OXFORD

Briefings in Bioinformatics, 2022, **23**, 1–12
https://doi.org/10.1093/bib/bbab523
Review

Artificial intelligence in clinical research of cancers

Dan Shao, Yinfei Dai, Nianfeng Li, Xuqing Cao, Wei Zhao, Li Cheng, Zhuqing Rong, Lan Huang, Yan Wang and Jing Zhao

Corresponding authors: Dr Yan Wang, Key laboratory of Symbol Computation and Knowledge Engineering of Ministry of Education, College of Computer Science and Technology, Jilin University, Changchun 130012, China. E-mail: wy6868@jlu.edu.cn; Dr Jing Zhao, Department of Biomedical Informatics, College of Medicine, The Ohio State University, Columbus, OH 43210, USA. E-mail: jing.zhao2@osumc.edu

Abstract

Several factors, including advances in computational algorithms, the availability of high-performance computing hardware, and the assembly of large community-based databases, have led to the extensive application of Artificial Intelligence (AI) in the biomedical domain for nearly 20 years. AI algorithms have attained expert-level performance in cancer research. However, only a few AI-based applications have been approved for use in the real world. Whether AI will eventually be capable of replacing medical experts has been a hot topic. In this article, we first summarize the cancer research status using AI in the past two decades, including the consensus on the procedure of AI based on an ideal paradigm and current efforts of the expertise and domain knowledge. Next, the available data of AI process in the biomedical domain are surveyed. Then, we review the methods and applications of AI in cancer clinical research categorized by the data types including radiographic imaging, cancer genome, medical records, drug information and biomedical literatures. At last, we discuss challenges in moving AI from theoretical research to real-world cancer research applications and the perspectives toward the future realization of AI participating cancer treatment.

Keywords: drug discovery, clinical research of cancers, deep learning, artificial intelligence

Introduction

Artificial Intelligence (AI) has been applied extensively to tasks in medical specialties [1]. The advent of AI technologies to basic biology, pharmacology and medicine have led to multiple performance breakthroughs and achieved performance comparable to human experts in some areas [2]. In a survey about the effect of AI, AI is expected to have a significant impact on many activities in areas such as health and science, and there is a 50% chance of AI outperforming human being in all tasks in 45 years [3].

Early AI approaches were dominated by traditional-symbol-based and information-based expert systems. Subsequently, the emergence of machine learning (ML) brought revolutionary progress to AI. ML, as a traditional AI technology, provides a plethora of algorithms that can improve the determination or prediction accuracy with abundant, high-quality data [4]. Deep Learning (DL) is a type of algorithms developed using neural network models to solve problems that are challenging to solve with traditional ML. Since the first application in image recognition in 2012 [5], DL has become the *de facto* approach for the analysis of computer vision and has been greatly improved in the accuracy of image recognition ever since [1]. DL approaches are classified based on the architecture, type of layers, updating algorithms of connecting weight, feedback mechanism, etc. The most common DL models are deep neural networks (DNNs) [6], convolutional neural networks (CNNs) [7] and recurrent neural networks (RNNs) [8].

Dan Shao is a professor at the College of Computer Science and Technology in Changchun University. Her research focuses on bioinformatics, data mining and machine learning.
Yinfei Dai is a professor at the College of Computer Science and Technology in Changchun University. Her research focuses on data mining and big data computing.
Nianfeng Li is a professor at the College of Computer Science and Technology in Changchun University. His research focuses on data mining and machine learning.
Xuqing Cao is an associate professor of medicine at the Department of Neurology in the People's Hospital of Ningxia Hui Autonomous Region. Her research focuses on the diagnosis and treatment of cerebrovascular disease and Parkinson's disease.
Wei Zhao is a professor at the Department of Biochemistry and Molecular Biology in Ningxia Medical University. Her research focuses on DNA damage repair and gonadal development.
Li Cheng is a doctor in charge at the Department of Electrical Diagnosis in the Affiliated Hospital of Changchun University of Traditional Chinese Medicine. Her research focuses on medical imaging technology.
Zhuqing Rong is a researcher at the School of Science in Changchun University. Her research focuses on statistical analysis.
Lan Huang is a professor at the College of Computer Science and Technology in Jilin University. Her research focuses on data mining and machine learning.
Yan Wang is a professor at the College of Computer Science and Technology in Jilin University. His research focuses on bioinformatics, data mining and machine learning.
Jing Zhao is a clinical assistant professor in the Department of Biomedical Informatics at the Ohio State University College of Medicine. Her research interests lie in statistical genomics and predictive modeling for disease progression.

Received: August 11, 2021. **Revised:** November 6, 2021. **Accepted:** November 13, 2021

Over the past two decades, AI has regained its popularity and shown promises across all dimensions of cancer research along the explosion of digital information. With the rapid growth of the variety and volume of data sets, a large amount of molecular-level tumor information from cancer patients can be readily acquired. The mission of many public databases is to enable the data sharing across cancer studies in support of precision medicine. For example, The Cancer Genome Atlas (TCGA) [9], a widely used public database for cancer research, characterizes over 20 000 cancer samples spanning 33 cancer types and generates various types of data and resources, including genomic data and digital slide repository. Moreover, the development of high-performance computing hardware, such as graphical processing units (GPUs), has provided the parallel processing power for AI algorithms in numerically intensive computations, e.g. multiple layers of abstraction and millions of computing nodes. As a result, AI excels at handling large volumes and complex data, and identifying characteristic from the data, which the human brain cannot recognize.

Although AI has been rapidly incorporated into oncologic research, the development of AI solutions is still in its infancy. Only a few AI-based applications have been approved for use in practice, e.g. hospitals, pharmaceutical companies, etc. It is still under debate whether AI is capable of replacing medical experts as professionals. Much of the popular discussion of AI focuses on progress so far in AI application for cancer clinical research areas.

Consequently, research of AI applications has accelerated and attained performance comparable to human experts in the biomedical field. Furthermore, AI will equip human experts with more information in decision-making and become an essential component of the medical team. This review provides an overview about the key concepts of AI in cancer clinical research, including clinical research status using AI in past two decades, available data, techniques and current applications. We describe the challenges faced in the translation of AI from theoretical studies to real-world clinical use. We hope to provide future perspectives to help drive meaningful investigations that will ultimately realize actual participation of AI in cancer treatment.

Cancer clinical research status using AI in the past two decades

AI has been used in cancer research for nearly 20 years. Significant advances in cancer research have begun to show promise, and expert-level performance has been reached [1]. There are thousands of papers on cancer clinical oncology by AI models, and some key studies covered in this Review are listed in Supplementary Table S1 available online at http://bib.oxfordjournals.org/. Furthermore, multiple companies and industry research groups have joined in using AI to detect, diagnose and treat cancer. IBM is the first company to make a major push to bring AI to the clinic. In 2014, IBM developed Watson to provide medical AI for cancer research [10]. Likewise, at Microsoft's research labs, a group of computer scientists and researchers are trying to use ML and Natural Language Processing (NLP) to program biology for cancer treatment [11]. As a result, AI is poised to make practice-changing impacts on improving accuracy and speed of diagnosis, assisting treatment options suggestions and recommendations, and leading to better prognosis outcomes (Figure 1).

Consensus on AI process in the cancer clinical research

Although some expect AI can replace human experts in diagnostic imaging, treatment decision making, etc., AI has only played a role in adjuvant medicine. Today, AI needs to address many issues in the development and validation of solutions in the cancer research domain. The potential of AI can primarily be confirmed in carefully designed experiments. Hence, an ideal schema will benefit improving practices of AI solutions.

Before beginning AI solution, it is critical to define and characterize the issues to be addressed, and anticipate whether it can be solved (or is worth solving) by AI [43]. Many clinical data exhibit an unbalanced natural distribution of samples between different classes. The unbalanced feature may create a challenge for classification algorithms that are generally designed for balanced classes. The simplest re-sampling methods are random over-sampling and random under-sampling. The former augments the minority class by duplicating the samples in the minority class, while the latter randomly deletes some samples in the majority class [44]. To get an unbiased assessment of AI model, all the available data are divided into three parts, i.e. training set, validation set and testing set. The proportion common is 60–70, 15–20 and 15–20%. Training dataset is used to train the model; verification dataset is used to adjust parameters and select features; and testing dataset is used to evaluate the performance of the trained AI model.

Not all of these features are helpful for AI model computing, while some noisy or irrelevant features may negatively impact the performance of the classifier. Feature selection can assign values regarding which features are most important for identifying favorable bioactivity. Many methods are available for feature selection—t-test, false discovery rate (FDR), recursive feature elimination (RFE), Z-score, Wilcoxon, etc. Cross-validation is a validation technique used to evaluate an AI model on limited samples. A given dataset is randomly split into K groups, and one group is set for testing, and the remaining $K-1$ groups are used for training. So, the procedure is often called K-fold cross-validation. A loss function or cost function can simply measure the absolute difference between the prediction result and the real value and represent some 'cost' associated with the event.

Choosing the most appropriate model could maximize the chances of success in AI solution. There are three

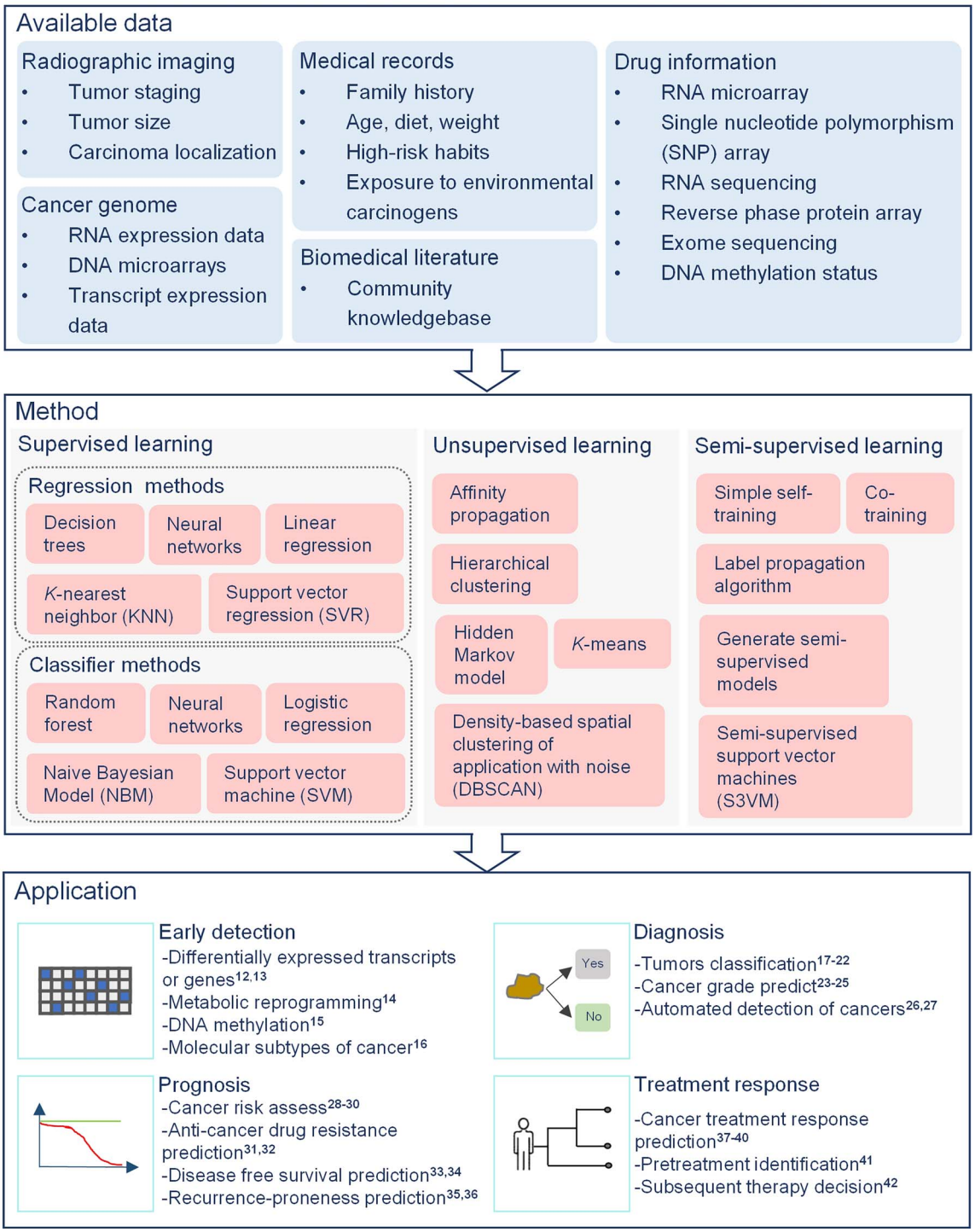

Figure 1. An overview of AI applications used in the clinical research of cancers. Within each data domain, there are still challenges related to the standard of data quality and normalization of AI models.

common types of models that are used in AI: supervised learning, unsupervised learning and semi-supervised learning, as depicted in Figure 1. Supervised learning can predict one or more targets associated with a given label, typical applications including regression and classification. Regression is the task of predicting a continuous real number while classification is the task of predicting a discrete class label. Popular regression algorithms include decision tree, linear regression, *K*-nearest neighbors (KNN), support vector regression (SVR). Meanwhile, popular classification algorithms include random forest, logistic regression, naive Bayes and support vector machines (SVM). There is also some algorithmic overlap, such as neural networks, which can be used for both classification and regression. In contrast, unsupervised learning allows the discovery of latent rules or trends in data, or clustering algorithms, to explore data collections and correlations among samples [45, 46].

Typical unsupervised learning methods include hierarchical clustering, affinity propagation, hidden Markov model, k-means clustering and density-based spatial clustering of application with noise (DBSCAN). Nevertheless, semi-supervised learning combines supervised learning and unsupervised learning, which uses a large amount of unlabeled data, as well as labeled data simultaneously, for pattern recognition. Simple self-training, co-training, label propagation algorithm, generate semi-supervised models and semi-supervised support vector machines (S3VM) are all semi-supervised learning methods. An AI model development is to learn its parameters and then make accurate predictions or determinations on unseen data.

The model's performance is evaluated by various metrics [47]. For a two-class problem, common performance methods include sensitivity, specificity, precision, accuracy, Matthew's correlation coefficient (MCC), *F*-value and the area under the curve (AUC). Sometimes, to assess the generalizability of the performance of AI model, the model is also applied to the independent validation set to assess the classifier's generalizability to new data sets.

Current efforts of the expertise and domain knowledge

Efforts of the expertise

There is an overall scarcity of expert labels available for a generate [2]. While raw data can be fed into the AI models directly, data sets still require manual annotation or at least curation [1]. Multiple subject experts should be involved in data annotation to provide an accurate assessment of data labels. For example, the annotation of medical images requires commitments from clinical experts for extracting region of interests (ROIs) domains in advance. Sometimes, the results of the training model also need qualitatively evaluated by the experts for adjusting the hyperparameters of the model. In particular, rare cases, which are very important in cancer research, need human experts to recognize benefiting from their training and experience. For instance, detecting lung cancer region from hematoxylin and eosin (H&E)-stained pathology images by AI is very difficult due to the complexity of lung cancer tissue structures. It needs human pathologists to circle accurate tumor boundaries and indicate all the tumor spread through air spaces (STASs) [48].

Examples of successful cases of expertise and modeling the cooperation between AI and experts include work by Fan *et al*. [49] and Cha *et al*. [37], in which cancer lesions were manually segmented in consensus by experienced expert radiologists and performances of models were compared to expert radiologists. This will likely be a common role for expertise in the near future.

Domain knowledge integration

Causal factors (e.g. relations between clinical events) often have common determinants at multiple levels. Integrated knowledge can enhance the quality of AI research on individual medical events with cancers, while also paving the way for the interpretability of AI models. Such integration has the dual advantage of generating greater knowledge about complicated mechanisms of cancers, as well as improving our understanding of the influences of disparate causal factors. No existing method dynamically exploits complicated medical knowledge.

Several successful case studies have now been published in which different knowledge domains work together to solve problems from medical practice. Recent advances using digital pathology images have been applied productively to imaging tasks (e.g. grading, prognosis and prediction) across dermatology, ophthalmology, radiology and histopathology. Nguyen *et al*. [50] developed a feasibility study for predicting optimal radiation therapy dose distributions of prostate cancer patients from patient anatomy using deep learning. Mobadersany *et al*. [51] combined histology with genomics to improve prognostic accuracy of molecular subtypes of glioma instead of basing on histology alone. Esteva *et al*. [52] demonstrated classification of skin lesions using CNN on a dataset of 129 450 clinical images.

Structural learning based on bio-networks

Biological networks, such as disease pathway networks, protein–protein interactions (PPIs) networks, and disease similarity networks, provide structured knowledge repositories for discovering the interactions and properties of biological systems [53]. Network approaches have been used in many tasks with remarkable discoveries in biology, including cancer diagnosis, genomic function prediction and drug discovery. Furthermore, these approaches have shown broad utility in uncovering new biology from single-cell to population level [53].

AI on graphs is an important and ubiquitous approach to use graph-structured data as feature information in classification and regression problems. Graph convolutional neural (GCN) network, one of these approaches, combines the graph structure with a neural network to solve biological and chemical problems. Rhee *et al*. [54] proposed a hybrid model that integrated two key components GCN and relation network (RN) to classify breast cancer subtype. Li *et al*. [55] developed a GCN network to predict the survival rate by rendering the optimal graph representations of lung and brain carcinoma whole slide images (WSIs).

Available data of AI process in biomedical domain

To optimize the desired outcomes, it is important to understand what kinds of datasets are needed for a potential utility and how to obtain these datasets. As shown in Figure 1, the sensitive and useful indicators or features for cancer research performed with AI commonly include: (i) radiographic imaging, (ii) cancer

genome, (iii) medical records, (iv) drug information and (v) biomedical literature.

Medical imaging techniques can provide a view of internal activities inside the human body without incisions by doctors, medical practitioners and researchers [56]. Image scanning methods have an important role in clinical research of cancers. magnetic resonance imaging (MRI), computed tomography (CT), ultraviolet, histopathological slides, X-ray and mammography are the most common image scanning methods in radiology. As a result, it is possible that AI could be used to distinguish disease cases from healthy controls by extracting information, such as tumor staging, tumor size, carcinoma localization and so on, from images.

Additionally, genomic analysis has focused on improving the test accuracy by including more disease relevant characteristics [22] and understanding the relationships between genetic makeups and disease states. Gene expression signatures may be used in finding potential biomarkers or therapeutic targets, which can be obtained from the analysis of DNA microarrays, RNA and transcript expression data, and have been demonstrated to be useful in the classification of cancers at the gene expression level [33]. Oligonucleotide chips and cDNA arrays are the two commonly used microarrays [57]. Some popular high-dimensional cancer microarray datasets, such as the small round blue cell tumor (SRBCT) dataset [58] and diffuse large B-cell lymphoma (DLBCL) [59], contain large number of experimental samples with corresponding genes expressions. Genomics can use the genetic changes in the patient's tumor to determine an adequate treatment plan for precision medicine [4]. Moreover, empowered by the advancement of high-throughput bio technologies, the numerous novel cancer biomarkers had been found in body-fluid proteomes [47], such as circular RNAs (circRNAs) that serve as biomarkers for prostate cancer and can be detected in urine [60]. Finding the most discriminative genomic expression across different stages of cancers is a major challenge for cancer research in the last few years.

Meanwhile, medical records, e.g. family history, age, diet, weight (obesity), high-risk habits (smoking, heavy drinking) and exposure to environmental carcinogens (UV radiation, radon or asbestos), may be relevant to the onset and progression of cancer. However, these existing data for training models are limited and unitary. Much ongoing clinical information in surgical pathology reports can be used to determine the eligibility of recruited patients for the study. For data science experts to obtain clinical data is difficult due to lack of opportunity for clinical practice or needing the approval from the institute. Additionally, many first-hand clinical data were hard to incorporate into models because of the difficulty in manual capture. Electronic Health Record (EHR) data would potentially improve the results of biomedical research. EHR systems integrate laboratory results, procedure and radiology reports and clinical narratives, such as primary care and gastroenterology clinic notes.

Recently, several promising results have been demonstrated using AI in drug development, drug-target profiling and drug repurposing/repositioning [32]. Pharmacodynamic, pharmacokinetic and toxicological can improve target specificity and selectivity in small-molecule design of drug. Different data types have been used in cancer-related drug discovery literature based on AI. Classical data types include drug chemical structures, physicochemical properties and molecular targets [61]. Especially, RNA microarray, single nucleotide polymorphism (SNP) array, RNA sequencing (RNA-Seq), reverse phase protein array, exome sequencing and DNA methylation status are available for finding biomarker and generating drug sensitivity predictive models [2]. Existing resource to facilitate the cancer drug discovery include DepMap [62], Genomics of Drug Sensitivity in Cancer (GDSC) [63], canSAR [64], Open Targets [65], TG-GATE [66], drugBank [67] and others. Using these databases and resources, drug sensitivity can be correlated and potential biomarkers of drug response can be provided.

Moreover, the biomedical literature is large and growing rapidly. Several successful applications of AI in various stages of cancer research have been published. To gather and uniformly present the available resources from high-throughput literature retrieval of the cancer clinical research, many centralized, freely accessible and opened community knowledge bases, e.g. TCGA, have been created to provide clinical and molecular data for clinicians and researchers. These knowledge bases integrate heterogeneous data including gene, protein and expression information in control and tumor tissues as well as radiographic imaging information of cancer patients.

AI in cancer clinical research: method and application

Coupled with increasing richness in modern biomedical data, AI and more specifically, DL has garnered some successes in cancer clinical research (Figure 1). AI-based methods are increasingly being used in various fields of cancer clinical research to improve accuracy and efficiency. These include the use of AI in cancer imaging recognition, genomic analysis, medical record mining, drug discovery and biomedical literature utility. We review below the different subareas of the cancer clinical research, which have benefitted from incorporating AI.

Cancer imaging recognition

With the advent of increased computational capabilities and algorithms, AI has been successfully applied in radiology to help the radiologist in defining disease [30]. The raw images, before feeding into model, may need to undergo basic preprocessing. For example, to avoid detecting irrelevant parts of the image, ROIs are extracted by segmenting the lesions on an image, and then only the image information within ROIs are predicted by the

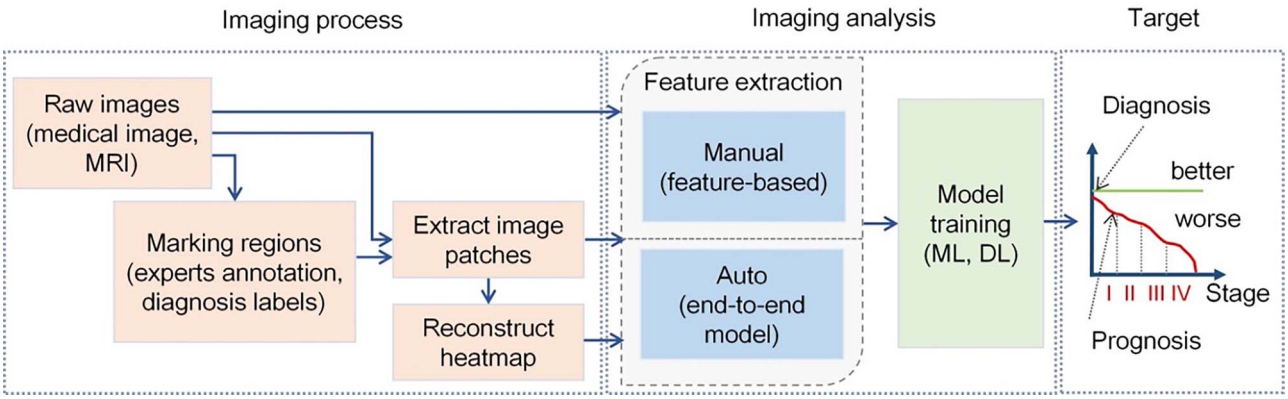

Figure 2. Computational imaging recognition for cancer clinical research. A predictive model for cancer (green box) can be generated using AI approaches on image data. The model uses data from clinical patient raw or preprocessed images. Once validated, the model could be used for predicting diagnosis and/or estimating tumor progression to aid physicians significantly in making decisions about the care and treatment of cancer patients.

model. The regions can be annotated by experts [30] or assigned by diagnosis labels [48]. Nevertheless, unlike other common medical image types, WSIs are too large to be processed by DL model in their entirety [68]. To overcome this defect, WSIs are cropped into numerous small image patches and then fuse these patch level predictions to obtain image level prediction [69]. Sometimes, a tumor probability heatmap can be used to perform geometrical and morphological feature selection, as input of a model, and to identify and characterize disease patterns on digitized tissue slides [70]. Traditional image recognition approaches use handcrafted, user-designed image features, such as texture, shape, color, the density of pixels and contrast/brightness, to capture tumor or cell morphology [37]. These featured-based algorithms suffer from some limitations: (i) these methods depend on the feature-extraction step [71] and (ii) these features are not always consistent under different scanning conditions [1]. Automatic feature extraction, dispensing with the initial feature-extraction step, enables the feeding of raw images into the model directly (end-to-end model) and perform image classification simultaneously (Figure 2) [72].

Many established imaging methods have generated good results in screening and treatment across different cancer types by AI. For example, Trebeschi *et al.* [26] built a CNN classifier using multiparametric MRI (mpMRI) to classify each voxel into tumor or non-tumor. MRI scans of 140 patients with locally advanced rectal cancer were included in their analysis, and two expert radiologists segmented each tumor. AUC of the resulting probability maps was very high, AUC = 0.99. Fan *et al.* [73] utilized a completely unsupervised Convex Analysis of Mixtures (CAM) method to predict breast cancer subtypes by the Decomposition of Contrast-Enhanced MRI (DCE-MRI) from heterogeneous tissues. Cancer recurrence proneness prediction has been identified by Wang *et al.* [36]. They trained a DL network in 8917 CT images from the feature learning cohort to extract the prognostic biomarkers of High-Grade Serous Ovarian Cancer (HGSOC). Afterward, a DL-Cox Proportional Hazard (Cox-PH) model was developed to predict the individual recurrence risk and 3-year recurrence probability of patients. Another valuable application of AI is the prediction of cancer outcomes, e.g. survivability, life expectancy, progression and tumor-drug sensitivity. In many cases, the availability of mammography has been confirmed as the main imaging test method used to screen breast cancer by many computational methods [74]. Li *et al.* [75] developed an improved DL approach for detection of thyroid papillary cancer in ultrasound images. In a study by Vang *et al.* [70], histopathological slides, e.g. H&E stained images, have also been used to classify multiclass breast cancer. Fan *et al.* [76] developed a 3D-mask region-based CNN (3D-Mask RCNN) Computer-Aided Diagnosis (CAD) system for breast based mass detection and segmentation in digital breast tomosynthesis (DBT). More recently, WSIs have been used for pathologic analysis, such as inferring molecular subtype, tumor grade or estrogen receptor status [77].

Genomic analysis

Fundamentally, genomics are systematic approaches to characterize the function of every genomic element of an organism [78]. Genome-wide association studies (GWAS) have successfully identified interacting genetic variants contributing to cancer risk [29]. Additionally, molecular profiling is essential for the identification of predictive biomarkers associated with cancer phenotypes, prognosis and clinical outcomes [79]. Single-cell resolution enables quantitative measurements of the cell types and molecular activity within a tumor [80]. Two computational analysis commonly used in cancer research based on genomics include: (i) gene selection (Figure 3A) and (ii) cancer classification (Figure 3B). Gene selection tries to select high-regulated or differential expression gene and remove poor ones from thousands of genes in microarray experiments by analyzing and measuring their effects upon constructing a classifier [81].

Recently, there has been much progress on AI in cancer research using various types of genomic data as input

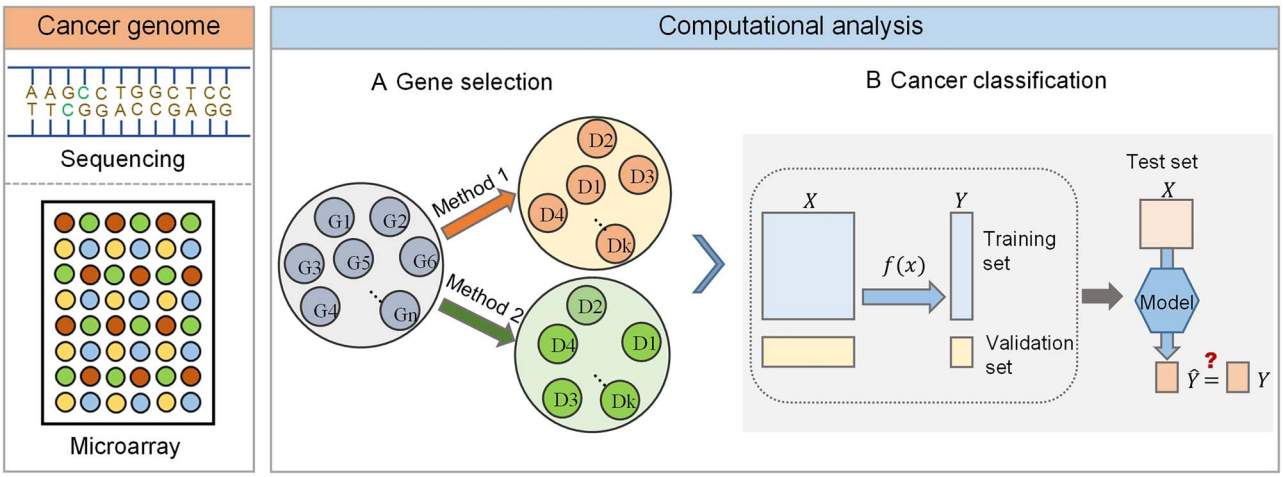

Figure 3. Computational analysis commonly used in cancer research on genomics. AI framework can select high-regulated or differentially expressed genes (A) in several basic cancer classification (B) tasks (such as early disease diagnostics, drug discovery or classification of tumors) using sequencing or microarray profiling data.

data in models. Morais-Rodrigues *et al.* [82] developed a modified logistic regression method to analyze the microarray gene expression for breast cancer progression. In addition, Maros *et al.* [83] developed machine-learning workflows to estimate class probabilities for cancer diagnostics on DNA methylation microarray data. More recently, Albaradei *et al.* [84] presented a DL-based model to differentiated pan-cancer metastasis status based on three heterogeneous data layers from TCGA, including RNA-Seq, microRNA-Seq and DNA methylation data. The model used convolutional variational autoencoder for feature extraction and DNN for classification. The results showed that integrating data can improve the performance compared with using mRNA data only. In other studies, AI models have focused on cancer grade prediction. In a study by Yamamoto *et al.* [17], a SVM classifier was trained on the morphometric classification of microenvironmental myoepithelial cells to quantitatively diagnose breast tumors. They quantitatively measured 11 661 nuclei on four histological types: normal cases, usual ductal hyperplasia and low/high-grade ductal carcinoma in situ (DCIS). At least three pathologists diagnosed and scored all cases independently, and this model was able to classify the four histological types with 90.9% accuracy. Notably, disease-related biomarkers can be identified from genomic data. For instance, Zeng *et al.* [85] used deep forests combined with positive-unlabeled learning methods to predict potential disease-related circRNAs. Radhakrishnan *et al.* [20] combined fluorescence imaging and deep learning to detect subtle changes in nuclear morphometrics at single-cell resolution and opened new avenues for early disease diagnostics and drug discovery.

Electronic medical record mining

Some AI-based models utilize integrated medical record data, including genomic information, unstructured health record and family history to improve the performance of cancer prediction [86]. As an example, to predict survival outcomes of lung cancer, a dataset of observed cancer-associated characteristics of individuals such as lung cancer pathology images, age, gender, smoking status and stage, should be considered. Furthermore, tumor shape, including area, perimeter, convex area, filled area, major axis length and minor axis length, have a role in causing the outcome [48]. For example, Tseng *et al.* [35] applied machine learning, including SVM, C5.0 and extreme learning machine (ELM) to predict the recurrence-proneness for cervical cancer based on the medical records and pathology. They found four most important recurrence-proneness factors that are pathologic stage, pathologic T, cell type and RT target summary.

However, a major challenge is to extract the potential input data from electronic medical records. NLP systems can capture much of the information for the cancer research project. For instance, a preprocessor integrated with an existing NLP system (MedLEE) was done in conjunction with an ongoing clinical research project that assesses disparities and risks of developing breast cancer for minority women [87]. NLP algorithm is used to identify primary and recurrent cancers by identifying and extracting information from electronic pathology reports [88]. In addition, NLP can improve the identification of cancer testing in the electronic medical record [89].

Drug discovery

The development of a new drug is a very complex, expensive and time-consuming process, while AI combined with new experimental technologies is expected to improve this process. AI may be applied for the prediction of clinical efficiency of certain drugs and treatment responses for individual patients (Figure 4). Much work has been done to apply AI to screen drug candidates by identifying a similar chemical structure computationally from large compound libraries [2]. In addition, the

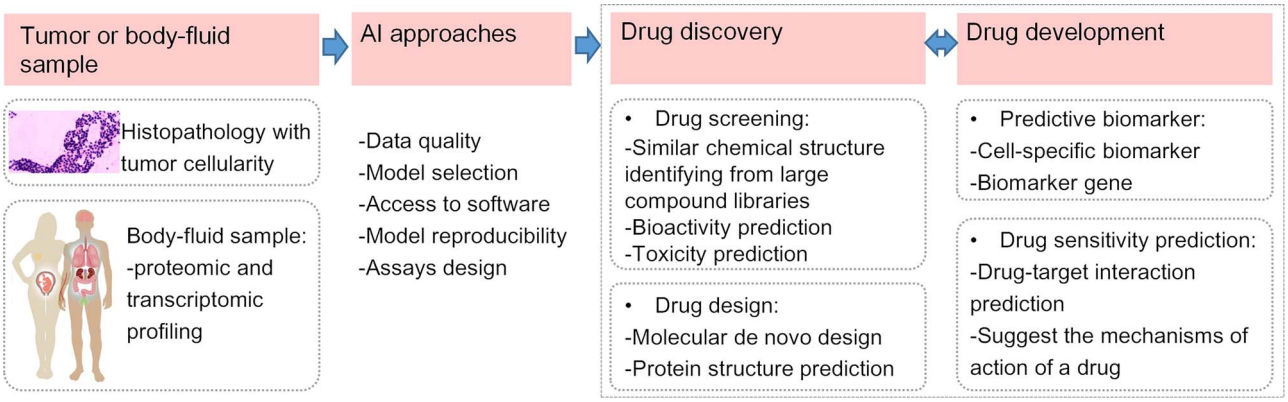

Figure 4. Utilizing AI to support drug discovery and development. Drug research models can be generated using AI approaches on tumor or body-fluid sample data. The model could be used for drug screening and/or design, as well as biomarker discovery and drug sensitivity prediction.

prediction of physical properties, such as bioactivity and toxicity, greatly improves the bioavailability of a candidate molecule [90]. For drug design, molecular de novo design is a valuable application [2]. Meanwhile, the 3D protein structure is extremely important for drug design because candidate molecules are generally designed according to the 3D chemical environment of a target protein [91]. Furthermore, to better understand the mechanism of action of a drug and help improve clinical success rates, some pharmaceutical companies have teamed up with IT companies to develop a platform for biomarker discovery and drug sensitivity prediction [92]. Indeed, molecular profiling data from tissue slices or body fluid have aimed at identifying genomic biomarkers predictive of anticancer drug response [32].

Cancer drug research can benefit from AI due to the availability of a large amount of public databases and resources, and have become more accurate and sophisticated. Choi *et al.* [31] developed a novel deep neural network model for improved prediction of drug resistance and identification of biomarkers related to drug response. Huang *et al.* [93] predicted the responses of 175 individual cancer patients to a variety of standard-of-care chemotherapeutic drugs from the gene-expression profiles (RNA-seq or microarray) of individual patient tumors. Borisov *et al.* [94] predicted the clinical efficiency of anti-cancer drugs for individual patients by transferring features obtained from the expression-based data from cell lines. Chang *et al.* [32] reported Cancer Drug Response profile scan (CDRscan) to predict anticancer drug responsiveness based on large-scale drug screening assay data, including genomic profiles of 787 human cancer cell lines and structural profiles of 244 drugs. Moreover, predicting and interpreting cancer drug response in single cell data based on computational biology approaches shows a clear significance. Yanagisawa *et al.* [95] constructed a CNN model to predict the efficiency of antitumor drugs at the single-cell level.

Many computational tools have been proposed in cancer-related drug discovery based on different AI methodologies. Examples of the applications include DeepChem [96], DeepTox [97], gene2drug [98], STITCH [99], AlphaFold [100] and/or DeepNeuralNetQSAR [101]. The DeepTox algorithm, for instance, based on ML computationally predict 12 000 environmental chemicals and drugs for 12 different toxic effects in specifically designed assays [97]. Otherwise, AlphaFold relied on DNNs is used to predict the 3D structure of a drug target protein [100]. The creation of these tools has helped reduce the cost of drug discovery.

Biomedical literature utility

In the past decades, with great effort by a few large consortiums, several community-based knowledge bases have been developed based on a large collection of published literature in the cancer clinical research field. For example, National Lung Screening Trial (NLST) [101] is a unified data sharing platform that allows users to search, browse, download and analyze tumor regions of lung adenocarcinoma (ADC) patients. The National Cancer Institute (NCI) Genomic Data Commons (GDC) [102] serves as a single knowledge base that unifies genomic and clinical data from different research programs for the cancer research community. In a typical study, a deep CNN model takes a systematic study of the detected tumor regions of lung cancer patients from NLST cohort inputs and is trained to automatically recognize tumor regions for lung cancer, whereas the model developed from the NLST cohort is independently validated in the TCGA cohort for prognostic performance [48].

In addition, some research teams have developed a growing number of databases to search for comprehensive information. For instance, the Sheikh Khalifa Bin Zayed Al Nahyan Institute for Personalized Cancer Therapy (IPCT) at MD Anderson Cancer Center has developed a knowledge base, which provides information on the functions of common genomic alterations and their therapeutic implications to guide personalized treatments in oncology [78]. The literature is manually reviewed by a precision oncology decision support (PODS) team that includes oncologists, geneticists, molecular biologists, computational scientists, computer programmers and bioinformaticians [103]. Moreover,

an integrated Precision Medicine Knowledgebase (PreMedKB) has been developed by seamlessly interpreting the four fundamental components of precision medicine: diseases, genes, variants and drugs [104]. Furthermore, CIVic is an expert-crowd sourced knowledgebase for clinical interpretation of variants in cancer [105]. Recently, body fluid proteome has been intensively studied as a primary source for cancer biomarker. For this reason, our research group have developed a new database of human body fluid proteome (HBFP) that archived 11,827 unique proteins reported by 164 scientific publications since 2001 [106]. By providing a wealth of information, these knowledgebases can be excellent resources and tools for the research community.

Challenges and future directions

AI has demonstrated comparable performance to that of an expert in common application fields across a range of biomedicine. However, although some AI solutions are already available, there are still many challenges for AI to move from theoretical studies to real-world applications.

Currently, one of the biggest challenges facing AI, in general, is data hungry. The acquisition of sufficient large, public, well-annotated cancer dataset is an ongoing need for AI. Although the inclusion of images, genomic data and clinical outcomes in some opened databases had a significant impact on enhancing computational clinical research. The scale, quality and diversity of the data types, such as patient history from prior reports, are potentially relevant to the risk and progression of cancer, but are time-consuming to collect. Data sharing agreement can play an important role in addressing the challenge above. Sharing of large datasets with the community can be enabled by cloud computing and advanced development of the next generation of predictive cancer models.

Additionally, the successful development of an AI model is dependent on the high-quality data. Notwithstanding the amount of available data is growing in volume and variety, the assessment of the quality of data is not standardized.

Moreover, some clinical tasks, such as prognosis prediction, are more unstructured than traditional deep learning tasks [107]. Sometimes, we have to give accurate predictions (e.g. survival times) from a combination of clinicopathologic, genomic markers and images that are much higher resolution. Furthermore, patients span a wide variety of cancer types, and are often missing some form of clinical, imaging or genomic data, making it difficult to apply AI.

Despite AI regularly achieving high performance in medical research, the adoption of AI in real cases is limited due to the somewhat opacity of the model. The machine could not explain how it knew and why it got this result. This is often referred to as the 'black box' problem [108]. It is difficult to present which features of the input data contribute to the output. For example, AI can predict the optimal treatment for a patient but not provide the reasoning it used to make that prediction. Interpretable DL is a trend in alleviating this limitation [109]. In addition, the knowledge gap between clinical and data science experts still presents significant challenges. Physicians have much experience with oncologic workup and management versus data scientists have high-level cognition in data science for understanding AI mechanisms. Further collaboration should be pursued between clinical and data science experts to bridge the gap between them.

Another important issue for AI is its role. It is almost impossible to run an AI without experts. AI should not be seen as a standalone solution in a completely unsupervised environment. On the contrary, it is a helpful assistant to experts, as well as a tool that can help in areas where human capabilities remain limited.

In the future, we believe that AI will participate in cancer treatment clinically and will be deployed to expedite diagnosis, treatment, and even a cure. Moreover, we expect the AI technology will be more widely available and applied to boost survival rates, improve treatment responses and reduce side effects.

Authors' contributions

D.S., Y.W. and J.Z. conceived, designed, and supervised the project. D.S., W.Z., X.C., C.L., L.H., N.L., Y.D. and Z.R. prepared the manuscript. All the authors read and improved the manuscript.

Key Points

- We analyzed cancer clinical research status using AI in the past two decades, which included the ideal schema on AI process in biomedical domain and the current efforts of the expertise and domain knowledge.
- We reviewed the successes in cancer clinical research using AI, focusing on available data, method and application.
- AI algorithms have attained expert level performance in tumors and other malignancies research. The challenge of AI from theoretical studies to real-world clinical use was discussed.

Supplementary data

Supplementary data are available online at https://academic.oup.com/bib.

Acknowledgements

The authors thank Juan Cui (Department of Computer Science and Engineering, University of Nebraska-Lincoln) for helpful comments.

Funding

National Natural Science Foundation of China (62072212), the Development Project of Jilin Province of China (20200401083GX, 2020C003, 2020LY500L06, 20200403172SF), and Guangdong Key Project for Applied Fundamental Research (2018KZDXM076). This work was also supported by Jilin Province Key Laboratory of Big Data Intelligent Computing (20180622002JC).

References

1. Levine AB, Schlosser C, Grewal J, *et al.* Rise of the machines: advances in deep learning for cancer diagnosis. *Trends Cancer* 2019;**5**:157–69.
2. Vamathevan J, Clark D, Czodrowski P, *et al.* Applications of machine learning in drug discovery and development. *Nat Rev Drug Discov* 2019;**18**:463–77.
3. Grace K, Salvatier J, Dafoe A, *et al.* When will AI exceed human performance? Evidence from AI experts. *J Artif Intell Res* 2018;**62**:729–54.
4. Tan AC, Gilbert D. Ensemble machine learning on gene expression data for cancer classification. *Appl Bioinformatics* 2003;**2**(3 Suppl):S75–83.
5. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. *Adv Neural Inform Process Syst* 2012;**25**:1097–105.
6. Larochelle H, Bengio Y, Louradour J, *et al.* Exploring strategies for training deep neural networks. *J Mach Learn Res* 2009;**10**: 1–40.
7. Gu J, Wang Z, Kuen J, *et al.* Recent advances in convolutional neural networks. *Pattern Recognition* 2018;**77**:354–77.
8. Medsker LR, Jain LC. *Recurrent neural networks: design and applications*. Los Angeles: CRC Press, 1999.
9. Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* 2008;**455**:1061–8.
10. Strickland E. IBM Watson, heal thyself: how IBM overpromised and underdelivered on AI health care. *IEEE Spectrum* 2019;**56**: 24–31.
11. Linn A. How Microsoft computer scientists and researchers are working to 'solve' cancer [Internet]. News.microsoft.com 2028. Available from: https://news.microsoft.com/stories/computingcancer/ (August 25 2018, date last accessed).
12. Singireddy S, Alkhateeb A, Rezaeian I, *et al.* Identifying differentially expressed transcripts associated with prostate cancer progression using RNA-Seq and machine learning techniques. *2015 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, IEEE 2015;1–5.
13. Wang D, Li JR, Zhang YH, *et al.* Identification of differentially expressed genes between original breast cancer and xenograft using machine learning algorithms. *Genes* 2018;**9**:155.
14. Zhang Y, Zhang XF, Lane AN, *et al.* TFmeta: a machine learning approach to uncover transcription factors governing metabolic reprogramming. *In Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics* ACM 2018; 351–9.
15. Koelsche C, Schrimpf D, Stichel D, *et al.* Sarcoma classification by DNA methylation profiling. *Nat Commun* 2021;**12**:498.
16. Ming F, He T, Peng Z, *et al.* Diffusion-weighted imaging features of breast tumours and the surrounding stroma reflect intrinsic heterogeneous characteristics of molecular subtypes in breast cancer. *NMR Biomed* 2018;**31**:e0189302.
17. Yamamoto Y, Saito A, Tateishi A, *et al.* Quantitative diagnosis of breast tumors by morphometric classification of microenvironmental myoepithelial cells using a machine learning approach. *Sci Rep* 2017;**7**:46732.
18. Ko J, Bhagwat N, Yee SS, *et al.* Combining machine learning and nanofluidic technology to diagnose pancreatic cancer using exosomes. *ACS Nano* 2017;**11**:11182–93.
19. Yuan Y, Shi Y, Li C, *et al.* DeepGene: an advanced cancer type classifier based on deep learning and somatic point mutations. *BMC Bioinformatics* 2016;**17**:243–56.
20. Radhakrishnan A, Damodaran K, Soylemezoglu AC, *et al.* Machine learning for nuclear mechano-morphometric biomarkers in cancer diagnosis. *Sci Rep* 2017;**7**:17946.
21. Hollon TC, Pandian B, Adapa AR, *et al.* Near real-time intraoperative brain tumor diagnosis using stimulated Raman histology and deep neural networks. *Nat Med* 2020;**26**:52–8.
22. Guillen P, Ebalunode J. Cancer classification based on microarray gene expression data using deep learning. *International Conference on Computational Science and Computational Intelligence (CSCI)*, IEEE 2016; 1403–5.
23. Couture HD, Williams LA, Geradts J, *et al.* Image analysis with deep learning to predict breast cancer grade, ER status, histologic subtype, and intrinsic subtype. *NPJ Breast Cancer* 2018;**4**:30.
24. Fan M, Yuan W, Zhao W, *et al.* Joint prediction of breast cancer histological grade and Ki-67 expression level based on DCE-MRI and DWI radiomics. *IEEE J Biomed Health Inform* 2019;**24**: 1632–42.
25. Fan M, Liu ZH, Xie SD, *et al.* Integration of dynamic contrast-enhanced magnetic resonance imaging and T2-weighted imaging radiomic features by a canonical correlation analysis-based feature fusion method to predict histological grade in ductal breast carcinoma. *Phys Med Biol* 2019;**64**:215001.
26. Trebeschi S, van Griethuysen JJM, Lambregts DMJ, *et al.* Deep learning for fully-automated localization and segmentation of rectal cancer on multiparametric MR. *Sci Rep* 2017;**7**:5301.
27. Schwyzer M, Ferraro DA, Muehlematter UJ, *et al.* Automated detection of lung cancer at ultralow dose PET/CT by deep neural networks–initial results. *Lung Cancer* 2018;**126**:170–3.
28. Li H, Giger ML, Huynh BQ, *et al.* Deep learning in breast cancer risk assessment: evaluation of convolutional neural networks on a clinical dataset of full-field digital mammograms. *J Med Imaging* 2017;**4**:041304.
29. Behravan H, Hartikainen JM, Tengström M, *et al.* Machine learning identifies interacting genetic variants contributing to breast cancer risk: a case study in Finnish cases and controls. *Sci Rep* 2018;**8**:13149.
30. Varghese B, Chen F, Hwang D, *et al.* Objective risk stratification of prostate cancer using machine learning and radiomics applied to multiparametric magnetic resonance images. *Sci Rep* 2019;**9**:1570.
31. Choi J, Park S, Ahn J. RefDNN: a reference drug based neural network for more accurate prediction of anticancer drug resistance. *Sci Rep* 2020;**10**:1861.
32. Chang Y, Park H, Yang HJ, *et al.* Cancer drug response profile scan (CDRscan): a deep learning model that predicts drug effectiveness from cancer genomic signature. *Sci Rep* 2018;**8**: 8857.

33. Ing N, Huang F, Conley A, *et al.* A novel machine learning approach reveals latent vascular phenotypes predictive of renal cancer outcome. *Sci Rep* 2017;**7**:13190.
34. Fan M, Cheng H, Zhang P, *et al.* DCE-MRI texture analysis with tumor subregion partitioning for predicting Ki-67 status of estrogen receptor-positive breast cancers. *J Magn Reson Imaging* 2018;**48**:237–47.
35. Tseng CJ, Lu CJ, Chang CC, *et al.* Application of machine learning to predict the recurrence-proneness for cervical cancer. *Neural Comput Applic* 2014;**24**:1311–6.
36. Wang S, Liu Z, Rong Y, *et al.* Deep learning provides a new computed tomography-based prognostic biomarker for recurrence prediction in high-grade serous ovarian cancer. *Radiother Oncol* 2019;**132**:171–7.
37. Cha KH, Hadjiiski L, Chan HP, *et al.* Bladder cancer treatment response assessment in CT using radiomics with deep-learning. *Sci Rep* 2017;**7**:8738.
38. Xu Y, Hosny A, Zeleznik R, *et al.* Deep learning predicts lung cancer treatment response from serial medical imaging. *Clin Cancer Res* 2019;**25**:3266–75.
39. Bibault JE, Giraud P, Housset M, *et al.* Deep learning and radiomics predict complete response after neo-adjuvant chemoradiation for locally advanced rectal cancer. *Sci Rep* 2018;**8**:12611.
40. Fan M, Chen H, You C, *et al.* Radiomics of tumor heterogeneity in longitudinal dynamic contrast-enhanced magnetic resonance imaging for predicting response to neoadjuvant chemotherapy in breast cancer. *Front Mol Biosci* 2021;**8**:622219.
41. Kann BH, Aneja S, Loganadane GV, *et al.* Pretreatment identification of head and neck cancer nodal metastasis and extranodal extension using deep learning neural networks. *Sci Rep* 2018;**8**:14036.
42. Nagpal K, Foote D, Liu Y, *et al.* Development and validation of a deep learning algorithm for improving Gleason scoring of prostate cancer. *NPJ Digit Med* 2019;**2**:48.
43. Matheny M, Israni T, Ahmed M, *et al. Artificial Intelligence in Health Care: The Hope, The Hype, The Promise, The Peril*. Washington, DC: National Academy of Medicine Press, 2019.
44. Lao J, Chen Y, Li ZC, *et al.* A deep learning-based radiomics model for prediction of survival in glioblastoma multiforme. *Sci Rep* 2017;**7**:10353.
45. Fan M, Xia P, Liu B, *et al.* Tumour heterogeneity revealed by unsupervised decomposition of dynamic contrast-enhanced magnetic resonance imaging is associated with underlying gene expression patterns and poor survival in breast cancer patients. *Breast Cancer Res* 2019;**21**:112.
46. Wang JY, Wang XL, Gao X. Non-negative matrix factorization by maximizing correntropy for cancer clustering. *BMC Bioinformatics* 2013;**14**:107.
47. Huang L, Shao D, Wang Y, *et al.* Human body-fluid proteome: quantitative profiling and computational prediction. *Brief Bioinform* 2021;**22**:315–33.
48. Wang S, Chen A, Yang L, *et al.* Comprehensive analysis of lung cancer pathology images to discover tumor shape and boundary features that predict survival outcome. *Sci Rep* 2018;**8**:10393.
49. Fan M, Liu ZH, Xu MS, *et al.* Generative adversarial network-based super-resolution of diffusion-weighted imaging: application to tumour radiomics in breast cancer. *NMR Biomed* 2020;**33**:e4345.
50. Nguyen D, Long T, Jia X, *et al.* A feasibility study for predicting optimal radiation therapy dose distributions of prostate cancer patients from patient anatomy using deep learning. *Sci Rep* 2019;**9**:1076.
51. Mobadersany P, Yousefi S, Amgad M, *et al.* Predicting cancer outcomes from histology and genomics using convolutional networks. *Proc Natl Acad Sci USA* 2018;**15**:E2970–9.
52. Esteva A, Kuprel B, Novoa RA, *et al.* Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017;**542**:115–8.
53. Camacho DM, Collins KM, Powers RK, *et al.* Next-generation machine learning for biological networks. *Cell* 2018;**173**:1581–92.
54. Rhee S, Seo S, Kim S. Hybrid approach of relation network and localized graph convolutional filtering for breast cancer subtype classification. 2017; *arXiv*:1711.05859.
55. Li R, Yao J, Zhu X, *et al.* Graph CNN for survival analysis on whole slide pathological images. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, Cham 2018; 11071:174–82.
56. Tandel GS, Biswas M, Kakde OG, *et al.* A review on a deep learning perspective in brain cancer classification. *Cancers (Basel)* 2019;**11**:111.
57. Dashtban M, Balafar M. Gene selection for microarray cancer classification using a new evolutionary method employing artificial intelligence concepts. *Genomics* 2017;**109**:91–107.
58. Khan J, Wei JS, Ringnér M, *et al.* Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat Med* 2001;**7**:673–9.
59. Alizadeh AA, Eisen MB, Davis RE, *et al.* Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* 2000;**403**:503–11.
60. Vo JN, Cieslik M, Zhang Y, *et al.* The landscape of circular RNA in cancer. *Cell* 2019;**176**:869–81.e13.
61. Mottini C, Napolitano F, Li ZX, *et al.* Computer-aided drug repurposing for cancer therapy: approaches and opportunities to challenge anticancer targets. *Semin Cancer Biol* 2021;**68**:59–74.
62. Tsherniak A, Vazquez F, Montgomery PG, *et al.* Defining a cancer dependency map. *Cell* 2017;**170**:564–76.
63. Iorio F, Knijnenburg TA, Vis DJ, *et al.* A landscape of pharmacogenomics interactions in cancer. *Cell* 2016;**166**:740–54.
64. Coker EA, Mitsopoulos C, Tym JE, *et al.* canSAR: update to the cancer translational research and drug discovery knowledgebase. *Nucleic Acids Res* 2018;**47**:D917–22.
65. Koscielny G, An P, Carvalho-Silva D, *et al.* Open targets: a platform for therapeutic target identification and validation. *Nucleic Acids Res* 2017;**45**:D985–94.
66. Igarashi Y, Nakatsu N, Yamashita T, *et al.* Open TG-GATEs: a large-scale toxicogenomics database. *Nucleic Acids Res* 2015;**43**:D921–7.
67. Wishart DS, Feunang YD, Guo AC, *et al.* DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res* 2018;**46**:D1074–82.
68. Aubreville M, Knipfer C, Oetter N, *et al.* Automatic classification of cancerous tissue in laserendomicroscopy images of the oral cavity using deep learning. *Sci Rep* 2017;**7**:11979.
69. Granter SR, Beck AH, Papke DJ. Alphago, deep learning, and the future of the human microscopist. *Arch Pathol Lab Med* 2017;**141**:619–21.
70. Vang YS, Chen Z, Xie X. Deep learning framework for multi-class breast cancer histology image classification. 2018; *arXiv*: 1802.00931.
71. Wang X, Yang W, Weinreb J, *et al.* Searching for prostate cancer by fully automated magnetic resonance imaging classification: deep learning versus non-deep learning. *Sci Rep* 2017;**7**:15415.

72. Han Z, Wei B, Zheng Y, *et al.* Breast cancer multi-classification from histopathological images with structured deep learning model. *Sci Rep* 2017;**7**:4172.
73. Fan M, Zhang P, Wang Y, *et al.* Radiomic analysis of imaging heterogeneity in tumours and the surrounding parenchyma based on unsupervised decomposition of DCE-MRI for predicting molecular subtypes of breast cancer. *Eur Radiol* 2019;**29**: 4456–67.
74. Geras KJ, Wolfson S, Shen Y, *et al.* High-resolution breast cancer screening with multi-view deep convolutional neural networks. 2017; *arXiv*: 1703.07047.
75. Li H, Weng J, Shi Y, *et al.* An improved deep learning approach for detection of thyroid papillary cancer in ultrasound images. *Sci Rep* 2018;**8**:6600.
76. Fan M, Zheng HZ, Zheng S, *et al.* Mass detection and segmentation in digital breast tomosynthesis using 3D-mask region-based convolutional neural network: a comparative analysis. *Front Mol Biosci* 2020;**7**:599333.
77. Gurcan MN, Boucheron LE, Can A, *et al.* Histopathological image analysis: a review. *IEEE Rev Biomed Eng* 2009;**2**:147–71.
78. Eraslan G, Avsec Ž, Gagneur J, *et al.* Deep learning: new computational modelling techniques for genomics. *Nat Rev Genet* 2019;**20**:389–403.
79. Dumbrava EI, Meric-Bernstam F. Personalized cancer therapy-leveraging a knowledge base for clinical decision-making. *Cold Spring Harb Mol Case Stud* 2018;**4**:a001578.
80. Fan J, Slowikowski K, Zhang F. Single-cell transcriptomics in cancer: computational challenges and opportunities. *Exp Mol Med* 2020;**52**:1452–65.
81. Shao D, Huang L, Wang Y, *et al.* DeepSec: a deep learning framework for secreted protein discovery in human body fluids. *Bioinformatics* 2021;**2021**:btab545. https://doi.org/10.1093/bioinformatics/btab545.
82. Morais-Rodrigues F, Silvério-Machado R, Kato RB, *et al.* Analysis of the microarray gene expression for breast cancer progression after the application modified logistic regression. *Gene* 2020;**726**:144168.
83. Maros ME, Capper D, Jones DTW, *et al.* Machine learning workflows to estimate class probabilities for precision cancer diagnostics on DNA methylation microarray data. *Nat Protoc* 2020;**15**:479–512.
84. Albaradei S, Napolitano F, Thafar MA, *et al.* MetaCancer: a deep learning-based pan-cancer metastasis prediction model developed using multi-omics data. *Comput Struct Biotechnol J* 2021;**19**:4404–11.
85. Zeng X, Zhong Y, Lin W, *et al.* Predicting disease-associated circular RNAs using deep forests combined with positive-unlabeled learning methods. *Brief Bioinform* 2020;**21**:1425–36.
86. Kann BH, Hosny A, Aerts HJWL. Artificial intelligence for clinical oncology. *Cancer Cell* 2021;**39**:916–27.
87. Xu H, Anderson K, Grann VR, *et al.* Facilitating cancer research using natural language processing of pathology reports. *Stud Health Technol Inform* 2004;**107**:565–72.
88. Karimi YH, Blayney DW, Kurian AW, *et al.* Development and use of natural language processing for identification of distant cancer recurrence and sites of distant recurrence using unstructured electronic health record data. *JCO Clin Cancer Info* 2021;**5**:469–78.
89. Zeng J, Banerjee I, Henry AS, *et al.* Natural language processing to identify cancer treatments with electronic medical records. *JCO Clin Cancer Info* 2021;**5**:379–93.
90. Chan HCS, Shan H, Dahoun T, *et al.* Advancing drug discovery via artificial intelligence. *Trends Pharmacol Sci* 2019;**40**: 592–604.
91. Workman P, Antolin AA, Al-Lazikani B. Transforming cancer drug discovery with big data and AI. *Expert Opin Drug Discovery* 2019;**14**:1089–95.
92. Paul D, Sanap G, Shenoy S, *et al.* Artificial intelligence in drug discovery and development. *Drug Discov Today* 2021;**26**: 80–93.
93. Huang C, Clayton EA, Matyunina LV, *et al.* Machine learning predicts individual cancer patient responses to therapeutic drugs with high accuracy. *Sci Rep* 2018;**8**:16444.
94. Borisov N, Tkachev V, Suntsova M, *et al.* A method of gene expression data transfer from cell lines to cancer patients for machine-learning prediction of drug efficiency. *Cell Cycle* 2018;**17**:486–91.
95. Yanagisawa K, Toratani M, Asai A, *et al.* Convolutional neural network can recognize drug resistance of single cancer cells. *Int J Mol Sci* 2020;**21**:3166.
96. Ramsundar B, Eastman P, Walters P, *et al. Deep Learning for the Life Sciences: Applying Deep Learning to Genomics, Microscopy, Drug Discovery and More*. Sebastopol, CA: O'Reilly Media, 2019.
97. Mayr A, Klambauer G, Unterthiner T, *et al.* DeepTox: toxicity prediction using deep learning. *Front Environ Sci* 2016;**3**:80.
98. Napolitano F, Carrella D, Mandriani B, *et al.* gene2drug: a computational tool for pathway-based rational drug repositioning. *Bioinformatics* 2017;**34**:1498–505.
99. Kuhn M, Szklarczyk D, Pletscher-Frankild S, *et al.* STITCH 4: integration of protein–chemical interactions with user data. *Nucleic Acids Res* 2014;**42**:D401–7.
100. Senior AW, Evans R, Jumper J, *et al.* Improved protein structure prediction using potentials from deep learning. *Nature* 2020;**577**:706–10.
101. Aberle DR, Berg CD, Black WC, *et al.* The national lung screening trial: overview and study design. *Radiology* 2011;**258**: 243–53.
102. Jensen MA, Ferretti V, Grossman RL, *et al.* The NCI genomic data commons as an engine for precision medicine. *Blood* 2017;**130**: 453–9.
103. Kurnit KC, Bailey AM, Zeng J, *et al.* "Personalized cancer therapy": a publicly available precision oncology resource. *Cancer Res* 2017;**77**:e123–6.
104. Yu Y, Wang Y, Xia Z, *et al.* PreMedKB: an integrated precision medicine knowledgebase for interpreting relationships between diseases, genes, variants and drugs. *Nucleic Acids Res* 2019;**47**:D1090–101.
105. Griffith M, Spies NC, Krysiak K, *et al.* CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. *Nat Genet* 2017;**49**:170–4.
106. Shao D, Huang L, Wang Y, *et al.* HBFP: a new repository for human body fluid proteome. *Database* 2021;**2021**:baab065. 10.1093/database/baab065.
107. Cheerla A, Gevaert O. Deep learning with multimodal representation for pancancer prognosis prediction. *Bioinformatics* 2019;**35**:i446–54.
108. Adadi A, Berrada M. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE Access* 2018;**6**: 52138–60.
109. Fortelny N, Bock C. Knowledge-primed neural networks enable biologically interpretable deep learning on single-cell sequencing data. *Genome Biol* 2020;**21**:190.