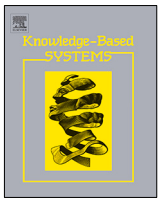
Knowledge-Based Systems 279 (2023) 110937

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys



Artificial intelligence-driven biomedical genomics

Kairui Guo [a,*], Mengjia Wu [a], Zelia Soo [a], Yue Yang [a], Yi Zhang [a], Qian Zhang [a], Hua Lin [b], Mark Grosser [b], Deon Venter [b], Guangquan Zhang [a], Jie Lu [a]

[a] Australian Artificial Intelligence Institute, Faculty of Engineering and Information Technology, University of Technology Sydney, 61 Broadway, Ultimo, 2007, New South Wales, Australia
[b] 23Strands, 26-32 Pirrama Road, Pyrmont, 2009, New South Wales, Australia

ARTICLE INFO

Article history:
Received 26 June 2023
Received in revised form 14 August 2023
Accepted 16 August 2023
Available online 7 September 2023

Keywords:
Artificial Intelligence
Machine learning
Genomics
Biomedicine

ABSTRACT

As genomic research becomes more complex and data-rich, artificial intelligence (AI) has emerged as a crucial tool for processing and analyzing high-dimensional genomic data, accelerating biomarker discovery, and enhancing genomic sequence annotations. Despite the increasing application of AI in genomic research, challenges persist, particularly regarding the integration of biomedical knowledge into algorithm development. We reviewed high-quality, AI-driven biomedical genomic studies from the past five years, covering applications in disease prediction, detection, diagnosis, and treatment. Each category highlights how different AI techniques are applied in biomedical contexts. Furthermore, we identify current challenges and potential solutions in AI-assisted biomedical genomics. This comprehensive review is designed to encourage collaboration among computer scientists, healthcare professionals, and interested communities, propelling the development of AI applications that can be smoothly integrated into routine medical services.

© 2023 Elsevier B.V. All rights reserved.

1. Introduction

Genetic and genomic research has evolved from Mendel's inheritance studies of pea plants to a complex clinical tool with diverse applications, including newborn screening, predictive and diagnostic methods for diseases, and precision medicine based on pharmacogenomic findings. Initiated in 1990 and completed in 2003, the Human Genome Project provided fundamental information about the human blueprint. Since then, genomics has accelerated the study of human biology and improved the practice of medicine [1]. In the post-genome era, efforts to apply genomic information to biomedicine have become increasingly active.

The emerging biomedical applications inspired by genomics have posed several challenges. Firstly, deciphering hidden information from the whole genome, gene transcription/expression, and phenotypic data requires intricate non-linear modeling. Secondly, the heavy interlinking and high dimensionality of different genomic data types exacerbate the difficulty. Noisy, incomplete, or unmatched data with insufficient labels further increases the difficulty of analysis. At the same time, much genomic discovery has occurred through individuals of European ancestry, with a limited representation of other populations [2]. Additionally, when these applications were tested in a clinical environment, issues such as the proliferation of variant-specific therapies and prognostic analyses increased the complexity of implementation for medical specialists [3].

Artificial intelligence (AI), which encompasses machine learning (ML), computer vision, neural networks, and natural language processing, has emerged as an indispensable tool for addressing these challenges. It empowers the processing, analysis, modeling, and interpretation of large-scale genomic data. Across various stages of healthcare service, AI algorithms have been employed to answer diverse questions, from biomarker discovery studies [4,5] to annotating genomic sequence elements [6].

Despite numerous reviews exploring the application of AI in genetic and genomic research across various diseases [6–11], there is still significant potential to improve the performance and usability of AI-driven genomic applications in biomedical and clinical settings. Two critical aspects that merit additional focus include integrating biomedical knowledge into genomic-specific AI, and ensuring that adequate training is provided for both the developers and users of these applications. Several survey papers have attempted to address these challenges from diverse perspectives. Some have focused on molecular medicine and systems biology [6,7,12], while others have delved into the progress and specific methods intersecting early multi-modal methodologies such as regression, computer vision and genomic enrichment in particular medical fields such as brain imaging genomics [13]. However, previous surveys have not comprehensively covered the use of more advanced techniques, including transfer learning and recommender systems.

* Corresponding author.
E-mail address: kairui.guo@uts.edu.au (K. Guo).

https://doi.org/10.1016/j.knosys.2023.110937
0950-7051/© 2023 Elsevier B.V. All rights reserved.

To further promote the integration of AI algorithms in biomedicine and establish a shared comprehension of two rapidly evolving fields – AI and genomics – we believe an up-to-date review centered on AI algorithm development and application in biomedical genomic analysis, will be invaluable. Therefore, this paper aims to serve as a starting point for collaboration among computer scientists, healthcare professionals, and other interested communities and organizations. Our goal is to stimulate joint efforts in developing AI applications that integrate into routine medical services.

This paper presents a systematic review of recent advances in AI-driven biomedical genomic applications, which can potentially improve clinicians' work. We developed a framework to identify relevant academic articles on AI-related biomedical applications. Applying this framework specifically to genomics, we reviewed 82 high-quality, AI-driven biomedical genomic studies from the past five years. Our analysis discusses the strengths of AI techniques, examines their usability at specific stages of healthcare for various health conditions, and addresses the challenges and potential solutions for future research in the field.

The main contributions of this paper are:

(i) It presents a timely review of AI applications in biomedical genomics over the past five years, emphasizing the contributions of AI in enhancing disease prediction, diagnosis, and treatment.
(ii) It devises a literature search framework for biomedical analysis using advanced bibliometric methods.
(iii) It identifies challenges in contemporary AI-driven genomic research and proposes potential solutions.
(iv) It serves as the training material for researchers, physicians, and industry partners who are interested in the convergence of AI and genomics, providing fundamental knowledge and the latest applications in this field.

The rest of the paper is structured as follows: Section 2 presents a bibliometric literature review framework for AI in medicine. Section 3 details the AI techniques applied in the genomic studies reviewed. Section 4 through 6 delve into biomedical and clinical applications, covering disease prediction, early detection, diagnosis, treatment, and prognosis analysis. Section 7 outlines current challenges and potential solutions in AI-assisted biomedical genomics, spanning from algorithm development to social issues. Our aim is to provide a thorough analysis showcasing the latest trends in AI-driven genomics and equip readers with the knowledge necessary for future progress in this domain.

2. Literature review and a framework for AI in biomedical genomics

Electronic databases, PubMed and Web of Science, were utilized to identify relevant clinical genomic applications. PubMed is considered to be the most extensive open-source database of biomedical literature worldwide, encompassing more than 30 million articles and online books. It is widely regarded as the optimal data source for gathering literature related to biomedical and life sciences [14]. On the other hand, the Web of Science is a widely recognized multidisciplinary scholarly database, which accumulates over 74.8 million scientific publications and concentrates on high impact journals [15]. To ensure comprehensive coverage of both biomedical and computer science domains, we have integrated data from both sources. The following major search keywords and MeSH terms were used to retrieve relevant studies: artificial intelligence (and representative techniques), genom*, disease, diagnos*, treatment, precision medicine, and personali* medicine. The keywords with an asterisk symbol (*) were used to include synonyms and solve the issue of American and British English spelling differences. The date range was from January 2018 to February 2023, representing studies reported within the last five years.

The literature search process yielded 442 papers of original research studies. Aiming to profile the research landscape of this domain, we mapped the collected papers to the OpenAlex database via digital objective identifier matching, through this process, we obtained a collection of hierarchically-organized concepts associated with the papers. These concepts represent Wikipedia entries that were mapped to the research papers using a topic modeling approach [16]. The hierarchical structure of the relevant concepts is visually presented in Fig. 1. The figure illustrates the identification of four distinct concept groups derived from the analyzed papers. These groups are categorized as follows: #1 biomedicine, #2 clinical issues, #3 computer science methods, and #4 other.

Group #1 encompasses a diverse range of genomic data utilized in modeling analysis and biomedical research, particularly within the context of AI and genomics applications.

Group #2 focuses on the clinical issues that can be addressed or improved through the application of AI in genomics. In addition to the cancers covered in Group #1, this branch highlights several other diseases, including dementia, heart disease, autism, and COVID-19. The inclusion of these diseases underscores the immense potential of utilizing genomics and AI in enhancing clinical diagnosis, treatment, and research efforts related to these conditions.

Group #3 represents a wide spectrum of computer science methods that are relevant to AI applications in genomics. The prevalence indicates the popularity and effectiveness in studies exploring the application of AI in genomics. Additionally, two other important concepts, natural language processing and computer vision, are identified.

Group #4 presents a comprehensive analysis of various societal and pertinent issues. The utilization of AI methodologies to enhance genomic research is not only rooted in clinical endeavors but also holds substantial implications for population health and the overall welfare of the public.

Fig. 2 provides valuable insights into the prevailing trends within the application of AI techniques in the realm of genomics-enabled clinical problems. It becomes evident that cancer research continues to command the most substantial attention across all AI methodologies. The integration of various sequencing data types, such as gene expression, human genome, microRNA, and DNA, as input sources for AI approaches, is a recurring practice. Within the domain of interdisciplinary studies, conventional machine learning methods find notable applications in gene expression analysis and genome data exploration. Deep learning architectures have been broadly applied in cancer research and the investigation of genome data.

As shown in Fig. 3, studies had to meet the following criteria to be eligible for inclusion: (1) full-text peer-reviewed publications in English. (2) Original studies that include both human genomic data as input and AI techniques as methods were selected, and reviews, editorials, or letters were removed. Duplicate studies were also removed. (3) Studies published before 2018 were removed. A minimum threshold of 20 citations was set to ensure the quality of the study. Finally, the eligible full texts were independently screened by four reviewers (KG, MW, ZS, and YY), who also conducted a manual search to include all relevant studies. In total, 82 papers fitting the established criteria were thoroughly examined and included in our analysis. Three categories, namely prediction and early detection, disease diagnosis, treatment and prognosis, were identified among the included papers. Under each category, AI techniques including conventional machine learning, deep neural network (DNN), transfer learning (TL), computer vision (CV), graph representation learning (GRL), and natural language processing (NLP) were recognized, and discussed in detail later.

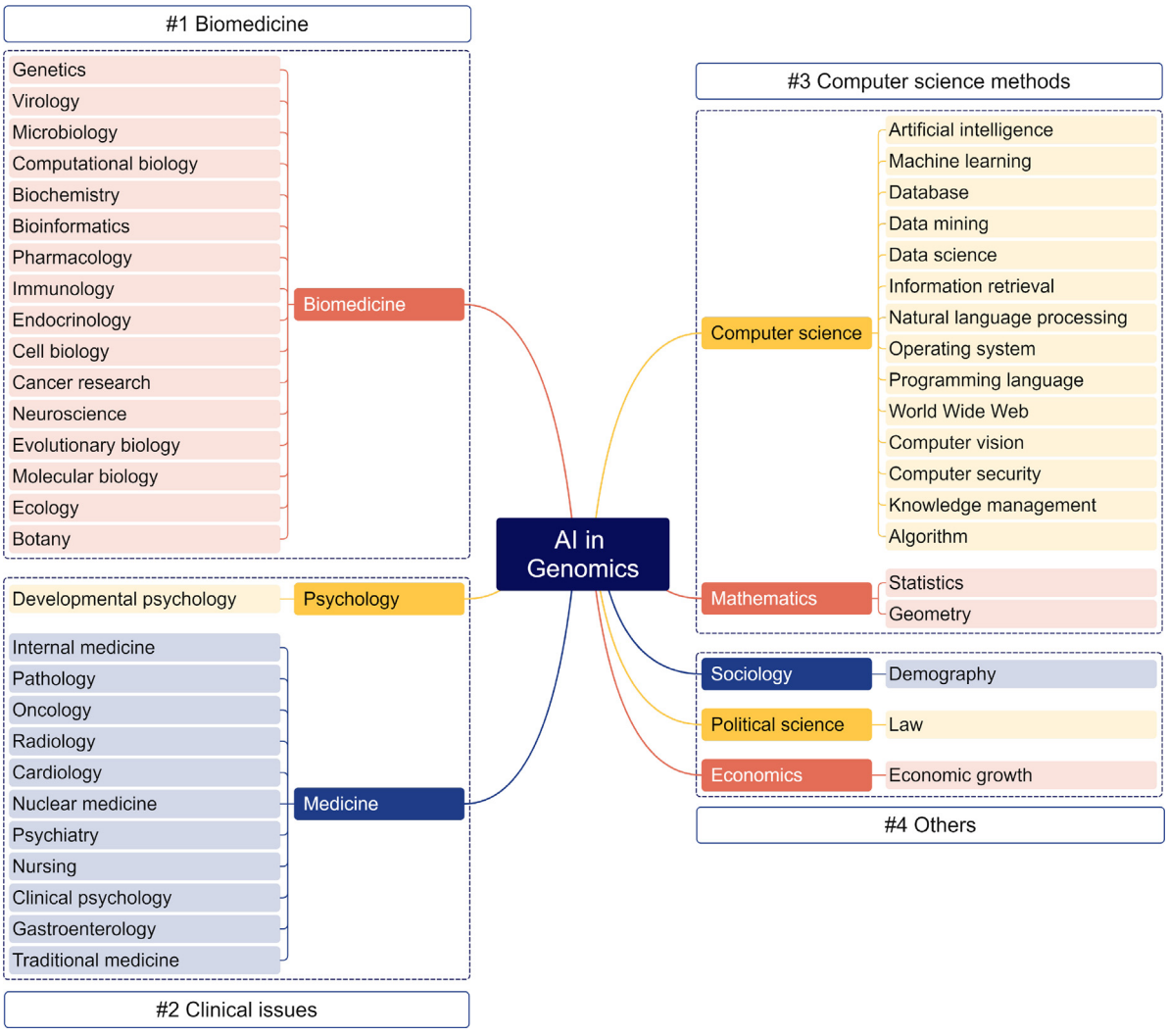


Fig. 1. The hierarchical structure of concepts showing the four groups.

3. Main AI techniques used in genomic applications

In AI-driven applications, a preprocessing step is essential before the actual implementation of the algorithms. Data preprocessing, a concept well-established in Big Data, ensures that the data conforms to the specifications required by the AI algorithms applied in subsequent stages [17]. In genomics, preprocessing addresses challenges such as noisy and missing data, as well as the classic 'curse of dimensionality' often encountered in AI. To address the challenges posed by noise from various sources and missing values, several software programs have been developed in recent years [18–20]. There are also specific tools tailored to handle the dimensionality issues associated with sequencing data [21,22].

In this section, we introduce the AI techniques used in clinical genomic applications from a conceptual level. We emphasize how each method analyses knowledge from data, obtains patterns by building unique models, and makes predictions and classifications on new data. CV and NLP are commonly recognized as perception tasks in computer science [23], as their role is to interpret texts and images to reveal hidden information. These two techniques are included because of their significant contribution to clinical genomic applications.

3.1. Conventional machine learning

Conventional machine learning refers to ML methods that facilitate computer learning from a given dataset, thereby improving task performance based on that learning. These methods, which include linear regression, logistic regression, decision trees, random forests (RF), support vector machines (SVMs), and neural networks, typically employ classic statistical models and equations to identify underlying patterns and relationships within the data. Several recent review articles [24–26] have illuminated these techniques in the context of medicine, leading to their popularity in biomedical genomic applications in numerous COVID-19 and cancer studies.

3.2. Deep neural networks

When the number of layers between the input layer and the output layer is large enough, we consider them as deep neural networks. Currently, the hidden layers of DNNs range from five to more than a thousand [27]. In AI-assisted genomic applications, one of the most popular DNNs, convolutional neural networks (CNNs), employ a unique architecture, convolution layers, together with pooling layers that minimize the number of network parameters, which helps avoid overfitting [28].

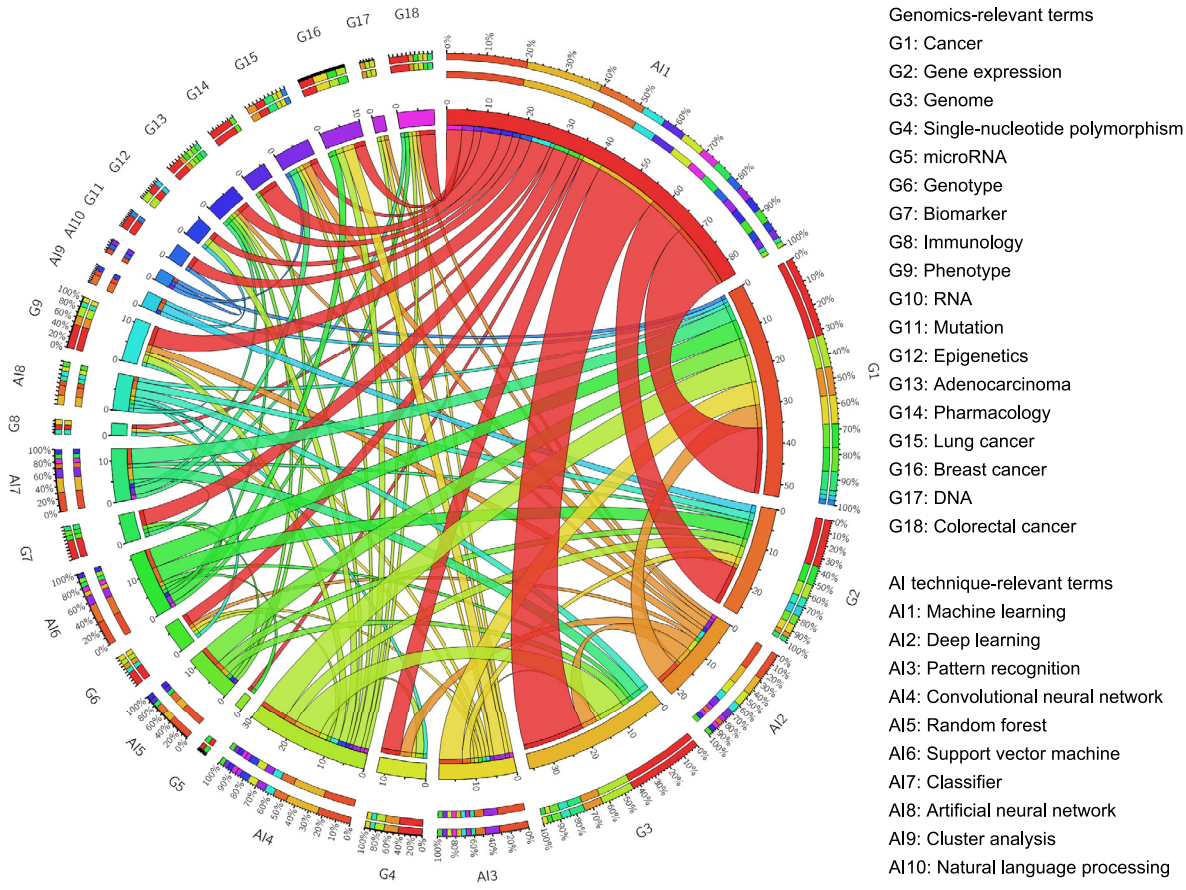


Fig. 2. The co-occurrence network of genomic/clinical and AI concepts.

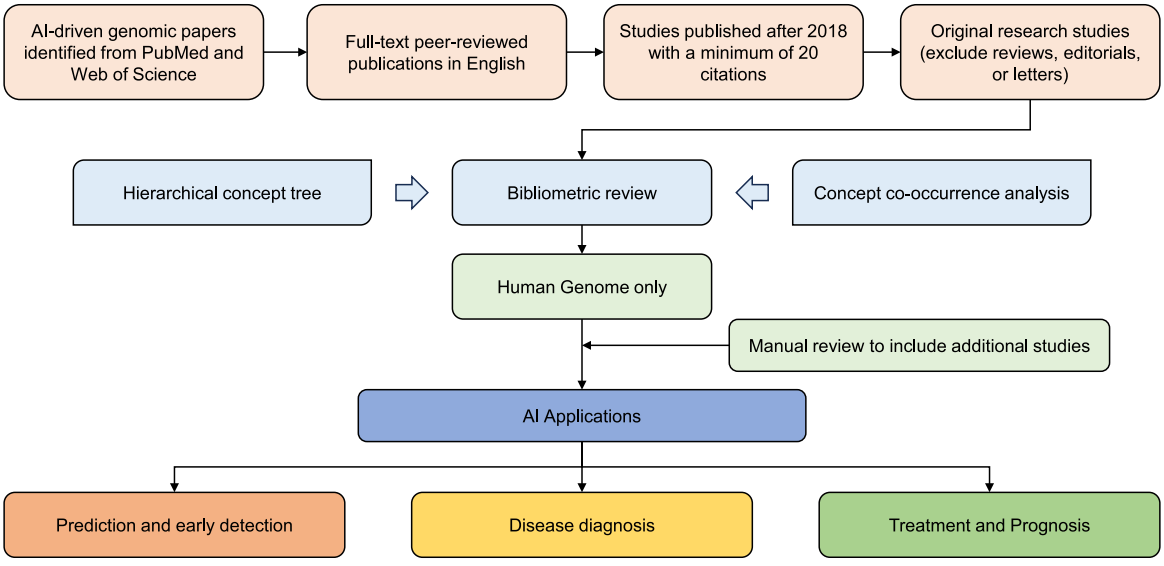


Fig. 3. The literature review framework and AI-driven biomedical genomics categories.

In situations where biomedical data have been collected over time, recurrent neural networks (RNNs) [29] offer a solution. While CNNs perform exceptionally in image-related research, they fail to account for the time dimension. RNNs incorporate recurrent connections in deep neural networks, thus capturing the time element. However, when sequential data spans a lengthy period, RNNs can struggle to capture features that present with a large time gap. To tackle this, the Long Short-Term Memory

(LSTM) structure [30] and the Gated Recurrent Units (GRU) [31] were introduced. These models use gates to control the information flow and are well-equipped to analyze biomedical data collected over extended periods, as demonstrated in a Type 2 Diabetes prediction study [32].

3.3. Transfer learning

Transfer learning aims to exploit the knowledge accumulated from data in auxiliary domains to improve learning in a target domain that does not have sufficient labeled data or none to train a model [33]. It has three main types: inductive, transductive, and unsupervised.

Inductive TL [34] has requirements that the target domain must have some labeled data and the same feature space as the source domain, although they have different data distributions. Moreover, the quantity of labels required for training the predictive function in the target domain highly depends on the complexity of the learning task, the similarity between the two domains, and the availability of labeled information from the source domain(s).

In contrast to inductive TL, transductive TL [35] allows the source and target domains to have different feature spaces. The most active sub-field of transductive TL is domain adaptation, which deals with scenarios where there is a limited amount of labeled source data and unlabeled target data available for training. Existing domain adaptation algorithms primarily concentrate on reducing the distance between marginal or conditional distributions through two approaches: symmetrical training, utilized in feature-based algorithms, and asymmetrical training, used in instance-based domain adaptation.

Unsupervised TL [34] can learn key feature representations of the data without target labels. Instead, the algorithm is tasked to find underlying patterns and structures within the data to create a more compact and meaningful representation. These representations can be thought of as compressed versions of the source data that preserve its essential features. One new technique used in unsupervised transfer learning is self-supervised learning, under which the model is trained to predict certain properties or relationships within the data itself [36].

Transfer learning is often used to improve the performance of machine learning models on smaller medical datasets by leveraging the knowledge obtained from well-labeled large datasets. In genomics, TL shows its particular utility in cross-population studies [37].

3.4. Computer vision

Computer vision is a discipline that employs mathematical techniques to extract the three-dimensional shape and appearance of objects in imagery [38]. The advancement of deep learning-based computer vision architectures has found extensive applications across diverse industries. In the context of biomedical genomic research, four popular architectures: AlexNet [39], VGGNet [40], GoogLeNet [41] and ResNet [42] are introduced.

AlexNet utilized five convolutional layers and three fully connected layers, integrating strategies like ReLU activation, local normalization, overlapping pooling, dropout techniques, and data augmentation. VGGNet comprises 16 convolutional and three dense layers [43]. Concurrently, GoogLeNet uses one-twelfth of the parameters that of AlexNet, offering a solution for resource-limited scenarios. ResNet emerged with a remarkably deep 152-layer architecture, introducing a deep residual learning framework to address the issue of accuracy saturation. Through the addition of shortcut connections, it allowed for skipping layers during the feedforward process, ensuring effective training even with hundreds of layers, without the addition of extra parameters or computational complexity. These architectures are frequently utilized in linking imaging phenotypes to the tumor genetic profile [44].

3.5. Graph representation learning

Graph representation learning methods aim to encapsulate graph data, including topology and attributes, into low-dimensional vectors [45]. Traditional techniques such as dimension reduction [46], random walk [47,48], and matrix factorization [49] focus on node topological similarity, but struggle with scalability and context-awareness in complex, large-scale graphs. More accurate representation models have been introduced through various graph neural network (GNN) architectures. For a deep dive into GRL and GNNs, see surveys [50–52]. GRL techniques have broad applicability in genomics. They enable the conversion of genomic sequencing data into graph structures by considering gene associations or expression-based similarities [52–54]. Moreover, the incorporation of supplementary molecular components or biomedical entities [55,56] and their established associations into the graph models enables the formulation of heterogeneous graph representations. The heterogeneous networks are particularly useful when dealing with various types of genomic data [57].

3.6. Natural language processing

Natural Language Processing focuses on developing techniques for human–computer interaction through natural language. Its core objective is to analyze and process vast amounts of unstructured textual data, enabling machines to comprehend human language and generate human-like responses [58]. The main tasks of NLP development involve realizing natural language understanding and generation [59]. To achieve these outcomes, a variety of language models and neural network architectures have been developed. For more on NLP technical progress, refer to [60,61]. In genomic analysis, NLP primarily serves as an auxiliary technique aimed at transforming potentially relevant textual data, such as Electronic Health Records (EHR), into computable features. These features can then be applied to Genome-Wide Association Studies (GWAS) and Phenome-Wide Association Studies (PheWAS).

The following three chapters will offer a comprehensive review of biomedical genomic applications with a focus on the AI techniques introduced in this chapter. A mind map, depicted in Fig. 4, is provided to guide readers.

4. AI-driven prediction and early detection through genomic data

In healthcare, disease prediction and early detection are critical in clinical decision-making, enabling physicians to identify and manage patients at high risk of adverse outcomes. Conventional machine learning, deep neural networks, and transfer learning have all found applications in prediction and early detection.

4.1. Conventional machine learning for prediction

Genomic analysis for disease predictions has often employed conventional ML classification and regression methods. Notable models were developed to study SARS-CoV-2. Shrock et al. [62] used an oligonucleotide library incorporating multiple strains' proteomic profiles of SARS-CoV-2 and human patient immune responses, analyzed via a gradient boosting algorithm and logistic regression classifier to establish the VirScan platform. This differentiated between severe and mild symptom cases, aiding disease

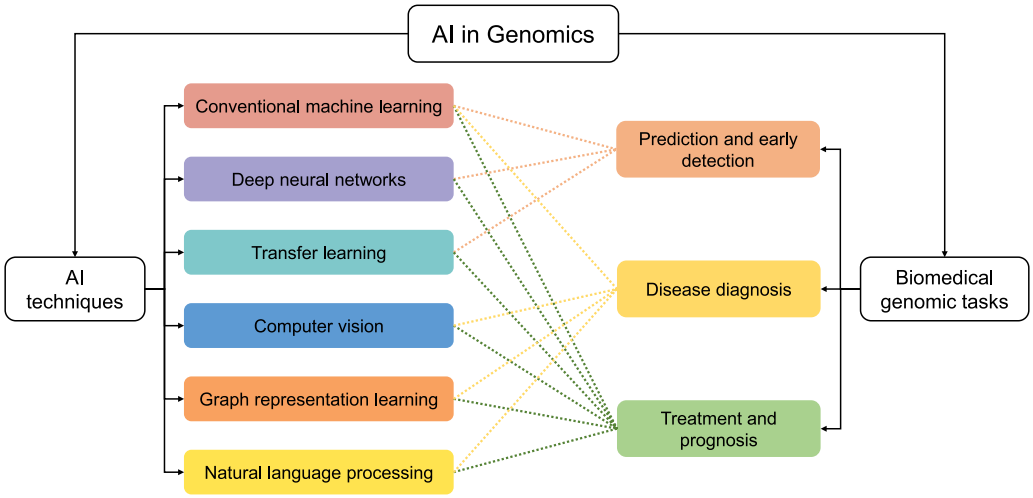


Fig. 4. A mind map of the AI techniques and biomedical topics reviewed.

severity prediction. Wang et al. [63] sequenced over 500,000 SARS-CoV-2 genomes, using gradient-boosting trees and protein–protein binding data to predict mutation changes and their impact on COVID-19 severity.

Cancer stage and survival predictions using genomic data have also been investigated. One study employed a regularized non-negative matrix factorization method using RNA-seq and protein interaction data from various breast cancer stages to predict patient outcomes and aid in personalized care planning [64]. A breast cancer protein classification study found a multilayer perceptron neural network with one hidden layer of 20 neurons to be the most effective [65]. In gastric cancer, a binary logistic regression model evaluated mRNA signatures for lymph node metastasis prediction [66], while the LASSO method identified biomarkers for differentiated thyroid cancer [67]. Random Forest was used for a prostate cancer progression biomarker study [68].

Conventional ML has been widely used in neurological condition studies. For instance, a Bayesian network identified a five-fold increase in amyotrophic lateral sclerosis-associated genes compared to GWAS-discovered loci [69]. Gradient-boosted decision trees were used for Parkinson's disease gene prediction analysis [70], and a neural network predicted multiple sclerosis risk based on single nucleotide polymorphisms [71].

Other applications, like kidney transplants [72], post-menopausal osteoporosis [73], and atherosclerosis [74], also used ensemble methods and SVMs for genetic biomarker identification. Despite the widespread adoption of conventional ML due to its simplicity, these studies often lack comprehensive model-building analysis, which can impede larger cohort experiments in later stages.

4.2. Deep learning for prediction and early detection

Deep learning is a valuable tool for genomic analysis in predicting common chronic diseases. In the Type-2 diabetes prediction application conducted by Srinivasu et al. [75], RNNs were selected to quickly memorize features from previous cycles and process genomic data. Bidirectional LSTM and GRU were tested using data from the UK Biobank. The results demonstrated that RNNs outperformed other AI algorithms in predicting the possibility of Type-2 diabetes occurring in the future.

In the field of cancer research, DNN-based applications have been applied for detection purposes. Zhang et al. [76] used cell-free DNA fragmentomics to achieve high sensitivity and specificity for the early detection of primary liver cancer. Extracted bioinformatic features were inputted into three algorithms: gradient-boosting machine, RF, and DNN, with DNN being able to identify the most significant features. Loeffler et al. [77] trained a deep neural network to detect FGFR3 mutations in bladder cancer using histology images. The proposed system using ShuffleNet [78] with two output layers could be used as a diagnostic tool for bladder cancer detection. Compared to basic machine learning methods, deep learning approaches enhance the performance of genomic-based disease prediction by introducing novel features. However, these applications also demand more computational power.

4.3. Transfer learning for disease prediction

Transfer learning has enhanced the predictive performance for the risk of developing numerous diseases by adapting pre-trained models to the specific characteristics of healthcare data and fine-tuning them on smaller, domain-specific datasets. For example, Jonsson et al. [79] applied TL to brain age prediction using a deep CNN pre-trained on a dataset of healthy Icelanders. They were able to improve the accuracy of brain-age prediction on two different datasets through fine-tuning, thereby demonstrating the potential of transfer learning to enhance clinical prediction tasks in healthcare.

Transfer learning has also been utilized to predict trends in patient data, providing new insights into clinical findings from genotypic information. Dong et al. [80] enhanced the accuracy of predicting functional variants in regulatory elements by incorporating knowledge from large genomic datasets using transfer learning. Similarly, Zhuang et al. [81] and Zhou et al. [82] applied TL to fine-tune models trained on extensive datasets to improve predictive performance on enhancer–promoter interactions and to impute missing RNA-sequencing data, respectively. Taroni et al. [83] introduced a transfer learning framework called MultiPLIER for predicting rare diseases using large-scale gene expression datasets. This improved the description of biological processes related to disease severity. These methodologies underscore the effectiveness of transfer learning in linking molecular features to phenotype, thereby enhancing our understanding of complex diseases and potential therapeutic targets in genomic research.

5. AI-driven disease diagnosis through genomic data

Early disease detection is a pivotal aspect of improving patient outcomes, and genomic data can provide significant assistance in

this process, potentially acting as a diagnostic tool in clinical practice. Algorithms utilizing conventional machine learning, transfer learning, and graph representation learning have been explored as potential early detection and diagnostic tools. Furthermore, graph neural networks and natural language processing have demonstrated their capabilities in gene–disease association studies.

5.1. Conventional machine learning for disease diagnosis

Conventional ML methods have been implemented in various diseases to assist the diagnostic process. In a COVID-19 study, the VirScan platform leveraged a gradient boosting algorithm and a logistic regression classifier to ascertain current or prior SARS-CoV-2 virus infection, achieving 99.1% sensitivity and 98.4% specificity [62]. A similar logistic regression classifier identified differentially expressed genes for sepsis-induced acute respiratory distress syndrome [84]. ML can also pinpoint chronic pain locations and contributing genes. A probabilistic neural network mapped the dorsal root ganglion from mouse models to a primate model and related Genome-Wide Association Study [85]. RF and neural networks have been instrumental in the diagnostic process, uncovering novel genetic risk factors for abdominal aortic aneurysms [86], sarcopenia [87], and non-obstructive azoospermia [88].

Cancer diagnosis is another key application of AI-assisted genomic analysis. A study used a stacked ensemble machine learning model for early liver cancer diagnosis [76]. Although prior methods using cell-free DNA had low sensitivity despite their non-invasive nature, researchers improved model performance by constructing a generalized linear model and enhancing it with ensemble models, surpassing the original design's efficacy. Ensemble models, along with SVMs and shallow neural networks, have been implemented as diagnostic tools in leukemia [89], breast and lung cancer [90,91], endometrial carcinoma [92], and various combinations of cancer types [93,94]. The intuitiveness and ease of use of conventional ML algorithms, courtesy of convenient packages in popular programming languages, have made it common to compare several predefined conventional ML algorithms within a single study.

5.2. Transfer learning for disease diagnosis

Transfer learning has showcased its remarkable potential in improving diagnostic accuracy in healthcare. Incorporating domain-specific knowledge, TL not only improves model interpretability but also supports clinical decision-making. The field of cancer detection has particularly benefitted from transfer learning, as it has been utilized to improve the accuracy of detecting recurrent cancer evolution and identifying recurring mutations. Specifically, Caravagna et al. [95] employed a TL approach, utilizing deep convolutional neural networks pre-trained on extensive genomic datasets and subsequently fine-tuned on multi-region tumor sequencing data. This resulted in improved accuracy in identifying recurrent mutations and inferring their sequence of occurrence. Furthermore, a TL-based algorithm, CTC-Tracer, was developed to address the distributional shift between primary cancer cells and circulating tumor cells (CTCs) [96]. This allowed the transfer of lesion labels from the primary cancer cell atlas to CTCs, thereby enhancing the detection of cancer types using CTCs obtained from liquid biopsies.

Transfer learning has also been employed in fields beyond cancer. For instance, Ge et al. [97] studied the role of circular RNAs in non-obstructive azoospermia, offering valuable insights into the molecular mechanisms underpinning the condition and underscoring the potential of circRNAs as diagnostic and therapeutic targets in male infertility. As such, transfer learning has emerged as a promising method for improving diagnostic accuracy in healthcare, proving its efficacy across a variety of disease-diagnostic tasks. Integrating domain-specific knowledge facilitates clinical decision-making and enhances model interpretability.

5.3. Graph representation learning for disease diagnosis and subtype classification

GRL techniques have been effectively utilized in various genomic biomedical graph datasets to facilitate clinical disease diagnosis. An illustrative example can be found in the work of [98], where graph attention networks were employed on single-cell RNA sequencing data for the purpose of diagnosing multiple sclerosis. The researchers curated a comprehensive dataset comprising 60,667 single-cell samples obtained from individuals affected by multiple sclerosis and proceeded to construct K-nearest neighbor graphs. By leveraging the graph attention network model, they attained an impressive diagnostic accuracy of 92% through the classification of individual cells. Furthermore, Ramirez et al. [99] introduced a novel architecture based on graph convolutional networks (GCNs) to address the multi-class classification of cancers for diagnostic purposes, utilizing gene expression data and protein–protein interaction (PPI) data. The researchers transformed the original RNA sequencing data from 10,340 cancer samples and 731 normal tissue samples, along with the PPI data, into four distinct biological graphs. Subsequently, the GCN model was employed to perform the downstream multi-class classification task. The outcomes of their investigation showcased classification accuracies ranging from 89.9% to 94.7%, thereby highlighting the substantial value of non-coding gene regulations in disease classification for diagnostic purposes.

In addition to distinguishing between healthy individuals and those with illnesses, the differential diagnosis of disease subtypes holds significant importance due to the varying treatment approaches and prognoses associated with different subtypes. GRL-based dimensionality reduction techniques have demonstrated their efficacy in identifying cancer subtypes. This can be exemplified by the work conducted by Rhee et al. [55], who developed a hybrid model that integrates a graph convolutional network and a relation network. The aim was to learn vector representations for patient gene expression profiles. The approach was applied to a cohort of 983 breast cancer patients encompassing four subtypes: Luminal A, Luminal B, HER2, and Basal-like. The hybrid model achieved an overall accuracy of 83.19% in correctly classifying the subtypes. Moreover, Wu et al. [100] collected gene expression data from 402 patients diagnosed with diffuse lower-grade glioma. Through unsupervised clustering analysis, they discovered three distinct immune subtypes: Im1, Im2, and Im3. Employing graph learning-based dimensionality reduction techniques on the data, they revealed the presence of intra-cluster heterogeneity within the Im2 subtype. Similarly, Li et al. [54] conducted unsupervised clustering analysis on gene expression data from a cohort of 1368 patients with squamous cell carcinoma. The clustering analysis unveiled six immune subtypes. To explore the intricacies within these subtypes, the researchers applied a graph structure-based dimensionality reduction method. The results revealed the underlying tree structures of patient immune profiles and identified intra-cluster heterogeneity within immune subtypes 1, 2, 4, and 6. By utilizing GRL-based dimensionality reduction techniques, these studies showcase the capacity to identify disease subtypes, unravel intra-cluster heterogeneity, and shed light on the complex nature of immune profiles in various cancer types.

5.4. Graph neural networks for gene–disease association prediction

Identifying disease genes that lead to the onset and progression of diseases is a long-time challenging issue due to the large volume of gene pleiotropy and limited known pathogenesis for diseases. The guilt-by-association concept proposed in recent biomedical studies indicates that genes with similar features to the causative genes are more likely to be associated with diseases [101]. Inspired by that, leveraging the identified gene–disease associations to learn from genomic data for discovering new candidate disease genes and prioritizing them for further clinical investigation [101–105]. This task is presented as a link prediction (or recommendation) issue in current most research works with graph data and GRL techniques dominantly used. The input data is formulated by homogeneous or heterogeneous nodes that consist of genes, diseases, and other biomedical entities such as phenotypes, proteins, chemicals, etc. Links connecting disease to gene nodes can originate from genome-wide association studies, literature reports, or the integration of multiple sources [106,107]. Moreover, the inclusion of heterogeneous associations, such as disease similarity [101], gene similarity, disease–phenotype associations [103], and gene–chemical associations [108], can provide additional information for context-aware network representation.

Multiple GRL techniques were employed in the present investigation, utilizing diverse types of input networks, resulting in varying prediction accuracy depending on the chosen input network construction scheme and GRL techniques used. In a study conducted by Rhee et al. [55], disease and gene similarity networks were individually constructed, and their embeddings were learned using graph convolutional networks. Subsequently, the disease and gene embeddings were concatenated pairwise to predict association scores between them. Another research endeavor by Wu et al. [100] aimed to integrate disease and gene nodes into the input network, alongside symptom and gene ontology nodes, thereby forming a heterogeneous network input. Furthermore, studies such as [54] expanded upon this approach by incorporating phenotype and pathway nodes into the heterogeneous network, utilizing identified associations from biomedical databases. By employing GRL techniques on the constructed graphs, it becomes possible to align disease and gene vector representations within the same spatial vector space, thereby facilitating downstream tasks, including gene–disease association classification and prediction. However, it is important to note that the development of an appropriate input network and the selection of suitable GRL techniques necessitate further exploration and investigation.

5.5. Natural language processing for genome and phenome association analysis

The utilization of NLP techniques in genomics has led to significant advancements in computational phenotyping, which is considered a crucial application within this domain. In the early stages, computational genotyping approaches heavily relied on keyword search and rule-based methods [109]. However, recent developments in NLP, such as the emergence of deep learning architectures, text embedding techniques, and language models, have transformed phenotyping into a text classification task. Consequently, novel NLP methods have been proposed to yield more accurate phenotyping results.

For instance, in a study conducted by Zhang et al. [110], an unsupervised deep learning model was developed to automatically annotate patient phenotypes within EHRs. The researchers hypothesized that the semantic content of EHRs contains valuable information about phenotypic abnormalities. To capture this information, they constructed an auto-encoder model augmented with a classifier, which facilitated the training of semantic representations of EHRs. This approach enabled the identification of phenotypic abnormalities that possess greater semantic significance within EHRs. Another notable study by Yang et al. [111] proposed the use of word- and sentence-level CNNs for phenotyping patient EHRs. This method allowed users to assess the semantic contributions of individual tokens and provided a certain level of interpretability to the phenotyping results based on token frequencies and categories. By incorporating these mechanisms, researchers were able to gain insights into the phenotypic characteristics present within patient EHRs.

Within the realm of genomic studies, cohorts derived from computational phenotyping play a crucial role in enabling subsequent GWAS and PheWAS. GWAS investigations establish connections between genomic data and the analyzed EHR outcomes, thereby uncovering associations between specific genes and phenotypes [112,113]. Conversely, PheWAS endeavors to examine phenotypes associated with particular genetic variants [114,115]. These association studies are of utmost importance in elucidating the underlying mechanisms behind disease onset and progression.

By leveraging computational phenotyping cohorts, researchers can effectively bridge the gap between genomic information and clinical outcomes, providing valuable insights into the relationship between genes and phenotypes. GWAS studies allow for the identification of genetic variants that are statistically associated with specific traits or diseases, thereby potentially shedding light on the genetic underpinnings of complex phenotypic traits. On the other hand, PheWAS investigations facilitate the exploration of diverse phenotypic manifestations that are linked to a particular genetic variant, enabling a comprehensive understanding of the impact of genetic variations on health and disease. These association studies serve as crucial tools in unraveling the complex interplay between genetic factors and phenotypic expressions, ultimately contributing to our understanding of disease etiology, progression, and potential therapeutic interventions.

In contrast to conventional genomic investigation approaches that necessitate manual annotation and phenotype extraction, the utilization of NLP techniques in research endeavors provides notable advantages in terms of time efficiency and efficacy. A prime illustration of this paradigm is exemplified in the scholarly investigation conducted by Clark et al. [116], wherein an NLP system was developed to extract intricate phenotypic characteristics from electronic health records of pediatric patients afflicted with genetic disorders. The researchers amalgamated the outcomes of genome sequencing with the degree of similarity between the patient's extracted phenome and the anticipated phenotypic attributes of various genetic diseases. Consequently, this fusion enabled the generation of a comprehensive ranking score, facilitating the diagnosis of genetic diseases.

6. AI-driven treatment and prognosis through genomic data

Genomics plays a vital role in personalized medicine by guiding the design of individualized therapies and providing insights into disease prognosis. All six main types of AI techniques mentioned in Section 3 have been implemented in AI-assisted treatment and prognosis analysis studies.

6.1. Conventional machine learning for treatment design and prognosis analysis

The use of conventional ML has been experimented with to perform classification and regression tasks. In a COVID-19 treatment study, Carapito et al. [117] highlighted biologically relevant genes that could be targeted for personalized medical treatment.

Seven machine learning algorithms were employed to calculate informative features using whole genome sequencing and RNA-seq data. The top 600 genes were selected for further analysis using a Bayesian belief network. The findings suggested that these genes were potential drivers in severe SARS-CoV-2 infections, thereby serving as therapeutic targets to improve COVID-19 treatment. A similar combination of conventional ML methods was applied in esophageal squamous cell carcinoma [118], yielding a ranked list of 17 prognostic molecules.

The treatment of cancer can be greatly enhanced with the aid of AI applications. In a notable study, Xiao et al. [119] explored the relationship between alterations in metabolic products and genetic mutations within breast cancer tissue. Through a combined application of LASSO and SVM algorithms, distinct subtypes of breast cancer were identified via these correlations, thus unveiling potential novel therapeutic targets. Additionally, a gene signature study conducted on stage I lung adenocarcinoma patients utilized decision trees to predict clinical outcomes and therapeutic responses [120]. Analyses of cancer stages in patients with lung adenocarcinoma [121] and hepatocellular carcinoma [122] were performed using shallow neural networks. Moreover, in an effort to improve renal cancer treatment, Motzer et al. employed RNA-seq to distinguish different subtypes of the disease [123]. Utilizing a non-negative matrix factorization algorithm and a random forest, they successfully clustered each subtype, unmasking distinct genetic mutations that can be specifically targeted in personalized therapies.

ML algorithms have also been applied to treatment plan development for less common diseases in biomedical genomics. For instance, one study sequenced the gut microbiome of patients with major depressive disorder (MDD) and healthy controls to identify potential treatment targets [124]. RF and the Boruta ML algorithm revealed differences between the biomarkers of MDD patients and controls, indicating potential bacterial targets for treating MDD. The classification of myelodysplastic syndromes and their subtypes is challenging with current disease classification guidelines [125]. Relying solely on morphological features results in low interobserver reproducibility due to the qualitative nature of identification descriptions. To understand genotype–phenotype correlations, 47 genes were screened and processed using Bayesian network analysis and Dirichlet processes. The resulting subtype clusters were analyzed using multivariate logistic regression to discern the effects of each genomic abnormality on the phenotype. This ability to easily classify subtypes could assist clinicians in making prognostic predictions and therapeutic decisions.

6.2. Deep learning for treatment and prognosis

DNNs have been implemented in the treatment phase of several diseases, such as cancer and COVID-19. Genetic alterations in tumors are the key to precision medicine for cancer treatment. Kather et al. [126] proposed a pan-cancer genetic alteration detection workflow. ShuffleNet was compared to other networks, including DenseNet [127], Inception [41] and ResNet, and this lightweight DNN not only demonstrate the best performance but also offered a straightforward solution for practical applications using mobile platforms. In cancer treatment, DNNs were applied in the prediction of neoantigens, which are one of the primary targets of immunotherapies. Sullivan et al. presented epitope discovery in cancer genomes (EDGE) to identify the neoantigens from routine clinical specimens [128]. EDGE contains a multilayer neural network integrating allele-interacting and non-interacting features based on prior knowledge of human leukocyte antigens. Additionally, a novel DNN architecture was proposed for breast cancer prognosis prediction using both gene expression data and clinical records [129].

DNNs were used for drug response classification and repurposing. Identification of therapeutic biomarkers is crucial to personalized anti-cancer drugs. A hybrid deep learning model, DeepResp-Forest, was proposed, where RFs were assembled in a DNN structure to predict drug sensitivity using gene expression and copy number alteration as inputs [130]. Deep-Resp-Forest was tested on 15 drugs with 400 input samples and compared to a standard SVM approach to show the excellent discriminative ability. In another drug-related study, a network-based deep learning methodology, called deepDTnet, was developed by Zeng et al. for drug repurposing [131]. deepDTnet used a PU-matrix completion algorithm in a DNN structure that consists of genotypic and phenotypic data trained on 732 FDA-approved drugs, showing high performance in the drug–target interaction application. deepDTnet is also capable of uncovering chemical structures, semantic relationships, and molecular targets of different types of drugs. Another drug reposition application used similarity network fusion and collective variational autoencoder [132]. Two case studies showed potential drugs for Alzheimer's disease and Juvenile rheumatoid arthritis.

Additionally, DNN-based prediction models have been developed to reveal protein–protein interactions, which contain significant information about biochemical pathways that can guide drug discovery. However, the challenge of binding affinity changes after mutation remains in this area of research. Wang et al. [133] developed a topology-based network tree with a hybrid deep learning algorithm that merges CNNs and gradient-boosting trees to predict PPIs. Another deep learning tool, PrismNet, was created to predict interactions between RNA-binding proteins (RBP) and RNA function [134]. By building convolutional layers that extract features from RBP data, PrismNet can explain the sequential and structural information of the RBP–RNA interaction. Both models have been applied to COVID-19 research. The former focused on the evolution of viral variants to predict the future trend of vaccine discovery [135], and the latter accurately predicts host proteins that bind to the SARS-CoV-2 [136].

6.3. Computer vision for prognosis analysis

COVID-19 is a disease that exhibits a complex and diverse range of clinical symptoms [137]. The severity of infection can vary significantly between individuals, with some showing flu-like symptoms or even being asymptomatic, while others with similar age, sex, and phenotypic characteristics may experience acute respiratory distress syndrome and require intensive care. In a study that recruited 47 critical patients, 25 non-critical patients, and 22 healthy controls, a specific gene, ADAM9, was identified as a driver of COVID-19 severity using DNN together with other machine learning algorithms [117]. SARS-CoV-2 variants can also cause different levels of severity between individuals. Wang et al. [63] integrated CNNs and gradient-boosting trees to understand the impact of viral mutations on protein–protein interactions and infectivity. The hybrid DNN model solved the issue of complex input data in the form of 3D structures.

The prognosis of cancer is studied at the molecular level, where the genome drives the genetic and molecular abnormalities in tumors [138]. Prognosticators of colorectal cancer were investigated using deep CNNs from histology slides and correlated with the gene expression signature of cancer-associated fibroblasts, which is the gold standard for cancer staging [139]. Five popular CNN models (VGG19, AlexNet, SqueezeNet, GoogLeNet, and ResNet50) were initially selected, and VGG19 yielded the best performance with an acceptable training time.

To predict the cancer survival outcome, Chen et al. [140] developed a multimodal fusion strategy that integrates image and genomic features. This supervised learning approach was

validated using glioma and clear cell renal cell carcinoma datasets from the Cancer Genome Atlas. The multimodal fusion strategy is particularly useful in clinical applications as the features extracted from different sources can be interpreted by providing the feature importance shifts. In another CV-based cancer study, the association between histopathological patterns and genomic alterations was analyzed using a deep transfer learning approach [141]. The correlation between images and genomes helped clinicians understand the tumor composition and locate tumor-infiltrating lymphocytes, which can be used as part of immune therapy to improve overall outcomes.

6.4. Transfer learning for treatment and prognosis

Identifying the most effective drugs for individual patients based on their genetic and molecular characteristics is a crucial aspect of cancer treatment. However, crafting accurate drug sensitivity prediction models remains a challenge due to the complex interactions between drugs and cancer cells. Transfer learning has been suggested as a potential solution to enhance drug sensitivity prediction accuracy. For instance, Turki et al. [142] proposed a TL algorithm for drug sensitivity prediction, leveraging the concept of domain adaptation from auxiliary data on a related task. Their results exhibited improved drug sensitivity prediction and statistical significance in several diseases.

Transfer learning has also shown potential in cancer prognosis analysis. Yogananda et al. [143] utilized an image-based deep-learning framework to predict the methylation status of the MGMT promoter in glioma patients. As an important prognostic marker, the MGMT promoter methylation status can guide treatment decisions in glioma patients. By fine-tuning a pre-trained DNN on a small dataset of glioma patients using TL, the authors achieved high accuracy in predicting the MGMT promoter's methylation status. Additionally, Zhang et al. examined the clonal architecture of mesothelioma and its impact on prognosis and tumor microenvironment [110]. The authors used single-cell DNA sequencing to identify clonal heterogeneity in mesothelioma samples and applied TL to classify subpopulations based on their gene expression profiles. They discovered that clonal architecture has prognostic implications, with tumors dominated by a single clone having a better prognosis.

6.5. Graph representation learning for treatment and prognosis

The utilization of GRL techniques in the domain of disease treatment involving graph inputs has shown significant potential. The study conducted by Zong et al. [144] focused on drug–target association tasks within biomedical entity networks encompassing seven distinct entity types, namely disease, drug, target, side effect, variant location, pathway, and haplotype, along with their documented biomedical associations. To accomplish this, the authors employed Node2vec, a classical yet straightforward GRL technique, to generate network embeddings and perform classification and prediction tasks for drug–target associations. Specifically, they evaluated their approach on 75 drug–target associations related to 20 diseases, achieving Area Under Curve scores above 0.9 for multiple drug–target prediction tasks. Apart from proposing a methodological framework for drug–target association analysis, their study also demonstrated the effectiveness of employing different construction schemes by conducting a comparative evaluation of performance across 32 subnetworks. Notably, the Drug–Target–Pathway–Side effect–Variant location network emerged as the most effective in this regard, providing evidence for future research endeavors to determine the optimal formulation of such a network within this task domain.

Survival prediction is non-trivial in disease prognosis analysis and presents a complex task. In an extensive study conducted by Chen et al. [140], GCNs were utilized to extract cell histological features, which were subsequently fused with genomic features. The researchers introduced a methodological framework termed "Pathomic Fusion" to effectively integrate multimodal features obtained from histopathology and genomics, thereby facilitating cancer prognosis prediction. To extract cell graphs from histology images, a K-nearest neighbors algorithm was employed. Subsequently, graph convolutional networks were utilized to learn the corresponding histological features. The original histology image features, cell graph features, and genomic features were then combined using the Kronecker product, along with a gate-based attention mechanism. The efficacy of the proposed methodology was validated through two distinct survival prediction experiments, focusing on glioma and clear cell renal cell carcinoma. These experiments demonstrated the capability of the Pathomic Fusion framework to successfully predict survival outcomes in these specific cancer types. This study demonstrates the utilization of graph convolutional networks and the Pathomic Fusion framework for integrating multimodal features from histopathology and genomics, ultimately enabling accurate survival prediction in cancer prognosis analysis.

6.6. Natural language processing for treatment design

Electronic health records text mining holds considerable potential for enhancing the process of selecting and evaluating clinical treatment decisions. Targeting this trajectory, NLP techniques can serve as a bridge, unveiling latent genetic characteristics embedded within the clinical text, thereby facilitating the investigation of treatment-related objectives. For instance, a recent investigation [145] employed Term Frequency - Inverse Document Frequency and word embedding techniques to extract pertinent textual features related to genetic testing from clinical progress notes of cancer patients. These extracted features were subsequently utilized as inputs for classification models, enabling predictions regarding potential alterations in genomic-related treatment for individual patients. This study elucidated the utility of an automated NLP workflow in evaluating the efficacy of genetic testing in optimizing patient treatment adjustments.

In a similar vein, Zhao et al. [146] employed an NLP-based approach to investigate the correlation between the genetic mutation status of BRCA1/2 and treatment choices in the context of breast cancer. The researchers utilized rule-based techniques and random forest tree methods to extract and classify relevant textual descriptions found in clinical notes that pertained to the specific genetic mutation. By leveraging these extracted textual features, they conducted a classification analysis to characterize the BRCA1/2 genetic mutation status and subsequently examined the association between the prescription of poly-ADP ribose polymerase inhibitors and the identified mutation information. Furthermore, they scrutinized the extent to which the identified mutation information aligned with clinical treatment decisions.

7. Challenges in AI-driven biomedical genomics

7.1. Machine learning algorithms specifically designed for genomic analysis

As previously stated, the progress and potential of AI-assisted genomic applications in tackling clinical problems such as disease prevention, screening, prediction, detection, and prognosis analysis exemplify the promising synergy between AI and genomics. These tools have the potential to significantly enhance healthcare

quality. However, several unresolved challenges must be overcome before these AI-assisted genomic tools can substantially influence everyday clinical practice.

Firstly, the majority of AI-assisted biomedical genomic applications operate in a 'Plug and Play' mode. Conventional machine learning methods and popular computer vision architectures are the most commonly used AI techniques, mainly due to the availability of abundant open-source codes and online toolboxes. Once a research question is proposed, researchers can conveniently input genomic data into these models, generating outputs directly applicable to disease prediction and classification. However, without proper 'configuration,' the performance of these 'Plug and Play' applications can be inconsistent, and their usability may be restricted. Therefore, designing genome-specific machine learning algorithms that address prevailing challenges in current genomic research is paramount for the future of AI-based genomics. These challenges include the disproportionate representation of certain populations and the questionable data quality resulting from sequencing technologies and medical records.

To address these challenges, we can look to the successful applications of CV and NLP in medicine as instructive examples. Medical image segmentation once required manual tracing of hundreds of slices to form a diagnostic conclusion. With the assistance of CV algorithms, it is now possible to achieve higher accuracy than human experts in a matter of minutes. More recently, computer vision applications have explored the multimodal approach, resulting in a further increase in performance by integrating medical images with data from other resources such as physiological signals and clinical notes. Multimodal representations unify data from different modalities into the same vector space for general downstream tasks [147]. Task-specific algorithms have been developed using supervised [148], unsupervised [149], zero-shot [150], and transformer-based learning [151]. By applying similar strategies with the inclusion of genomic information, these AI-driven applications are expected to reach new heights.

For downstream tasks in CV and NLP applications, a vast number of products have been developed by not only large corporations like IBM and Google but also startups such as Viz.ai. For instance, Viz.ai's Vascular Suite, which combines medical images with clinical data and electrocardiograms to search for suspected vascular diseases, has demonstrated a sensitivity of 94.2% and a specificity of 97.3% in real-world clinical studies that collected 1303 CT scans [152]. To replicate this success in genomics, collaborating with clinicians to gather real-world evidence that validates performance is essential to pave the way for acceptance in clinical settings.

Based on the previous review, it has become evident that transfer learning and graph representation learning demonstrate outstanding performance across these tasks. These methodologies, with their inherent capabilities of handling complex data structures and leveraging pre-existing knowledge, have proven superior in managing complex genomics data [153]. This revelation points to an exciting direction for future research and application in genomics and precision medicine. We strongly advocate for the increased use and exploration of these techniques in addressing genomic and biomedical challenges.

7.2. Building knowledge base in biomedical genomics

Although the development of AI algorithms is progressing at a rapid pace, their successful implementation in practical settings heavily depends on the quality of the data used for training. This issue is further complicated in the field of biomedical genomics for two reasons: the availability of vast amounts of highly sensitive data and the limitations of researchers' expertise in handling multifaceted problems.

Today, biomedical genomic analysis goes beyond merely understanding genotypic data. The sources of data are diverse, ranging from medical images to text-based data such as clinical notes, academic publications, regulatory documents, and even ontologies that represent consensus views of knowledge in biology and medicine. The first challenge lies in resolving licensing issues associated with integrating data collected and stored at different organizations, which may include Information and Communication Technologies departments from hospitals, research groups in universities, laboratories funded by research institutes, and private biotechnology companies. Once all parties reach an agreement, the database structure must be designed to accommodate the data processing conventions of AI algorithms. As noted in the above sections, the algorithms developed over the past five years have primarily focused on single data formats due to limited accessibility to comprehensive medical data. By fostering collaborations between research, clinical, and commercial bodies, the integration of high-quality medical data from different sources using multi-modal AI algorithms can enhance system performance. This comprehensive knowledge base can serve as the foundation of a recommender system [154], providing personalized medical service recommendations to alleviate the burden on healthcare systems. Advanced multilayer network embedding algorithms have shown their capabilities in improving the accuracy of recommendations [155]. Such a system can not only streamline the decision-making process but also enhance patient care by offering tailored medical advice.

From a user's perspective, a multisource knowledge base should be designed to be accessible without necessitating extensive domain knowledge. This can be achieved by developing a user-friendly graphical interface, supplemented by a comprehensive manual and easy-to-follow tutorials. Regular maintenance of the interface is also essential as data, software programs, and operating systems are continually updated. Few of the reviewed biomedical genomic applications come with a graphical interface. Given the rapid rate of new genomic discoveries, it is crucial for researchers and biotechnology companies – who develop and maintain these software programs – to collaborate closely. This collaboration ensures the constant updating and enhancement of these tools, keeping them at the forefront of the latest developments in genomic research. Furthermore, rather than relying solely on retrospective experiments, conducting product testing procedures facilitated by biotechnology companies could be an effective approach to ensuring the feasibility of AI-assisted applications.

7.3. AI chatbots, trusting the AI output and ethical consideration

Since November 2022, an AI chatbot powered by Generative Pre-trained Transformers (GPT) has made significant strides globally. Within six months following the release of the public chatbot, ChatGPT, it attracted more than 1 billion users. These users employed it for tasks ranging from code writing and translation to crafting speeches and emails. Recently, the concept of using GPT-4 as a medical AI chatbot garnered substantial attention [156]. Potential applications, such as medical note-taking and consultation, were tested with GPT-4. However, tasks specifically related to genomics remain largely unexplored. One particular task that could be well-suited for a medical chatbot is interpreting genomic findings for patients—a task usually undertaken by genetic counselors. Nonetheless, before deploying chatbots in clinical settings, it is crucial to understand clinicians' perspectives on incorporating these AI-generated results into their practice — in other words, to what extent can clinicians trust the output of AI?

As emphasized in previous review sections, the trustworthiness of developed AI applications remains to be substantiated,

and regulation of AI-assisted biomedical applications is still uncertain. Regulatory bodies such as the U.S. Food and Drug Administration (FDA) and Australia's Therapeutic Goods Administration (TGA) have yet to provide explicit guidelines for using AI in medicine. Although current documentation for software-based medical devices mandates standard random double-blinded clinical trials for approval, this framework may not be suitable for AI-based genomic applications. Hence, it is imperative that clear guidelines for AI applications in clinical genomics are established by regulatory agencies like the FDA and TGA, both for commercial viability and ethical compliance. It is crucial to acknowledge that establishing such standards and regulations should not fall to a single agency alone. Specifically in the field of computer science, AI algorithm developers ought to assume a pivotal role in cultivating a responsible AI environment. By fostering ethical practices, prioritizing transparency, and ensuring fairness as fundamental requirements of AI modeling, developers can significantly contribute to the responsible use and progression of AI in genomics and healthcare.

8. Summary

This review utilized our proposed framework for identifying relevant resources for AI-driven biomedical genomic research, focusing on human genomic analysis using a variety of machine learning algorithms. The AI techniques discussed include conventional machine learning, deep neural networks, convolutional neural networks, transfer learning, computer vision architectures, graph representation learning, and natural language processing. Each has proven its value in genomic analysis. Conventional machine learning offers an intuitive and explainable method to present results. DNNs and CNNs delve into the hierarchical features of genomic data, leading to improved performance at the cost of increased computational power. Transfer learning addresses the challenge of imbalanced data. Meanwhile, advancements in computer vision, natural language processing, and graph representation learning have explored various data formats to address specific challenges.

Next, we highlighted the types of biomedical applications – ranging from disease prevention and early detection to diagnosis, treatment, and prognosis evaluation – discussed in these AI-centric studies. We identified challenges from algorithm development and knowledge base creation to ethics, alongside potential solutions. We aim for this work to spark further collaboration between researchers and industry professionals, leading to the development of more high-performance, practical genomic applications in the biomedical field.

CRediT authorship contribution statement

Kairui Guo: Writing – original draft, Visualization, Validation, Resources, Methodology, Investigation, Formal analysis, Conceptualization. **Mengjia Wu:** Writing – original draft, Methodology, Data curation. **Zelia Soo:** Writing – original draft. **Yue Yang:** Writing – original draft. **Yi Zhang:** Writing – review & editing, Project administration, Conceptualization. **Qian Zhang:** Writing – original draft. **Hua Lin:** Writing – review & editing, Supervision, Project administration, Conceptualization. **Mark Grosser:** Writing – review & editing, Supervision, Project administration, Funding acquisition. **Deon Venter:** Writing – review & editing, Conceptualization. **Guangquan Zhang:** Writing – review & editing, Supervision. **Jie Lu:** Writing – review & editing, Supervision, Project administration, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Disclosure Statement

Given her role as Editor in chief Jie Lu, had no involvement in the peer-review of this article and has no access to information regarding its peer-review. Full responsibility for the editorial process for this article was delegated to Zhong LI.

Data availability

No data was used for the research described in the article.

Acknowledgments

This research is supported by the Australian Research Council Linkage Project: LP210100414.

References

[1] Francis S. Collins, Victor A. McKusick, Implications of the human genome project for medical science, JAMA 285 (5) (2001) 540–544.
[2] Deepti Gurdasani, Inês Barroso, Eleftheria Zeggini, Manjinder S. Sandhu, Genomics of disease risk in globally diverse populations, Nature Rev. Genet. 20 (9) (2019) 520–535.
[3] Zeinab Safarpour Lima, Mostafa Ghadamzadeh, Farzad Tahmasebi Arashloo, Ghazaleh Amjad, Mohammad Reza Ebadi, Ladan Younesi, Recent advances of therapeutic targets based on the molecular signature in breast cancer: Genetic mutations and implications for current treatment paradigms, J. Hematol. Oncol. 12 (2019) 1–25.
[4] Brinton Seashore-Ludlow, Matthew G Rees, Jaime H. Cheah, Murat Cokol, Edmund V. Price, Matthew E Coletti, Victor Jones, Nicole E. Bodycombe, Christian K. Soule, Joshua Gould, Harnessing connectivity in a large-scale small-molecule sensitivity dataset, Cancer Discov. 5 (11) (2015) 1210–1223.
[5] Yasin Mamatjan, Sameer Agnihotri, Anna Goldenberg, Peter Tonge, Sheila Mansouri, Gelareh Zadeh, Kenneth Aldape, Molecular signatures for tumor classification: An analysis of the cancer genome atlas data, J. Molecul. Diagn. 19 (6) (2017) 881–891.
[6] Maxwell W. Libbrecht, William Stafford Noble, Machine learning applications in genetics and genomics, Nature Rev. Genet. 16 (6) (2015) 321–332.
[7] Raquel Dias, Ali Torkamani, Artificial intelligence in clinical and genomic diagnostics, Genome Med. 11 (1) (2019) 1–12.
[8] Ahmad Alimadadi, Sachin Aryal, Ishan Manandhar, Patricia B. Munroe, Bina Joe, Xi Cheng, Artificial intelligence and machine learning to fight COVID-19, 2020, pp. 200–202.
[9] Antonio De Marvao, Timothy J.W. Dawes, Declan P. O'Regan, Artificial intelligence for cardiac imaging-genetics research, Front. Cardiovasc. Med. 6 (2020) 195.
[10] Jia Xu, Pengwei Yang, Shang Xue, Bhuvan Sharma, Marta Sanchez-Martin, Fang Wang, Kirk A. Beaty, Elinor Dehan, Baiju Parikh, Translating cancer genomics into precision medicine with artificial intelligence: Applications, challenges and future perspectives, Hum. Genetics 138 (2) (2019) 109–124.
[11] J. Yan, L. Du, X. Yao, Li Shen, Machine learning in brain imaging genomics, in: Machine Learning and Medical Imaging, Elsevier, 2016, pp. 411–434.
[12] Bruna Gomes, Euan A. Ashley, Artificial intelligence in molecular medicine, N. Engl. J. Med. 388 (26) (2023) 2456–2465.
[13] Li Shen, Paul M. Thompson, Brain imaging genomics: Integrated analysis and machine learning, Proc. IEEE 108 (1) (2019) 125–162.
[14] Matthew E. Falagas, Eleni I. Pitsouni, George A. Malietzis, Georgios Pappas, Comparison of PubMed, scopus, web of science, and google scholar: strengths and weaknesses, FASEB J. 22 (2) (2008) 338–342.
[15] Arezoo Aghaei Chadegani, Hadi Salehi, Melor Md Yunus, Hadi Farhadi, Masood Fooladi, Maryam Farhadi, Nader Ale Ebrahim, A comparison between two main academic literature collections: Web of science and scopus databases, 2013, arXiv preprint arXiv:1305.0377.
[16] Zhihong Shen, Hao Ma, Kuansan Wang, A web-scale system for scientific knowledge exploration, 2018, arXiv preprint arXiv:1805.12216.
[17] Salvador García, Sergio Ramírez-Gallego, Julián Luengo, José Manuel Benítez, Francisco Herrera, Big data preprocessing: Methods and prospects, Big Data Anal. 1 (1) (2016) 1–22.

[18] Ilya Y. Zhbannikov, Samuel S. Hunter, James A. Foster, Matthew L. Settles, SeqyClean: A pipeline for high-throughput sequence data preprocessing, in: Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, 2017, pp. 407–416.
[19] Stephanie C. Hicks, F. William Townes, Mingxiang Teng, Rafael A. Irizarry, Missing data and technical variability in single-cell RNA-sequencing experiments, Biostatistics 19 (4) (2018) 562–578.
[20] Seungbyn Baek, Insuk Lee, Single-cell ATAC sequencing analysis: From data preprocessing to hypothesis generation, Comput. Struct. Biotechnol. J. 18 (2020) 1429–1439.
[21] Amrita Chattopadhyay, Tzu-Pin Lu, Gene-gene interaction: The curse of dimensionality, Ann. Transl. Med. 7 (24) (2019).
[22] Anne-Laure Boulesteix, Korbinian Strimmer, Partial least squares: A versatile tool for the analysis of high-dimensional genomic data, Brief. Bioinform. 8 (1) (2007) 32–44.
[23] Qian Zhang, Jie Lu, Yaochu Jin, Artificial intelligence in recommender systems, Complex Intell. Syst. 7 (2021) 439–457.
[24] Rahul C. Deo, Machine learning in medicine, Circulation 132 (20) (2015) 1920–1930.
[25] Alvin Rajkomar, Jeffrey Dean, Isaac Kohane, Machine learning in medicine, N. Engl. J. Med. 380 (14) (2019) 1347–1358.
[26] Jenni A.M. Sidey-Gibbons, Chris J. Sidey-Gibbons, Machine learning in medicine: A practical introduction, BMC Med. Res. Methodol. 19 (2019) 1–18.
[27] Atharva Sharma, Xiuwen Liu, Xiaojun Yang, Di Shi, A patch-based convolutional neural network for remote sensing image classification, Neural Netw. 95 (2017) 19–28.
[28] Afia Zafar, Muhammad Aamir, Nazri Mohd Nawi, Ali Arshad, Saman Riaz, Abdulrahman Alruban, Ashit Kumar Dutta, Sultan Almotairi, A comparison of pooling methods for convolutional neural networks, Appl. Sci. 12 (17) (2022) 8643.
[29] Hojjat Salehinejad, Sharan Sankar, Joseph Barfett, Errol Colak, Shahrokh Valaee, Recent advances in recurrent neural networks, 2017, arXiv preprint arXiv:1801.01078.
[30] Yong Yu, Xiaosheng Si, Changhua Hu, Jianxun Zhang, A review of recurrent neural networks: LSTM cells and network architectures, Neural Comput. 31 (7) (2019) 1235–1270.
[31] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, Yoshua Bengio, Empirical evaluation of gated recurrent neural networks on sequence modeling, 2014, arXiv preprint arXiv:1412.3555.
[32] Parvathaneni Naga Srinivasu, Jana Shafi, T. Balamurali Krishna, Canavoy Narahari Sujatha, S. Phani Praveen, Muhammad Fazal Ijaz, Using recurrent neural networks for predicting type-2 diabetes from genomic and tabular data, Diagnostics 12 (12) (2022) 3067.
[33] Jie Lu, Vahid Behbood, Peng Hao, Hua Zuo, Shan Xue, Guangquan Zhang, Transfer learning using computational intelligence: A survey, Knowl.-Based Syst. 80 (2015) 14–23.
[34] Sinno Jialin Pan, Qiang Yang, A survey on transfer learning, IEEE Trans. Knowl. Data Eng. 22 (10) (2010) 1345–1359.
[35] Andrew Arnold, Ramesh Nallapati, William W. Cohen, A comparative study of methods for transductive transfer learning, in: Seventh IEEE International Conference on Data Mining Workshops, ICDMW 2007, IEEE, 2007, pp. 77–82.
[36] Zheng Wang, Yangqiu Song, Changshui Zhang, Transferred dimensionality reduction, in: Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2008, Antwerp, Belgium, September 15-19, 2008, Proceedings, Part II 19, Springer, 2008, pp. 550–565.
[37] Zhangchen Zhao, Lars G. Fritsche, Jennifer A. Smith, Bhramar Mukherjee, Seunggeun Lee, The construction of cross-population polygenic risk scores using transfer learning, Am. J. Hum. Genet. 109 (11) (2022) 1998–2008.
[38] Richard Szeliski, Computer Vision: Algorithms and Applications, Springer Nature, 2022.
[39] Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton, Imagenet classification with deep convolutional neural networks, Commun. ACM 60 (6) (2017) 84–90.
[40] Karen Simonyan, Andrew Zisserman, Very deep convolutional networks for large-scale image recognition, 2014, arXiv preprint arXiv:1409.1556.
[41] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, Andrew Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1–9.
[42] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 770–778.
[43] Vijay Badrinarayanan, Alex Kendall, Roberto Cipolla, Segnet: A deep convolutional encoder-decoder architecture for image segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 39 (12) (2017) 2481–2495.
[44] Zuhir Bodalal, Stefano Trebeschi, Thi Dan Linh Nguyen-Kim, Winnie Schats, Regina Beets-Tan, Radiogenomics: Bridging imaging and genomics, Abdom. Radiol. 44 (6) (2019) 1960–1984.
[45] William L. Hamilton, Rex Ying, Jure Leskovec, Representation learning on graphs: Methods and applications, 2017, arXiv preprint arXiv:1709.05584.
[46] Hervé Abdi, Lynne J. Williams, Principal component analysis, Wiley Interdiscipl. Rev.: Comput. Stat. 2 (4) (2010) 433–459.
[47] Aditya Grover, Jure Leskovec, Node2vec: Scalable feature learning for networks, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 855–864.
[48] Mingdong Ou, Peng Cui, Jian Pei, Ziwei Zhang, Wenwu Zhu, Asymmetric transitivity preserving graph embedding, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1105–1114.
[49] Amr Ahmed, Nino Shervashidze, Shravan Narayanamurthy, Vanja Josifovski, Alexander J. Smola, Distributed large-scale natural graph factorization, in: Proceedings of the 22nd International Conference on World Wide Web, 2013, pp. 37–48.
[50] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, S. Yu Philip, A comprehensive survey on graph neural networks, IEEE Trans. Neural Netw. Learn. Syst. 32 (1) (2020) 4–24.
[51] Daokun Zhang, Jie Yin, Xingquan Zhu, Chengqi Zhang, Network representation learning: A survey, IEEE Trans. Big Data 6 (1) (2018) 3–28.
[52] Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, Maosong Sun, Graph neural networks: A review of methods and applications, AI Open 1 (2020) 57–81.
[53] Xiangyu Li, Weizheng Chen, Yang Chen, Xuegong Zhang, Jin Gu, Michael Q. Zhang, Network embedding-based representation learning for single cell RNA-seq data, Nucleic Acids Res. (2017).
[54] Bailiang Li, Yi Cui, Dhanya K. Nambiar, John B. Sunwoo, Ruijiang Li, The immune subtypes and landscape of squamous cell CarcinomaImmune landscape of squamous cell carcinoma, Clin. Cancer Res. 25 (12) (2019) 3528–3537.
[55] Sungmin Rhee, Seokjun Seo, Sun Kim, Hybrid approach of relation network and localized graph convolutional filtering for breast cancer subtype classification, 2017, arXiv preprint arXiv:1711.05859.
[56] Hai-Cheng Yi, Zhu-Hong You, Zhen-Hao Guo, De-Shuang Huang, Keith C.C. Chan, Learning representation of molecules in association network for predicting intermolecular associations, IEEE/ACM Trans. Comput. Biol. Bioinform. 18 (6) (2020) 2546–2554.
[57] Shuting Jin, Xiangxiang Zeng, Feng Xia, Wei Huang, Xiangrong Liu, Application of deep learning methods in biological networks, Brief. Bioinform. 22 (2) (2021) 1902–1917.
[58] Saskia Locke, Anthony Bashall, Sarah Al-Adely, John Moore, Anthony Wilson, Gareth B. Kitchen, Natural language processing in medicine: A review, Trends Anaesthesia Crit. Care 38 (2021) 4–9.
[59] Albert Gatt, Emiel Krahmer, Survey of the state of the art in natural language generation: Core tasks, applications and evaluation, J. Artificial Intelligence Res. 61 (2018) 65–170.
[60] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, Graham Neubig, Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing, ACM Comput. Surv. 55 (9) (2023) 1–35.
[61] Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, Xuanjing Huang, Pre-trained models for natural language processing: A survey, Sci. China Technol. Sci. 63 (10) (2020) 1872–1897.
[62] Ellen Shrock, Eric Fujimura, Tomasz Kula, Richard T. Timms, I-Hsiu Lee, Yumei Leng, Matthew L. Robinson, Brandon M. Sie, Mamie Z. Li, Yuezhou Chen, et al., Viral epitope profiling of COVID-19 patients reveals cross-reactivity and correlates of severity, Science 370 (6520) (2020) eabd4250.
[63] Rui Wang, Jiahui Chen, Kaifu Gao, Guo-Wei Wei, Vaccine-escape and fast-growing mutations in the United Kingdom, the United States, Singapore, Spain, India, and other COVID-19-devastated countries, Genomics 113 (4) (2021) 2158–2170.
[64] Xiaoke Ma, Penggang Sun, Maoguo Gong, An integrative framework of heterogeneous genomic data for cancer dynamic modules based on matrix decomposition, IEEE/ACM Trans. Comput. Biol. Bioinform. 19 (1) (2020) 305–316.
[65] Andrés López-Cortés, Alejandro Cabrera-Andrade, José M. Vázquez-Naya, Alejandro Pazos, Humberto Gonzáles-Díaz, César Paz-y Miño, Santiago Guerrero, Yunierkis Pérez-Castillo, Eduardo Tejera, Cristian R. Munteanu, Prediction of breast cancer proteins involved in immunotherapy, metastasis, and RNA-binding using molecular descriptors and artificial neural networks, Sci. Rep. 10 (1) (2020) 8515.
[66] Shilong Li, Zongxian Zhao, Huaxiang Yang, Daohan Wang, Weilin Sun, Shuliang Li, Zhaoxiong Zhang, Weihua Fu, Construction and validation of a nomogram for the preoperative prediction of lymph node metastasis in gastric cancer, Cancer Control 28 (2021) 10732748211027160.
[67] Min Wu, Deng-Jie Ou-Yang, Bo Wei, Pei Chen, Qi-man Shi, Hai-long Tan, Bo-qiang Huang, Mian Liu, Zi-en Qin, Ning Li, et al., A prognostic model of differentiated thyroid cancer based on up-regulated glycolysis-related genes, Front. Endocrinol. 13 (2022).

[68] Reka Toth, Heiko Schiffmann, Claudia Hube-Magg, Franziska Büscheck, Doris Höflmayer, Sören Weidemann, Patrick Lebok, Christoph Fraune, Sarah Minner, Thorsten Schlomm, et al., Random forest-based modelling to detect biomarkers for prostate cancer progression, Clin. Epigenetics 11 (2019) 1–15.
[69] Sai Zhang, Johnathan Cooper-Knock, Annika K. Weimer, Minyi Shi, Tobias Moll, Jack N.G. Marshall, Calum Harvey, Helia Ghahremani Nezhad, John Franklin, Cleide dos Santos Souza, et al., Genome-wide identification of the genetic basis of amyotrophic lateral sclerosis, Neuron 110 (6) (2022) 992–1008.
[70] Marwa Helmy, Eman Eldaydamony, Nagham Mekky, Mohammed Elmogy, Hassan Soliman, Predicting parkinson disease related genes based on PyFeat and gradient boosted decision tree, Sci. Rep. 12 (1) (2022) 10004.
[71] Soudeh Ghafouri-Fard, Mohammad Taheri, Mir Davood Omrani, Amir Daaee, Hossein Mohammad-Rahimi, Application of artificial neural network for prediction of risk of multiple sclerosis based on single nucleotide polymorphism genotypes, J. Mol. Neurosci. 70 (2020) 1081–1087.
[72] Philip F. Halloran, Jeff Reeve, Katelynn S. Madill-Thomsen, Zachary Demko, Adam Prewett, Paul Billings, Trifecta Investigators, et al., The trifecta study: Comparing plasma levels of donor-derived cell-free DNA with the molecular phenotype of kidney transplant biopsies, J. Am. Soc. Nephrol. 33 (2) (2022) 387–400.
[73] Chenggang Yang, Jing Ren, Bangling Li, Chuandi Jin, Cui Ma, Cheng Cheng, Yaolan Sun, Xiaofeng Shi, Identification of gene biomarkers in patients with postmenopausal osteoporosis, Mol. Med. Rep. 19 (2) (2019) 1065–1073.
[74] Feng Zhu, Lili Zuo, Rui Hu, Jin Wang, Zhihua Yang, Xin Qi, Limin Feng, A ten-genes-based diagnostic signature for atherosclerosis, BMC Cardiovasc. Disorders 21 (1) (2021) 1–8.
[75] Parvathaneni Naga Srinivasu, Jana Shafi, T. Balamurali Krishna, Canavoy Narahari Sujatha, S. Phani Praveen, Muhammad Fazal Ijaz, Using recurrent neural networks for predicting type-2 diabetes from genomic and tabular data, Diagnostics 12 (12) (2022) 3067.
[76] Xiangyu Zhang, Zheng Wang, Wanxiangfu Tang, Xinyu Wang, Rui Liu, Hua Bao, Xin Chen, Yulin Wei, Shuyu Wu, Hairong Bao, et al., Ultrasensitive and affordable assay for early detection of primary liver cancer using plasma cell-free DNA fragmentomics, Hepatology 76 (2) (2022) 317–329.
[77] Chiara Maria Lavinia Loeffler, Nadina Ortiz Bruechle, Max Jung, Lancelot Seillier, Michael Rose, Narmin Ghaffari Laleh, Ruth Knuechel, Titus J. Brinker, Christian Trautwein, Nadine T. Gaisa, Artificial intelligence–based detection of FGFR3 mutational status directly from routine histology in bladder cancer: A possible preselection for molecular testing? Eur. Urol. Focus 8 (2) (2022) 472–479.
[78] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, Jian Sun, Shufflenet: An extremely efficient convolutional neural network for mobile devices, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6848–6856.
[79] Benedikt Atli Jónsson, Gyda Bjornsdottir, T.E. Thorgeirsson, Lotta María Ellingsen, G. Bragi Walters, DF Gudbjartsson, Hreinn Stefansson, Kari Stefansson, M.O. Ulfarsson, Brain age prediction using deep learning uncovers associated sequence variants, Nat. Commun. 10 (1) (2019) 5409.
[80] Shengcheng Dong, Alan P. Boyle, Predicting functional variants in enhancer and promoter elements using regulomedb, Hum. Mutat. 40 (9) (2019) 1292–1298.
[81] Zhong Zhuang, Xiaotong Shen, Wei Pan, A simple convolutional neural network for prediction of enhancer–promoter interactions with DNA sequence data, Bioinformatics 35 (17) (2019) 2899–2906.
[82] Xiang Zhou, Hua Chai, Huiying Zhao, Ching-Hsing Luo, Yuedong Yang, Imputing missing RNA-sequencing data from DNA methylation by using a transfer learning–based neural network, GigaScience 9 (7) (2020) giaa076.
[83] Jaclyn N. Taroni, Peter C. Grayson, Qiwen Hu, Sean Eddy, Matthias Kretzler, Peter A. Merkel, Casey S. Greene, MultiPLIER: A transfer learning framework for transcriptomics reveals systemic features of rare disease, Cell Syst. 8 (5) (2019) 380–394. e4.
[84] Ning Xu, Hui Guo, Xurui Li, Qian Zhao, Jianguo Li, A five-genes based diagnostic signature for sepsis-induced ARDS, Pathol. Oncol. Res. (2021) 102.
[85] Jussi Kupari, Dmitry Usoskin, Marc Parisien, Daohua Lou, Yizhou Hu, Michael Fatt, Peter Lönnerberg, Mats Spångberg, Bengt Eriksson, Nikolaos Barkas, et al., Single cell transcriptomics of primate sensory neurons identifies cell types associated with chronic pain, Nat. Commun. 12 (1) (2021) 1510.
[86] Yixuan Duan, Enrui Xie, Chang Liu, Jingjing Sun, Jie Deng, Establishment of a combined diagnostic model of abdominal aortic aneurysm with random forest and artificial neural network, BioMed Res. Int. 2022 (2022).
[87] Shangjin Lin, Cong Chen, Xiaoxi Cai, Fengjian Yang, Yongqian Fan, Development and verification of a combined diagnostic model for sarcopenia with random forest and artificial neural network, Comput. Math. Methods Med. 2022 (2022).
[88] Fan Peng, Bahaerguli Muhuitijiang, Jiawei Zhou, Haoyu Liang, Yu Zhang, Ranran Zhou, An artificial neural network model to diagnose non-obstructive azoospermia based on RNA-binding protein-related genes, Aging (Albany NY) 15 (8) (2023) 3120.
[89] Juntao Li, Ke Liang, Xuekun Song, Logistic regression with adaptive sparse group lasso penalty and its application in acute leukemia diagnosis, Comput. Biol. Med. 141 (2022) 105154.
[90] Hui Yu, David C. Samuels, Ying-yong Zhao, Yan Guo, Architectures and accuracy of artificial neural network for disease classification from omics data, BMC Genomics 20 (2019) 1–12.
[91] Zohre Arabi Bulaghi, Ahmad Habibizad Navin, Mehdi Hosseinzadeh, Ali Rezaee, World competitive contest-based artificial neural network: A new class-specific method for classification of clinical and biological datasets, Genomics 113 (1) (2021) 541–552.
[92] Dongli Zhao, Zhe Zhang, Zhonghuang Wang, Zhenglin Du, Meng Wu, Tingting Zhang, Jialu Zhou, Wenming Zhao, Yuanguang Meng, Diagnosis and prediction of endometrial carcinoma using machine learning and artificial neural networks based on public databases, Genes 13 (6) (2022) 935.
[93] Aiguo Wang, Ning An, Guilin Chen, Li Liu, Gil Alterovitz, Subtype dependent biomarker identification and tumor classification from gene expression profiles, Knowl.-Based Syst. 146 (2018) 104–117.
[94] Zahid Halim, et al., An ensemble filter-based heuristic approach for cancerous gene expression classification, Knowl.-Based Syst. 234 (2021) 107560.
[95] Giulio Caravagna, Ylenia Giarratano, Daniele Ramazzotti, Ian Tomlinson, Trevor A. Graham, Guido Sanguinetti, Andrea Sottoriva, Detecting repeated cancer evolution from multi-region tumor sequencing data, Nat. Methods 15 (9) (2018) 707–714.
[96] Xiaoxu Guo, Fanghe Lin, Chuanyou Yi, Juan Song, Di Sun, Li Lin, Zhixing Zhong, Zhaorun Wu, Xiaoyu Wang, Yingkun Zhang, Deep transfer learning enables lesion tracing of circulating tumor cells, Nature Commun. 13 (1) (2022) 7687.
[97] Pan Ge, Jian Zhang, Liang Zhou, Mo-qi Lv, Yi-xin Li, Jin Wang, Dang-xia Zhou, CircRNA expression profile and functional analysis in testicular tissue of patients with non-obstructive azoospermia, Reproductive Biol. Endocrinol. 17 (1) (2019) 1–10.
[98] Neal Ravindra, Arijit Sehanobish, Jenna L. Pappalardo, David A. Hafler, David van Dijk, Disease state prediction from single-cell data using graph attention networks, in: Proceedings of the ACM Conference on Health, Inference, and Learning, 2020, pp. 121–130.
[99] Ricardo Ramirez, Yu-Chiao Chiu, Allen Hererra, Milad Mostavi, Joshua Ramirez, Yidong Chen, Yufei Huang, Yu-Fang Jin, Classification of cancer types using graph convolutional neural networks, Front. Phys. 8 (2020) 203.
[100] Fan Wu, Zhi-Liang Wang, Kuan-Yu Wang, Guan-Zhang Li, Rui-Chao Chai, Yu-Qing Liu, Hao-Yu Jiang, You Zhai, Yue-Mei Feng, Zheng Zhao, Classification of diffuse lower-grade glioma based on immunological profiling, Mol. Oncol. 14 (9) (2020) 2081–2095.
[101] Peng Han, Peng Yang, Peilin Zhao, Shuo Shang, Yong Liu, Jiayu Zhou, Xin Gao, Panos Kalnis, GCN-MF: Disease-gene association identification by graph convolutional networks and matrix factorization, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2019, pp. 705–713.
[102] Aditya Rao, Thomas Joseph, Vangala G. Saipradeep, Sujatha Kotte, Naveen Sivadasan, Rajgopal Srinivasan, PRIORI-T: A tool for rare disease gene prioritization using MEDLINE, PLoS One 15 (4) (2020) e0231728.
[103] Aditya Rao, Saipradeep Vg, Thomas Joseph, Sujatha Kotte, Naveen Sivadasan, Rajgopal Srinivasan, Phenotype-driven gene prioritization for rare diseases using graph convolution on heterogeneous networks, BMC Med. Genomics 11 (1) (2018) 1–12.
[104] Vikash Singh, Pietro Lio, Towards probabilistic generative models harnessing graph neural networks for disease-gene prediction, 2019, arXiv preprint arXiv:1907.05628.
[105] Kuo Yang, Ruyu Wang, Guangming Liu, Zixin Shu, Ning Wang, Runshun Zhang, Jian Yu, Jianxin Chen, Xiaodong Li, Xuezhong Zhou, HerGePred: Heterogeneous network embedding representation for disease gene prediction, IEEE J. Biomed. Health Inform. 23 (4) (2018) 1805–1815.
[106] Janet Piñero, Núria Queralt-Rosinach, Alex Bravo, Jordi Deu-Pons, Anna Bauer-Mehren, Martin Baron, Ferran Sanz, Laura I. Furlong, DisGeNET: A discovery platform for the dynamical exploration of human diseases and their genes, Database 2015 (2015).
[107] Noa Rappaport, Michal Twik, Inbar Plaschkes, Ron Nudel, Tsippi Iny Stein, Jacob Levitt, Moran Gershoni, C. Paul Morrey, Marilyn Safran, Doron Lancet, MalaCards: An amalgamated human disease compendium with diverse clinical and genetic annotation and structured search, Nucleic Acids Res. 45 (D1) (2017) D877–D887.
[108] Xiaochan Wang, Yuchong Gong, Jing Yi, Wen Zhang, Predicting gene-disease associations from the heterogeneous network using graph embedding, in: 2019 IEEE International Conference on Bioinformatics and Biomedicine, BIBM, IEEE, 2019, pp. 504–511.

[109] Olivier Bodenreider, The unified medical language system (UMLS): Integrating biomedical terminology, Nucleic Acids Res. 32 (suppl_1) (2004) D267–D270.

[110] Jingqing Zhang, Xiaoyu Zhang, Kai Sun, Xian Yang, Chengliang Dai, Yike Guo, Unsupervised annotation of phenotypic abnormalities via semantic latent representations on electronic health records, in: 2019 IEEE International Conference on Bioinformatics and Biomedicine, BIBM, IEEE, 2019, pp. 598–603.

[111] Zhen Yang, Matthias Dehmer, Olli Yli-Harja, Frank Emmert-Streib, Combining deep learning with token selection for patient phenotyping from electronic health records, Sci. Rep. 10 (1) (2020) 1–18.

[112] Derek Klarin, Julie Lynch, Krishna Aragam, Mark Chaffin, Themistocles L. Assimes, Jie Huang, Kyung Min Lee, Qing Shao, Jennifer E. Huffman, Pradeep Natarajan, Genome-wide association study of peripheral artery disease in the million veteran program, Nat. Med. 25 (8) (2019) 1274–1279.

[113] Brooke N. Wolford, Cristen J. Willer, Ida Surakka, Electronic health records: The next wave of complex disease genetics, Hum. Mol. Genetics 27 (R1) (2018) R14–R21.

[114] Joshua C. Denny, Marylyn D. Ritchie, Melissa A. Basford, Jill M. Pulley, Lisa Bastarache, Kristin Brown-Gentry, Deede Wang, Dan R. Masys, Dan M Roden, Dana C. Crawford, PheWAS: Demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations, Bioinformatics 26 (9) (2010) 1205–1210.

[115] Anurag Verma, Lisa Bang, Jason E. Miller, Yanfei Zhang, Ming Ta Michael Lee, Yu Zhang, Marta Byrska-Bishop, David J. Carey, Marylyn D. Ritchie, Sarah A. Pendergrass, Human-disease phenotype map derived from PheWAS across 38,682 individuals, Am. J. Hum. Genet. 104 (1) (2019) 55–64.

[116] Michelle M Clark, Amber Hildreth, Sergey Batalov, Yan Ding, Shimul Chowdhury, Kelly Watkins, Katarzyna Ellsworth, Brandon Camp, Cyrielle I. Kint, Calum Yacoubian, Diagnosis of genetic diseases in seriously ill children by rapid whole-genome sequencing and automated phenotyping and interpretation, Sci. Transl. Med. 11 (489) (2019) eaat6177.

[117] Raphael Carapito, Richard Li, Julie Helms, Christine Carapito, Sharvari Gujja, Véronique Rolli, Raony Guimaraes, Jose Malagon-Lopez, Perrine Spinnhirny, Alexandre Lederle, Identification of driver genes for critical forms of COVID-19 in a deeply phenotyped young patient cohort, Sci. Transl. Med. 14 (628) (2022) eabj7521.

[118] Meng-Xiang Li, Xiao-Meng Sun, Wei-Gang Cheng, Hao-Jie Ruan, Ke Liu, Pan Chen, Hai-Jun Xu, She-Gan Gao, Xiao-Shan Feng, Yi-Jun Qi, Using a machine learning approach to identify key prognostic molecules for esophageal squamous cell carcinoma, BMC Cancer 21 (1) (2021) 1–11.

[119] Yi Xiao, Ding Ma, Yun-Song Yang, Fan Yang, Jia-Han Ding, Yue Gong, Lin Jiang, Li-Ping Ge, Song-Yang Wu, Qiang Yu, et al., Comprehensive metabolomics expands precision medicine for triple-negative breast cancer, Cell Res. 32 (5) (2022) 477–490.

[120] Run Shi, Xuanwen Bao, Kristian Unger, Jing Sun, Shun Lu, Farkhad Manapov, Xuanbin Wang, Claus Belka, Minglun Li, Identification and validation of hypoxia-derived gene signatures to predict clinical outcomes and therapeutic responses in stage I lung adenocarcinoma patients, Theranostics 11 (10) (2021) 5061.

[121] Na Jiang, Xianrong Xu, Exploring the survival prognosis of lung adenocarcinoma based on the cancer genome atlas database using artificial neural network, Medicine 98 (20) (2019).

[122] Yiqiao Luo, Huaicheng Tan, Ting Yu, Jiangfang Tian, Huashan Shi, A novel artificial neural network prognostic model based on a cancer-associated fibroblast activation score system in hepatocellular carcinoma, Front. Immunol. 13 (2022) 927041.

[123] Robert J. Motzer, Romain Banchereau, Habib Hamidi, Thomas Powles, David McDermott, Michael B. Atkins, Bernard Escudier, Li-Fen Liu, Ning Leng, Alexander R. Abbas, et al., Molecular subsets in renal cancer determine outcome to checkpoint and angiogenesis blockade, Cancer Cell 38 (6) (2020) 803–817.

[124] Wen-tao Lai, Wen-feng Deng, Shu-xian Xu, Jie Zhao, Dan Xu, Yang-hui Liu, Yuan-yuan Guo, Ming-bang Wang, Fu-sheng He, Shu-wei Ye, et al., Shotgun metagenomics reveals both taxonomic and tryptophan pathway differences of gut microbiota in major depressive disorder patients, Psychol. Med. 51 (1) (2021) 90–101.

[125] Matteo Bersanelli, Erica Travaglino, Manja Meggendorfer, Tommaso Matteuzzi, Claudia Sala, Ettore Mosca, Chiara Chiereghin, Noemi Di Nanni, Matteo Gnocchi, Matteo Zampini, et al., Classification and personalized prognostic assessment on the basis of clinical and genomic features in myelodysplastic syndromes, J. Clin. Oncol. 39 (11) (2021) 1223.

[126] Jakob Nikolas Kather, Lara R. Heij, Heike I. Grabsch, Chiara Loeffler, Amelie Echle, Hannah Sophie Muti, Jeremias Krause, Jan M. Niehues, Kai A.J. Sommer, Peter Bankhead, Pan-cancer image-based detection of clinically actionable genetic alterations, Nat. Cancer 1 (8) (2020) 789–799.

[127] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, Kilian Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4700–4708.

[128] Brendan Bulik-Sullivan, Jennifer Busby, Christine D. Palmer, Matthew J. Davis, Tyler Murphy, Andrew Clark, Michele Busby, Fujiko Duke, Aaron Yang, Lauren Young, Deep learning using tumor HLA peptide mass spectrometry datasets improves neoantigen identification, Nature Biotechnol. 37 (1) (2019) 55–63.

[129] Nikhilanand Arya, Sriparna Saha, Multi-modal advanced deep learning architectures for breast cancer survival prediction, Knowl.-Based Syst. 221 (2021) 106965.

[130] Ran Su, Xinyi Liu, Leyi Wei, Quan Zou, Deep-resp-forest: A deep forest model to predict anti-cancer drug response, Methods 166 (2019) 91–102.

[131] Xiangxiang Zeng, Siyi Zhu, Weiqiang Lu, Zehui Liu, Jin Huang, Yadi Zhou, Jiansong Fang, Yin Huang, Huimin Guo, Lang Li, Target identification among known drugs by deep learning from heterogeneous networks, Chem. Sci. 11 (7) (2020) 1775–1797.

[132] Tamer N. Jarada, Jon G. Rokne, Reda Alhajj, SNF–CVAE: Computational method to predict drug–disease interactions using similarity network fusion and collective variational autoencoder, Knowl.-Based Syst. 212 (2021) 106585.

[133] Menglun Wang, Zixuan Cang, Guo-Wei Wei, A topology-based network tree for the prediction of protein–protein binding affinity changes following mutation, Nat. Mach. Intell. 2 (2) (2020) 116–123.

[134] Lei Sun, Kui Xu, Wenze Huang, Yucheng T. Yang, Pan Li, Lei Tang, Tuanlin Xiong, Qiangfeng Cliff Zhang, Predicting dynamic cellular protein–RNA interactions by deep learning using in vivo RNA structures, Cell Res. 31 (5) (2021) 495–516.

[135] Rui Wang, Jiahui Chen, Yuta Hozumi, Changchuan Yin, Guo-Wei Wei, Emerging vaccine-breakthrough SARS-CoV-2 variants, ACS Infect. Dis. 8 (3) (2022) 546–556.

[136] Lei Sun, Pan Li, Xiaohui Ju, Jian Rao, Wenze Huang, Lili Ren, Shaojun Zhang, Tuanlin Xiong, Kui Xu, Xiaolin Zhou, In vivo structural characterization of the SARS-CoV-2 RNA genome identifies host proteins vulnerable to repurposed drugs, Cell 184 (7) (2021) 1865–1883. e20.

[137] Richard Horton, Offline: COVID-19 is not a pandemic, The Lancet 396 (10255) (2020) 874.

[138] Philippe G. Aftimos, Philippe Barthelemy, Ahmad Awada, Molecular biology in medical oncology: Diagnosis, prognosis, and precision medicine, Discov. Med. 17 (92) (2014) 81–91.

[139] Jakob Nikolas Kather, Johannes Krisam, Pornpimol Charoentong, Tom Luedde, Esther Herpel, Cleo-Aron Weis, Timo Gaiser, Alexander Marx, Nektarios A. Valous, Dyke Ferber, Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study, PLoS Med. 16 (1) (2019) e1002730.

[140] Richard J. Chen, Ming Y. Lu, Jingwen Wang, Drew F.K. Williamson, Scott J. Rodig, Neal I. Lindeman, Faisal Mahmood, Pathomic fusion: An integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis, IEEE Trans. Med. Imaging 41 (4) (2020) 757–770.

[141] Yu Fu, Alexander W. Jung, Ramon Viñas Torne, Santiago Gonzalez, Harald Vöhringer, Artem Shmatko, Lucy R. Yates, Mercedes Jimenez-Linan, Luiza Moore, Moritz Gerstung, Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis, Nat. Cancer 1 (8) (2020) 800–810.

[142] Turki Turki, Zhi Wei, Jason T.L. Wang, A transfer learning approach via procrustes analysis and mean shift for cancer drug sensitivity prediction, J. Bioinform. Comput. Biol. 16 (03) (2018) 1840014.

[143] C.G.B. Yogananda, B.R. Shah, S.S. Nalawade, G.K. Murugesan, F.F. Yu, M.C. Pinho, B.C. Wagner, B. Mickey, T.R. Patel, B. Fei, MRI-based deep-learning method for determining glioma MGMT promoter methylation status, Am. J. Neuroradiol. 42 (5) (2021) 845–852.

[144] Nansu Zong, Rachael Sze Nga Wong, Yue Yu, Andrew Wen, Ming Huang, Ning Li, Drug–target prediction utilizing heterogeneous bio-linked network embeddings, Brief. Bioinform. 22 (1) (2021) 568–580.

[145] Meijian Guan, Samuel Cho, Robin Petro, Wei Zhang, Boris Pasche, Umit Topaloglu, Natural language processing and recurrent network models for identifying genomic mutation-associated cancer treatment change from patient progress notes, JAMIA Open 2 (1) (2019) 139–149.

[146] Yiqing Zhao, Saravut J. Weroha, Ellen L Goode, Hongfang Liu, Chen Wang, Generating real-world evidence from unstructured clinical notes to examine clinical utility of genetic tests: Use case in BRCAness, BMC Med. Inform. Decis. Mak. 21 (2021) 1–13.

[147] Chao Zhang, Zichao Yang, Xiaodong He, Li Deng, Multimodal intelligence: Representation learning, information fusion, and applications, IEEE J. Sel. Top. Sign. Proces. 14 (3) (2020) 478–493.

[148] Xiaodong Liu, Pengcheng He, Weizhu Chen, Jianfeng Gao, Multi-task deep neural networks for natural language understanding, 2019, arXiv preprint arXiv:1901.11504.

[149] David Golub, Roberto Martin-Martin, Ahmed El-Kishky, Silvio Savarese, Leveraging pretrained image classifiers for language-based segmentation, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2020, pp. 2010–2019.
[150] Duy-Kien Nguyen, Takayuki Okatani, Multi-task learning of hierarchical vision-language representation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 10492–10501.
[151] Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K. Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, et al., From captions to visual concepts and back, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1473–1482.
[152] David P. Stonko, Jonathan J. Morrison, Caitlin W. Hicks, A review of mature machine learning and artificial intelligence enabled applications in aortic surgery, JVS-Vasc. Insights (2023) 100016.
[153] Jie Lu, Hua Zuo, Guangquan Zhang, Fuzzy multiple-source transfer learning, IEEE Trans. Fuzzy Syst. 28 (12) (2019) 3418–3431.
[154] Jie Lu, Dianshuang Wu, Mingsong Mao, Wei Wang, Guangquan Zhang, Recommender system application developments: A survey, Decis. Support Syst. 74 (2015) 12–32.
[155] Jie Lu, Junyu Xuan, Guangquan Zhang, Xiangfeng Luo, Structural property-aware multilayer network embedding for latent factor analysis, Pattern Recognit. 76 (2018) 228–241.
[156] Peter Lee, Sebastien Bubeck, Joseph Petro, Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine, N. Engl. J. Med. 388 (13) (2023) 1233–1239.