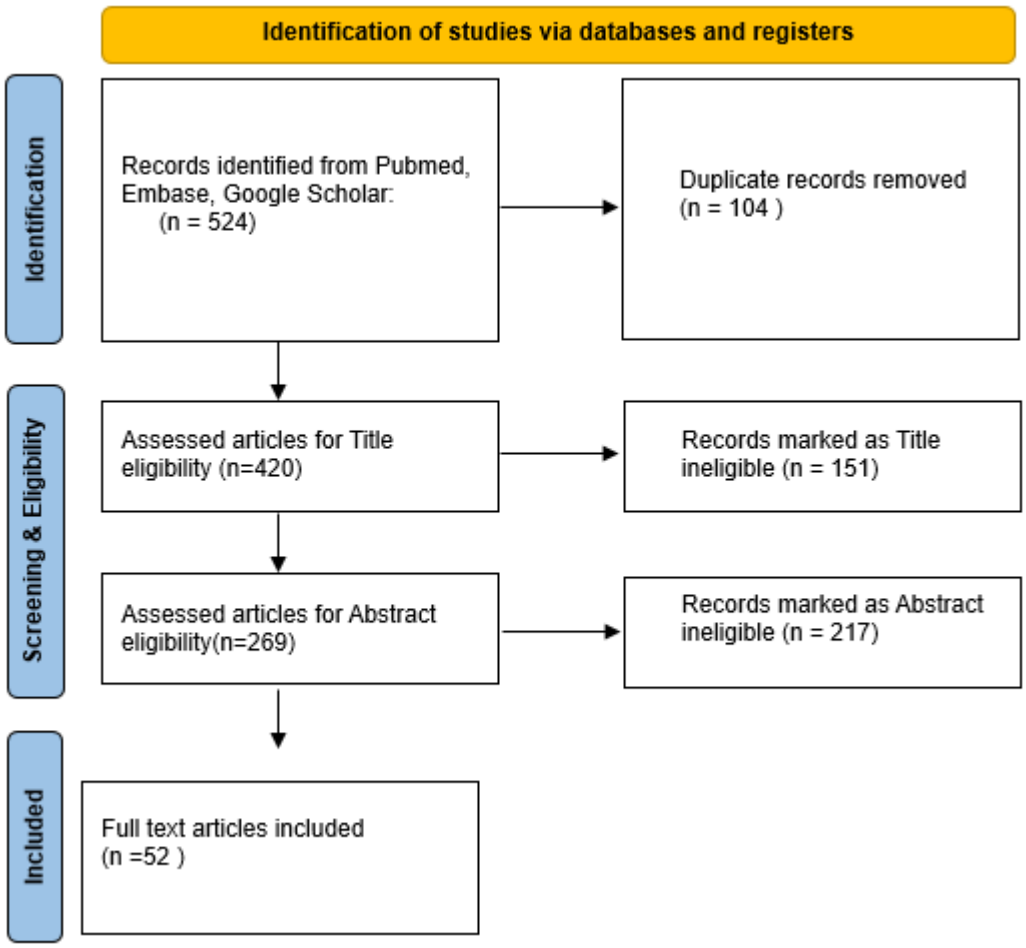

Figure 1. PRISMA flowchart of study design

Classical bioinformatics traditionally relied on rule-based algorithms, statistical methods, and manual interpretation of biological data. However, the exponential growth of biological data generated by high-throughput technologies such as next-generation sequencing has rendered classical methods inadequate for managing the complexity and scale of data (11). The shift from classical to smart bioinformatics marks a significant advancement in the field, primarily driven by breakthroughs in AI, ML, and big data analytics. The integration of AI and ML into bioinformatics not only enhances the efficiency and accuracy of analyses but also opens new avenues for understanding complex biological systems (12, 13).

Classical bioinformatics methods have notable limitations, including their struggles with handling intricate data, reliance on manual feature engineering, and lack of adaptability. Also, classical bioinformatics methods often encounter challenges when dealing with noisy data. Noisy data refers to data that contain random fluctuations or errors, which can arise from various sources such as experimental limitations, measurement inaccuracies, or biological variability. These noisy elements can distort the accuracy and reliability of analyses conducted using classical bioinformatics techniques (14). Thus, AI-based bioinformatics approaches offer several advantages, addressing the shortcomings of classical methods: (1) Automated feature learning: AI-based techniques can autonomously identify relevant features from data, eliminating the need for manual feature engineering. This capability has the potential to unveil hidden patterns and relationships within the data. (2) Adaptability: AI-based methods exhibit high adaptability, capable of being deployed across a diverse array of biological tasks such as classification, regression, clustering, and sequence analysis. (3) Cutting-edge performance: AI-based approaches have often demonstrated state-of-the-art performance across various bioinformatics tasks, including but not limited to protein structure prediction, gene expression analysis, and drug discovery (15-17).

Combining AI methods with bioinformatics effectively addresses various limitations and issues encountered in both fields. This integration allows for a more comprehensive analysis that leverages the strengths of both disciplines. Firstly, by integrating AI techniques such as machine learning and deep learning with bioinformatics, researchers can overcome limitations in traditional computational methods. AI algorithms can efficiently handle large-scale biological data, enabling more accurate and detailed modeling of biological processes (18). Secondly, this combination enhances the accuracy, specificity, and sensitivity of predictive models. AI algorithms can identify complex patterns and relationships within biological data that may not be apparent through conventional methods alone. By integrating AI with bioinformatics, researchers can develop predictive models that are not only more accurate but also more robust and reliable (19). Furthermore, the synergy between AI and bioinformatics enables the discovery of novel insights and associations in biological systems. This integrated approach facilitates the identification of biomarkers, drug targets, and disease mechanisms, leading to advancements in personalized medicine and drug development (20, 21).

The role of AI in bioinformatics

Prediction of molecular interactions, drug discovery, and modeling

Despite advancements in high-throughput screening of molecular interactions, the vast number of potential drug-target combinations and the complexity of model prediction pose a significant challenge in drug discovery, making it time-consuming and costly (22). AI is revolutionizing drug discovery through various applications. Graph neural networks (GNNs) predict how potential drug molecules interact with targets and guide drug selection. Generative adversarial networks (GANs) as another ML (ML) method, design novel drug molecules with desired properties, accelerating the discovery process. Additionally, AI models simulate and predict the outcome of drug treatment, enabling personalized medicine and optimized targeted therapy strategies (23, 24).

Omics

"Omics," as a wide concept, includes fields such as genomics, proteomics, and metabolomics, studying the vast molecular landscape within living organisms. AI by recruiting ML and DL (DL) algorithms, is revolutionizing Omics by analyzing these complex datasets, helping researchers understand biological

processes, diagnose diseases, and even design personalized therapies (25).

Genomics

AI methods, particularly ML algorithms, have shown promise in tasks such as variant calling, comparative genomics, and gene prediction. Convolutional neural networks (CNNs) and recurrent neural networks (RNNs), two types of DL models, have shown impressive results in sequence analysis tasks, such as DNA motif, transcription factor binding site, and non-coding RNA element prediction. For example, CNNs break the sequence into pieces, analyze them with filters, and identify important features. These features are then used to predict labels (e.g., binding targets in DNA). In comparison, the RNNs process the sequence one element at a time, considering the context of each element based on its neighbors and their predicted labels (e.g., differentiating between exons and introns in DNA) (26). DeepVariant is an AI-powered tool for identifying genetic variations in DNA. Traditional methods struggle with various factors including sample preparation, sequencing technology, and biological variations. DeepVariant utilizes CNNs to analyze raw sequencing data directly, avoiding these biases and achieving higher accuracy in variant calling (26). In a recent study, various genomic datasets including RNASeq, SNP, and CNV data were combined with clinical follow-up information sourced from TCGA. The dataset was then randomly divided into training and validation sets. Utilizing Random Forest (RF) analysis, six key genes (CD24, RCC2, MRGPRX, CASP8, PRRG1, and IQSEC3) were identified as top-ranking features. This unique 6-gene signature, developed within the study, emerged as a promising prognostic biomarker for breast cancer (BC) patients. It represents a novel prognostic marker with the potential to offer both diagnostic and prognostic insights, as well as serve as a foundation for identifying therapeutic targets across this patient population (27).

Transcriptomics

Transcriptomics, the comprehensive assessment of RNA transcripts, including their structure, interactions, and functions is crucial in understanding gene expression and regulation. The advent of AI has revolutionized this field by enabling researchers to analyze the vast and complex data generated from RNA-sequencing technologies (28). AI algorithms, particularly CNNs and RNNs, are tackling diverse challenges in transcriptomics including RNA structure prediction, RNA-protein interaction prediction, and Non-coding RNA (ncRNA) analysis (29, 30). Takeshita et al. pioneered an ML-based prognostic framework tailored to hormone receptor-positive breast cancer (BC). Employing logistic regression (LR) and hierarchical clustering algorithms, they pinpointed a distinctive nine-gene expression profile. This signature, comprising genes such as C1orf64, AGL, CYP4F22, KIF20A, TUBA3D, S100P, PRC1, LAD1, and HNMT was found to be closely associated with the prognosis of triple-negative breast cancer (TNBC) patients. Notably, this model demonstrated its efficacy in stratifying prognosis across diverse patient cohorts, effectively reflecting critical BC therapeutic pathways and the tumor immune microenvironment. Further, the model's effectiveness was substantiated by the observed therapeutic responses to chemotherapy and endocrine therapy (31).

Epigenomics

AI empowers researchers in epigenomics, the study of modifications that affect gene expression without altering the DNA sequence, to analyze complex data from chromatin immunoprecipitation sequencing (ChIP-seq) and DNA methylation profiling. Algorithms such as CNNs and SVM handle diverse tasks such as identifying regulatory elements. For example, CNNs determine DNA regions influencing gene expression, including enhancers, crucial for understanding gene regulation and developing therapies. In addition, SVM models are used for predicting disease risk to identify individuals at higher risk for specific diseases, paving the way for early intervention strategies (32, 33). A recent study revealed that LIHC data encompass four distinct categories: methylation, histone modifications, human genome information, and RNA sequences. These data were accessed using open-source technologies within the R programming language, leveraging The Cancer Genome Atlas (TCGA). The study employed a methodology that assessed 1,000 features spanning these four data types. Nine distinct feature selection techniques were employed, alongside the comparison of eight classification methods, in order to identify the optimal model using 5-fold cross-validation and various training-to-test ratios. The most

effective model emerged when utilizing 140 features selected through ReliefF feature selection in conjunction with the XGBoost classification method, yielding an AUC of 1.0 and an accuracy of 99.67% in predicting liver cancer (34).

Proteomics

Within the field of proteomics, which investigates the structure, function, and interactions of proteins, AI is rapidly emerging as a powerful tool for deciphering the intricate complexities of these multifaceted macromolecules. Algorithms such as CNNs and GNNs tackle various challenges, including predicting structures of proteins, crucial for drug development (35), understanding protein-protein interactions to unravel cellular processes (36), analyzing post-translational modifications (PTMs) for targeted therapies (37), as well as identifying and classifying enzymes for drug discovery (38). In a recent investigation, researchers fused machine learning with microfluidic technologies to scrutinize extracellular vesicles (EVs) in TNBC. They inspected the EV proteomes of 100 individuals with breast cancer and 30 individuals without any illness. By employing a microfluidic chip-based technique to extract tumor-derived EVs from minimal plasma samples, they successfully eliminated impurities such as albumin and immunoglobulins. By leveraging a machine learning algorithm, they pinpointed a distinctive pattern composed of three EV proteins—extracellular matrix protein 1 (ECM1), biotinidase (BTD), and mannose-binding lectin 2 (MBL2), which accurately distinguished TNBC patients from healthy individuals. This pattern not only served as a diagnostic instrument but also exhibited associations with unfavorable prognosis and heightened recurrence rates in TNBC (39).

Metagenomics

The growing area of metagenomics, analyzing the collective genomes of microbial communities, is being revolutionized by AI. ML algorithms, such as random forests and DL models, are tackling complex tasks such as taxonomic classification (identifying microbial species) and functional prediction (understanding the roles of microbes)(40). In a study, Harris et al. explored the application of random forest models to analyze massive datasets of microbial communities (metagenomes). They achieved high accuracy (91%) in both identifying the origin of a metagenome sample and predicting unknown samples based on their taxonomic profile. This exemplifies the immense potential of AI in unlocking the secrets of microbiomes as well as their impact on human health (41). Another research demonstrated enhancing the accuracy of colorectal cancer disease status prediction through random forest classification utilizing metagenomic shotgun sequencing data. The study indicated that utilizing microbial relative abundance profiles estimated by Centrifuge generally yields superior prediction performance compared to those estimated by MetaPhlAn2 and Bracken. Furthermore, a pioneering approach has been devised to amalgamate relative abundance profiles of both established and newly discovered microbial organisms, thereby amplifying the predictive capability for colorectal cancer detection from metagenomic datasets (42).

Phylogenetic assessments

Evolutionary biology also benefits from the transformative power of AI, particularly in dealing with large genomic data and inaccuracies in interpretation resulting from missing data. Researchers are increasingly utilizing AI methods such as phylogenetic network inference and ancestral state reconstruction to analyze complex evolutionary relationships and understand the intricate web of life. For instance, Bhattacharjee et al. successfully applied two ML methods i.e. matrix factorization (MF) and autoencoder (AE) to impute missing entries for estimating phylogenetic trees, especially in the case of missing data in distance matrices. The authors suggest that machine-learning techniques can improve the accuracy of phylogenetic trees (43). An investigation explored tumor classification through phylogenetic methods applied to expression data. Their approach was tested on two distinct datasets: one comprising 87 tissues, predominantly small, round, blue-cell tumors (SRBCTs), and another consisting of 22 breast tumors. In the first dataset, the method effectively categorized samples into four major clusters, precisely aligning with neuroblastomas, rhabdomyosarcomas, Burkitt's lymphomas, and Ewing's family of tumors. Utilizing the breast cancer data, the classification tree differentiated tumors with BRCA1 mutations from those with BRCA2 mutations, along with sporadic tumors, which were also distinguished from each other.

They also observed the adaptability of our class discovery method through standard resampling techniques such as jackknifing and noise perturbation. To address the class prediction challenge, we constructed a classification tree using the learning set and then optimized the placement of each test sample within this tree. This method was successfully validated on the SRBCT dataset, accurately classifying each tumor (44). Azer et al. demonstrated that in tumor phylogeny reconstruction using single-cell sequencing, the primary objective is to construct the most accurate phylogenetic tree from the genotype matrix, which captures genotype information from individual cells but is inherently noisy. They propose rapid deep-learning methodologies to address the challenges of determining whether the most probable tree exhibits a linear (chain-like) or branching structure, as well as assessing the feasibility of constructing a perfect phylogeny from a given genotype matrix. Additionally, they introduced a reinforcement learning technique for reconstructing the most probable tumor phylogeny. This preliminary research underscores the potential of data-driven approaches in capturing crucial aspects of tumor evolution (45).

System Biology

Systems and network biology is a branch of biology focused on comprehensively understanding biological systems, encompassing their structures, functions, and interactions. Multi-omics data empower researchers to uncover fresh perspectives on the intricate interplays within biological systems, such as the relationships among genes, proteins, and cells. AI techniques such as ML and DL are used to model complex biological systems. These models can integrate various types of omics data (genomics, transcriptomics, proteomics, etc.) to capture the interactions and dynamics within biological systems (46, 47). A key objective in bioinformatics is to integrate various types of biological data, known as omics data. However, AI currently faces limitations in effectively handling this integration, which is expected to be addressed in the future. Thus, there is a growing recognition of the necessity for enhanced interpretability in the integration methods employed. To achieve this, ongoing efforts involve combining machine learning algorithms with methods that incorporate biological knowledge, such as graph representations or imposing strict constraints on parameter representations (20, 48). A study unveiled the potential of machine learning algorithms and systems biology analysis in predicting the chemoresistance trait of cancer cell lines. Among the six classifiers trained to differentiate between cisplatin-resistant and sensitive samples, Naïve Bayes and KNN algorithms emerged as the most promising tools based on various evaluation metrics. Additionally, systems biology analysis identified several genes associated with chemoresistance, with CTNNB1, IFNG, YWHAH, CTNNB1, EDNRB, ANKRD50, ACSL6, and PTGER3 being highlighted as particularly significant in terms of network topology. These findings lay the groundwork for further experimental investigations in this field (49). AI offers significant potential to revolutionize systems biology by expediting data analysis, predictive modeling, and personalized medicine. Nevertheless, several challenges must be tackled to fully exploit its capabilities. These include ensuring data quality, managing the complexity of data interpretation, addressing ethical and privacy concerns related to data usage, and mitigating automation bias (50).

AI and personalized medicine

A personalized medicine, or precision medicine, approach tailors medical treatment and prevention strategies by considering the unique genetic makeup, environmental factors, and lifestyle of each individual. AI is revolutionizing the field of personalized medicine, offering unique opportunities for medical treatment and care to individual patients (51, 52). AI techniques, including ML and DL algorithms, facilitate the discovery and validation of biomarkers by efficiently analyzing large-scale biological and clinical datasets, identifying complex patterns, and generating actionable insights that can inform personalized treatment decisions (53, 54). The merging of AI technologies with advancements in molecular biology, genetics, and clinical research is enhancing personalized medicine, making it more accurate and efficient. This approach provides targeted therapies for individual patients, especially in detecting and using diagnostic, prognostic, and predictive biomarkers (7, 55). Khalili-Tanha et al. utilized bioinformatic analysis to anticipate significant gene markers in colorectal

cancer (CRC) by employing various ML algorithms including SVM, k-nearest neighbors (KNN), logistic regression, decision trees (DTs), and random forest (RF). They reported that RF algorithm demonstrated the highest accuracy and Area Under the Curve (AUC), making it the most suitable algorithm for this prediction task. Different studies utilized several ML techniques such as RF, max voting, adaboost, gradient boosting machines (GBM), and extreme gradient boosting (XGB) to analyze differentially expressed genes (DEGs) and discover new prognostic biomarkers in pancreatic cancer. Among these methods, XGBoost was found to be particularly effective, offering improved performance and faster processing compared to other ML algorithms and deep learning models (56). Chang et al. utilized ML techniques including support vector machine, logistic regression, random forest, and naïve Bayes to develop a reliable predictive model for detecting biomarkers in Alzheimer's Disease patients (57). Sayed et al. employed various machine learning algorithms, such as XGboost, bagging, AdaBoost, SVM, and lightGBM, to detect vocal biomarkers for Parkinson's disease. Among these algorithms, lightGBM demonstrated the highest sensitivity and specificity in identification (58).

Diagnostic Biomarkers: These are biological markers that serve to either detect or validate the existence of a particular disease or condition or to pinpoint an individual belonging to a specific subtype of the ailment. AI algorithms can analyze vast amounts of patient data including genetic information, medical imaging results, and clinical data to identify patterns and correlations that may serve as diagnostic biomarkers. For instance, AI algorithms can analyze imaging scans to detect early signs of diseases such as cancer or analyze genetic data to identify mutations associated with specific disorders (59).

Prognostic Biomarkers: Prognostic biomarkers provide information about the likely course or outcome of a disease. AI can help in identifying patterns and relationships within patient data that can predict disease progression or treatment response. By analyzing various factors such as genetic profiles, medical history, lifestyle factors, and environmental influences, AI algorithms can generate personalized prognostic assessments for individual patients, allowing healthcare providers to tailor treatment plans accordingly (56, 60).

Predictive Biomarkers: Predictive biomarkers indicate the likelihood of a patient's response to a particular treatment or intervention. AI-driven predictive modeling can analyze diverse datasets including genomic data, clinical records, and treatment outcomes to identify biomarkers associated with response or resistance to specific therapies. This information enables clinicians to select the most effective treatment strategies for individual patients, optimizing therapeutic outcomes while minimizing potential adverse effects (60, 61).

Medical Visual Data

Biomedical imaging

Biomedical imaging involves visualizing the body's interior, enabling healthcare professionals to see various structures, functions, and processes within the body, aiding in disease detection, diagnosis, and monitoring. Common modalities of biomedical imaging include X-ray Imaging, computed tomography (CT), magnetic resonance imaging (MRI), ultrasound imaging, single photon emission computed tomography (SPECT), positron emission tomography (PET), fluorescence Imaging, and optical coherence tomography (OCT) (62). AI techniques, in particular DL, have demonstrated significant success in analyzing medical images. They can aid in tasks such as segmentation (identifying and delineating structures or regions of interest), classification (identifying diseases or abnormalities), registration (aligning images from various modalities or time points), image reconstruction (enhancing image quality, removing noise and artifacts, and improving resolution), and real-time image analysis (offering immediate feedback to clinicians during procedures or examinations) (63, 64). De Haan et al. investigated the progress of computational microscopy and optical sensing systems using deep neural networks. They explained the basics of solving inverse problems in optical microscopy and suggested DL as a viable solution, particularly through supervised methods. The authors demonstrated how DL can be applied to achieve single-image super-resolution and improve images in these datasets, showcasing the advancements made in the field (65).

Elsewhere, Gong et al. discussed the increasing use of machine learning alongside rapid waveform digitizers to forecast the location and arrival time of high-energy

photons. They highlighted its role in quantitative image reconstruction, where it aids in estimating correction factors such as scattered events and attenuation images, as well as reducing statistical noise in reconstructed images. This involves replacing time-intensive computations with faster machine learning alternatives, such as in scatter estimation, or employing data-driven approaches to define functions, including estimating attenuation maps for PET/MR scans (66).

Medical video analysis is a burgeoning field where computer vision and machine learning techniques are employed to scrutinize medical videos for diverse purposes. Among these, video action recognition garners increasing interest from computer vision researchers. A recent study examined the efficacy, reproducibility across tests, inter-modality consistency, as well as the correlation between deep features and clinical indicators such as tumor volume and TNM staging. Radiomics, serving as the benchmark image biomarker, was integrated into the analysis. CT scans were transformed into videos for deep feature extraction, utilizing the pre-trained Inflated 3D ConvNet (I3D) video classification network architecture. The findings revealed that deep features can surpass radiomics in predicting tumor prognosis from alternative perspectives compared to traditional measures such as tumor volume and TNM staging. However, deep features exhibit lower reproducibility and interpretability compared to radiomic features (67).

Signal processing

Signal processing refers to the manipulation, analysis, and interpretation of signals. In the context of biomedical applications, signals can include physiological data including ECG (Electrocardiogram), EEG (Electroencephalogram), and EMG (Electromyogram), as well as data from biomedical imaging modalities. AI methods can complement signal processing by learning complex patterns from the data to enhance the quality of signals, extract relevant information, remove noise, detect patterns or abnormalities, and monitor health conditions (68-70). An et al. utilized a deep belief network (DBN) to analyze the frequency components of EEG signals to distinguish between left and right-hand motor imagery. Their study compared the recognition accuracy of the DBN classifier with that of the Support Vector Machine (SVM), indicating that the DBN classifier consistently outperformed the SVM in all tested scenarios. This research introduces a novel deep-learning approach for accurately classifying EEG data based on motor imagery (71). Jia et al. introduced an innovative semi-supervised deep learning approach for recognizing affective states from EEG signals. Their framework, distinct from prior models, is specifically tailored to address EEG classification challenges, offering enhanced adaptability and performance (72).

Biomedical Text Mining

The extensive biomedical literature is a valuable knowledge source for researchers. Text mining techniques are increasingly used to extract and analyze information from various biomedical texts, including research articles, clinical notes, and electronic health records. This process involves natural language processing (NLP) algorithms for text understanding, ML models for information classification and extraction, and data mining techniques for identifying patterns and relationships in the data. In addition to textual content, figures such as biological pathways in publications convey valuable knowledge, offering visual representations of molecular events in biological processes or diseases. The objective is to expedite biomedical research, support clinical decision-making, and enhance healthcare outcomes by efficiently utilizing the wealth of information present in textual form within the biomedical field (73, 74). Named Entity Recognition (NER) indeed plays a vital role in extracting knowledge, especially in domains such as biomedicine where precise identification of terms such as genes, proteins, diseases, and drugs is crucial for further analysis and understanding. NER involves identifying and classifying named entities within a body of text, which enables automated systems to understand and extract relevant information accurately (75). ML methods, such as SVM, hidden Markov models (HMM), conditional random fields (CRFs), and maximum entropy (ME), are currently extensively employed for named entity recognition (76-81).

Discussion

Classic bioinformatics methods, while foundational, have limitations in handling the ever-growing biological data. Traditional approaches often rely on pre-defined

rules and require significant human expertise, leading to slow and potentially overlooking complex associations. In contrast, AI-powered bioinformatics offers exciting possibilities. Machine learning algorithms can handle massive datasets, uncovering hidden patterns and generating more accurate predictions. There are still challenges such as model interpretability and potential biases; nevertheless, AI is opening a new era in bioinformatics with the potential for better healthcare. These advancements hold immense promise for accelerating drug discovery, personalizing medicine, and unlocking a deeper understanding of biological systems (82, 83) (101-103).

Conclusion

In conclusion, this mini-review briefly outlined the essential role of AI within the realm of bioinformatics, where computational techniques meet the analysis of biological data. Our examination encompassed the diverse applications of AI techniques such as ML, DL, and NLP, elucidating their contributions to tasks such as genome sequencing, protein structure prediction, drug discovery, system biology, personalized medicine, imaging, signal processing, and text mining. Additionally, we delved into the evolving landscape of AI-driven tools and algorithms, highlighting their significance in accelerating research, enhancing data interpretation, and fostering innovations in biomedical sciences. In the future, the application of artificial intelligence in medicine will involve integrating various types of multi-omics data. These encompass genomics, epigenomics, transcriptomics, proteomics, metabolomics, single-cell multi-omics, microbiomics, radiomics, and spatial transcriptomics. This integration will be accomplished by combining various types of machine learning algorithms aimed at enhancing modeling and predictive capabilities.

Conflict of interests

There is no conflict of interest for authors of this article.

References

1. Weng JT, Wu LC, Chang WC, Chang TH, Akutsu T, Lee TY. Novel bioinformatics approaches for analysis of high-throughput biological data. Biomed Res Int 2014;2014:814092.

2. Zhang Z, Li J, He T, Ding J. Bioinformatics identified 17 immune genes as prognostic biomarkers for breast cancer: application study based on artificial intelligence algorithms. Front Oncol 2020;10:330.

3. Patel P, Pillai N, Toby I. No-boundary thinking for artificial intelligence in bioinformatics and education. Front Bioinform 2023;3:1332902.

4. Vadapalli S, Abdelhalim H, Zeeshan S, Ahmed Z. Artificial intelligence and machine learning approaches using gene expression and variant data for personalized medicine. Brief Bioinform 2022;23:191.

5. Lancellotti C, Cancian P, Savevski V, Kotha SRR, Fraggetta F, Graziano P, et al. Artificial intelligence & tissue biomarkers: advantages, risks and perspectives for pathology. Cells 2021;10:787.

6. Secinaro S, Calandra D, Secinaro A, Muthurangu V, Biancone P. The role of artificial intelligence in healthcare: a structured literature review. BMC Med Inform Decis Mak 2021;21:125.

7. Issa NT, Byers SW, Dakshanamurthy S. Big data: the next frontier for innovation in therapeutics and healthcare. Expert Rev Clin Pharmacol 2014;7:293-8.

8. Ching T, Himmelstein DS, Beaulieu-Jones BK, Kalinin AA, Do BT, Way GP, et al. Opportunities and obstacles for deep learning in biology and medicine. J R Soc Interface 2018;15:20170387.

9. Khan M. Artificial intelligence in bioinformatics: advancements and applications. 2023.

10. Ezziane Z. Applications of artificial intelligence in bioinformatics: a review. Expert Syst Appl 2006;30:2-10.

11. Paliwal S, Sharma A, Jain S, Sharma S. Machine learning and deep learning in bioinformatics. Bioinformatics and Computational Biology: Chapman and Hall/CRC; 2024. p. 63-74.

12. Gupta A, Kumar S, Kumar A. Big data in bioinformatics and computational biology: basic insights. Reverse Engineering of Regulatory Networks: Springer; 2023. p. 153-66.

13. Atas Guvenilir H, Dogan T. How to approach machine learning-based prediction of drug/compound-target interactions. J Cheminform 2023;15:16.

14. Askr H, Elgeldawi E, Aboul Ella H, Elshaier Y, Gomaa MM, Hassanien AE. Deep learning in drug discovery: an integrative review and future challenges. Artif Intell Rev 2023;56:5975-6037.

15. Blanchard AE, Stanley C, Bhowmik D. Using GANs with adaptive training data to search for new molecules. J Cheminform 2021;13:14.

16. Manzoni C, Kia DA, Vandrovcova J, Hardy J, Wood NW, Lewis PA, et al. Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences. Brief Bioinform 2018;19:286-302.

17. Dias R, Torkamani A. Artificial intelligence in clinical and genomic diagnostics. Genome Med 2019;11:70.

18. Sato K, Hamada M. Recent trends in RNA informatics: a review of machine learning and deep learning for RNA secondary structure prediction and RNA drug discovery. Brief Bioinform 2023;24:186.

19. Liu XQ, Li BX, Zeng GR, Liu QY, Ai DM. Prediction of long non-coding RNAs based on deep learning. Genes 2019;10:273.

20. Lam JH, Li Y, Zhu L, Umarov R, Jiang H, Heliou A, et al. A deep learning framework to predict binding preference of RNA constituents on protein surface. Nat Commun 2019;10:4941.

21. Rauschert S, Raubenheimer K, Melton PE, Huang RC. Machine learning and clinical epigenetics: a review of challenges for diagnosis and classification. Clin Epigenetics 2020;12:51.

22. Huang S, Cai N, Pacheco PP, Narrandes S, Wang Y, Xu W. Applications of Support Vector Machine (SVM) learning in cancer genomics. Cancer Genom Proteom 2018;15:41-51.

23. Schauperl M, Denny RA. AI-based protein structure prediction in drug discovery: impacts and challenges. J Chem Inf Model 2022;62:3142-56.

24. Liu L, Zhu X, Ma Y, Piao H, Yang Y, Hao X, et al. Combining sequence and network information to enhance protein-protein interaction prediction. BMC Bioinformatics 2020;21:537.

25. Pakhrin SC, Pokharel S, Saigo H, Kc DB. Deep learning-based advances in protein posttranslational modification site and protein cleavage prediction. Methods Mol Biol 2022;2499:285-322.

26. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature 2021;596:583-9.

27. Mathieu A, Leclercq M, Sanabria M, Perin O, Droit A. Machine learning and deep learning applications in metagenomic taxonomy and functional annotation. Front Microbiol 2022;13:811495.

28. Harris ZN, Dhungel E, Mosior M, Ahn TH. Massive metagenomic data analysis using abundance-based machine learning. Biol Direct 2019;14:12.

29. Bhattacharjee A, Bayzid MS. Machine learning based imputation techniques for estimating phylogenetic trees from incomplete distance matrices. BMC Genomics 2020;21:497.

30. Dasguptaa A, Deb RK. Artificial intelligence in systems biology. Artificial Intelligence 2023;49:153.

31. Goshisht MK. machine learning and deep learning in synthetic biology: key architectures, applications, and challenges. ACS Omega 2024;9:9921-9945.

32. Strianese O, Rizzo F, Ciccarelli M, Galasso G, D'Agostino Y, Salvati A, et al. Precision and personalized medicine: how genomic approach improves the management of cardiovascular and neurodegenerative disease. Genes 2020;11:747.

33. Parekh A-DE, Shaikh OA, Simran F, Manan S, Al Hasibuzzaman M. AI in personalized medicine: AI-generated personalized therapy regimens based on genetic and medical history, short communication. Ann Med Surg 2023;85:5831-5833.

34. Sun TH, Wang CC, Wu YL, Hsu KC, Lee TH. Machine learning approaches for biomarker discovery to predict large-artery atherosclerosis. Sci Rep 2023;13:15139.

35. Mathema VB, Sen P, Lamichhane S, Orešič M, Khoomrung S. Deep learning facilitates multi-data type analysis and predictive biomarker discovery in cancer precision medicine. Comput Struct Biotechnol J 2023;21:1372-1382.

36. Raikar GVS, Raikar AS, Somnache SN. Advancements in artificial intelligence and machine learning in revolutionising biomarker discovery. Braz J Pharm Sci 2023;59:23146.

37. Khalili-Tanha G, Mohit R, Asadnia A, Khazaei M, Dashtiahangar M, Maftooh M, et al. Identification of ZMYND19 as a novel biomarker of colorectal cancer: RNA-sequencing and machine learning analysis. Cell Commun Signal 2023;17:1469-85.

38. Khojasteh-Leylakoohi F, Mohit R, Khalili-Tanha N, Asadnia A, Naderi H, Pourali G, et al. Down regulation of Cathepsin W is associated with poor prognosis in pancreatic cancer. Sci Rep 2023;13:16678.

39. Chang CH, Lin CH, Lane HY. Machine learning and novel biomarkers for the diagnosis of Alzheimer's disease. Int J Mol Sci 2021;22:2761.

40. Sayed MA, Cao DM, Islam MT, Tayaba M, Pavel MEUI, Mia MT, et al. Parkinson's disease detection through vocal biomarkers and advanced machine learning algorithms. Journal of Computer Science and Technology Studies 2023;5:142-9.

41. Califf RM. Biomarker definitions and their applications. Exp Biol Med 2018;243:213-21.

42. Nalejska E, Mączyńska E, Lewandowska MA. Prognostic and predictive biomarkers: tools in personalized oncology. Mol Diagn Ther 2014;18:273-84.

43. Prelaj A, Miskovic V, Zanitti M, Trovo F, Genova C, Viscardi G, et al. Artificial intelligence for predictive biomarker discovery in immuno-oncology: a systematic review. Ann Oncol 2024;35:29-65.

44. Hussain S, Mubeen I, Ullah N, Shah SSUD, Khan BA, Zahoor M, et al. Modern diagnostic imaging technique applications and risk factors in the medical field: A review. Biomed Res Int 2022;2022.

45. Min S, Lee B, Yoon S. Deep learning in bioinformatics. Brief Bioinform 2017;18:851-69.

46. Lee J-G, Jun S, Cho Y-W, Lee H, Kim GB, Seo JB, et al. Deep learning in medical imaging: general overviewKorean J Radiol 2017;18:570-84.

47. de Haan K, Rivenson Y, Wu Y, Ozcan A. Deep-learning-based image reconstruction and enhancement in optical microscopy. Proc IEEE 2019;108:30-50.

48. Gong K, Berg E, Cherry SR, Qi J. Machine learning in PET: from photon detection to quantitative image reconstruction. Proc IEEE 2019;108:51-68.

49. Ardeti VA, Kolluru VR, Varghese GT, Patjoshi RK. An Overview on State-of-the-Art electrocardiogram signal processing methods: traditional to ai-based approaches. Expert Syst Appl 2023:119561.

50. Huang Z, Wang M. A review of electroencephalogram signal processing methods for brain-controlled robots. Cognitive Robotics 2021;1:111-24.

51. AlHinai N. Introduction to biomedical signal processing and artificial intelligence. Biomedical signal processing and artificial intelligence in healthcare: Elsevier; 2020. p. 1-28.

52. An X, Kuang D, Guo X, Zhao Y, He L, editors. A deep learning method for classification of EEG data based on motor imagery. Intelligent Computing in Bioinformatics: 10th International Conference, ICIC 2014, Taiyuan, China, August 3-6, 2014 Proceedings 10; 2014: Springer.

53. Jia X, Li K, Li X, Zhang A, editors. A novel semi-supervised deep learning framework for affective state recognition on EEG signals. 2014 IEEE international conference on bioinformatics and bioengineering; 2014: IEEE.

54. He F, Liu K, Yang Z, Hannink M, Hammer RD, Popescu M, et al. Applications of cutting-edge artificial intelligence technologies in biomedical literature and document mining. Med Rev 2023;3:200-4.

55. Zhu F, Patumcharoenpol P, Zhang C, Yang Y, Chan J, Meechai A, et al. Biomedical text mining and its applications in cancer research. J Biomed Inform 2013;46:200-11.

56. Leser U, Hakenberg J. What makes a gene name? Named entity recognition in the biomedical literature. Brief Bioinform 2005;6:357-69.

57. Ephraim Y, Merhav N. Hidden markov processes. IEEE Transactions on information theory. 2002;48:1518-69.

58. Habib MS, Kalita J. Scalable biomedical Named Entity Recognition: investigation of a database-supported SVM approach. Int J Bioinform Res Appl 2010;6:191-208.

59. He Y, Kayaalp M, editors. Biological entity recognition with conditional random fields. AMIA Annual Symposium Proceedings; 2008: American Medical Informatics Association.

60. Saha SK, Sarkar S, Mitra P. Feature selection techniques for maximum entropy based biomedical named entity recognition. J Biomed Inform 2009;42:905-11.

61. Makino T, Ohta Y, Tsujii Ji, editors. Tuning support vector machines for biomedical named entity recognition. Proceedings of the ACL-02 workshop on Natural language processing in the biomedical domain; 2002.

62. Tsai Th, Chou WC, Wu SH, Sung TY, Hsiang J, Hsu WL. Integrating linguistic knowledge into a conditional random field framework to identify biomedical named entities. Expert Syst Appl 2006;30:117-28.

63. Chafai N, Bonizzi L, Botti S, Badaoui B. Emerging applications of machine learning in genomic medicine and healthcare. Crit Rev Clin Lab Sci 2024;61:140-63.

64. Han H, Liu W. The coming era of artificial intelligence in biological data science. BMC Bioinformatics 2019;20:712.

65. Cai YD, Liu XJ, Xu X, Zhou GP. Support vector machines for predicting protein structural class. BMC Bioinformatics 2001;2:3.

66. Strokach A, Kim PM. Deep generative modeling for protein design. Curr Opin Struct Biol 2022;72:226-36.

67. Wu FX. Genetic weighted k-means algorithm for clustering large-scale gene expression data. BMC Bioinformatics 2008;9:12.

68. Wei D, Jiang Q, Wei Y, Wang S. A novel hierarchical clustering algorithm for gene sequences. BMC Bioinformatics 2012;13:174.

69. Pragya, Govarthan PK, Sinha K, Mukherjee S, Agastinose Ronickom JF. Differential gene expression data analysis of asd using random forest. Stud Health Technol Inform 2023;302:1047-51.

70. Hoque MN, Sarkar MMH, Khan MA, Hossain MA, Hasan MI, Rahman MH, et al. Differential gene expression profiling reveals potential biomarkers and pharmacological compounds against SARS-CoV-2: Insights from machine learning and bioinformatics approaches. Front Immunol 2022;13:918692.

71. Abbas M, El-Manzalawy Y. Machine learning based refined differential gene expression analysis of pediatric sepsis. BMC Med Genomics 2020;13:122.

72. Collienne L, Gavryushkin A. Computing nearest neighbour interchange distances between ranked phylogenetic trees. J Math Biol 2021;82:8.

73. Lin Y, Hu F, Tang J, Moret BM. Maximum likelihood phylogenetic reconstruction from high-resolution whole-genome data and a tree of 68 eukaryotes. Pac Symp Biocomput 2013:285-96.

74. Larjo A, Shmulevich I, Lahdesmaki H. Structure learning for Bayesian networks as models of biological networks. Methods Mol Biol 2013;939:35-45.

75. Liu F, Heiner M, Gilbert D. Hybrid modelling of biological systems: current progress and future prospects. Brief Bioinform 2022;23.

76. Peng J, Jury EC, Donnes P, Ciurtin C. Machine learning techniques for personalised medicine approaches in immune-mediated chronic inflammatory diseases: applications and challenges. Front Pharmacol 2021;12:720694.

77. Guadiana-Alvarez JL, Hussain F, Morales-Menendez R, Rojas-Flores E, Garcia-Zendejas A, Escobar CA, et al. Prognosis patients with COVID-19 using deep learning. BMC Med Inform Decis Mak 2022;22:78.

78. Sarvamangala DR, Kulkarni RV. Convolutional neural networks in medical image understanding: a survey. Evol Intell 2022;15:1-22.

79. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. Med Image Anal 2017;42:60-88.

80. Zeng Z, Shi H, Wu Y, Hong Z. Survey of natural language processing techniques in bioinformatics. Comput Math Methods Med 2015;2015:674296.

81. Liu L, Tang L, Dong W, Yao S, Zhou W. An overview of topic modeling and its current applications in bioinformatics. Springerplus 2016;5:1608.