Archives of Computational Methods in Engineering (2021) 28:4875–4896
https://doi.org/10.1007/s11831-021-09556-z

SURVEY ARTICLE

A Systematic Review of Applications of Machine Learning in Cancer Prediction and Diagnosis

Aman Sharma[1] · Rinkle Rani[2]

Received: 2 August 2020 / Accepted: 12 January 2021 / Published online: 25 January 2021

Abstract

Advancement in genome sequencing technology has empowered researchers to think beyond their imagination. Researchers are trying their hard to fight against various genetic diseases such as cancer. Artificial intelligence has empowered research in the healthcare sector. The availability of open-source healthcare datasets has motivated the researchers to develop applications which helps in early diagnosis and prognosis of diseases. Further, Next-generation sequencing has helped to look into detailed intricacies of biological systems. It has provided an efficient and cost-effective approach with higher accuracy. The advent of microRNAs also known as small noncoding genes has begun the paradigm shift in oncological research. We are now able to profile expression profiles of RNAs using RNA-seq data. microRNA profiling has helped in uncovering their relationship in various genetic and biological processes. Here in this paper, we present a review of the machine learning perspective in cancer research. The best way to develop effective cancer treatment/drugs is to better understand the intricacies and complexities involved in the cancer microenvironment. Although there has been a plethora of methods and techniques proposed in the literature, still the deadliness of cancer can't be reduced. In such a situation Artificial intelligence (AI) or machine learning is providing a reliable, fast, and efficient way to deal with such stringent diseases.

1 Introduction

Bioinformatics is playing a critical role in fighting against various severe diseases such as cancer, diabetics, Alzheimer's, etc.. Cancer is caused as a result of mutations and variations in the genetic microenvironment of an individual. There is huge amount of complexity in cancer microenvironment which results in treatment difficulty. Even if patients have same type of cancer still they will response differently towards same type of therapy. Clinical trials and the traditional drug discovery process is a time demanding and tedious task. Hence, researchers are trying their hard to design optimal treatment options for such stringent diseases.

1.1 Cancer research as machine learning problem

Availability of a huge amount of oncological and pharmacogenomics online data sources has boosted the research in this field. Unlike traditional statistical and computational approaches, bioinformaticians are using machine learning techniques to improve the treatment options in genetic diseases. Cells are the basic building block of all living organisms. There are variety of cells available in the human body such as blood cells, muscle cells, fat cells, etc. Genes are responsible for variation in these cells. Gene helps to carry heredity information and is responsible for various physical and functional processes in the body. Genes are responsible for heterogeneity in genotype and phenotype traits among species. All the information regarding the inheritance of phenotypic traits is carried by genes. Overall if one wants to fight against genetic disease then their root cause i.e. genes need to be studied. Advancement in computational biology and high throughput sequencing is helping to find biomarkers (genes) that are responsible for various diseases.

Further, chip technology in healthcare is considered the future of the healthcare industry which also provided lab-on-a-chip devices. These chips help in proper diagnosis and prognosis of patients based on their genetic profiles.

✉ Aman Sharma
aman.sharma@juit.ac.in; amans.3008@gmail.com

[1] CS/IT, Jaypee University of Information Technology, Solan, H.P, India

[2] Computer Science and Engineering Department, Thapar Institute of Engineering and Technology, Patiala 147001, India

Various researchers are trying hard to find gene or gene set that are causing genetic diseases. Microarray technology helps to measure the gene expression levels of a particular micro-environment. Along with gene expression data, we can collect (genome, transcriptome, and proteome) data such as copy number variations, gene mutation, etc.. Gene expression, drug response data is extensively used in identifying anti-cancer drugs, drug targets, and biomarkers. Some researchers are working to explore various biological pathways corresponding to genetic diseases. The ratio of the expression level of an individual gene under two variable conditions, obtained by DNA microarray hybridization is called gene expression value. The quantity of mRNA released by gene determines the gene expression value of the individual gene. This quantity may vary based on external stimuli. mRNA helps to carry the information from the genes about protein synthesis. Gene expression data has enormous potential in biological research. It can help to identify the genomic reason behind the occurrence of the physical process. Disease biomarkers can be identified with the help of differentiating genomic traits. Genomic assays of MNase-seq, m-RNA, DNase-seq can be fed to machine learning models to predict a variety of disease-related information. Figure 1 explains the central dogma of molecular biology explains the flow of genetic information. Many researchers are exploiting gene expression data related to genetic diseases like cancer to better understand the microenvironment. Figure 2 shows the omic data used for machine learning modeling. Cancer is a complex genetic disease involving various subtypes. There is a need to develop computational approaches that could aid in the treatment of tumor subtypes. Over the past decade, oncological research has gained serious attention and researchers are trying to personalize treatment therapies for cancer patients [1]. Apart from biomarker identification researchers are also working for developing computational (in-silico) models/algorithms that can predict disease-specific drug responses, drug synergy, and drug-target interactions.

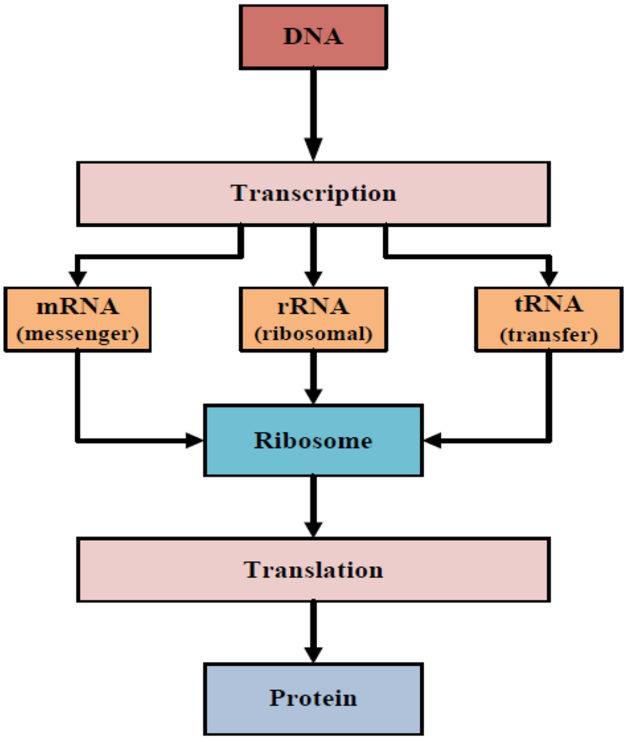

Fig. 1 Central Dogma of Biology

Many researchers are using machine learning algorithms to solve biological research problems. The supervised machine learning method is divided into three stages: learning, training, testing. In the learning phase, the machine learning algorithm is developed. In the training phase, a large amount of data is fed to the machine learning model to help it in making generalized rules out of it. In the testing phase, new data is fed to test the accuracy of the model prediction. Whereas, in unsupervised learning, data points are given but no labels are provided. The problem is to partition the data point in such a way that there should be maximum relevance and minimum redundancy.

The best way to develop effective cancer drugs is to better understand the intricacies and complexities involved in the cancer microenvironment. Although there has been a plethora of methods and techniques proposed in the literature, still the deadliness of cancer can't be reduced. In such a situation Artificial intelligence (AI) or machine learning is providing a reliable, fast, and efficient way to deal with such stringent diseases. For example, PathAI is one of the powerful AI-based tools, which is helping in the field of pathology. AI-enabled diagnosis is more fast, reliable, and accurate. AI can help to reduce the time lapse of clinical trials and the success rate of clinical trials can be predicted well in advance.

The task of the pathologist is to take out mass or lesion from a patient to put it on a glass slide for further observations. One slide can contain thousands of different cells. Even if there are one or two tumor cells in the sample, they are also important in the patient's treatment. So, pathologists have to deal with a large number of slides manually each day before making any decision regarding patients. Apart from this there are lots of other challenges such as there may be few cells to look upon because of the small size of a tissue or there may be so many cells that it is impossible to identify cancerous cells. So as a pathologist it is very challenging to pick cancerous cells out of normal cells.

In such a situation pathology slides can be digitized into digital pathology images. These images can be fed to the computer to recognize cancer cells Vs normal cells. Once the computer has finished earning you can apply the algorithms across all the images in your dataset. AI

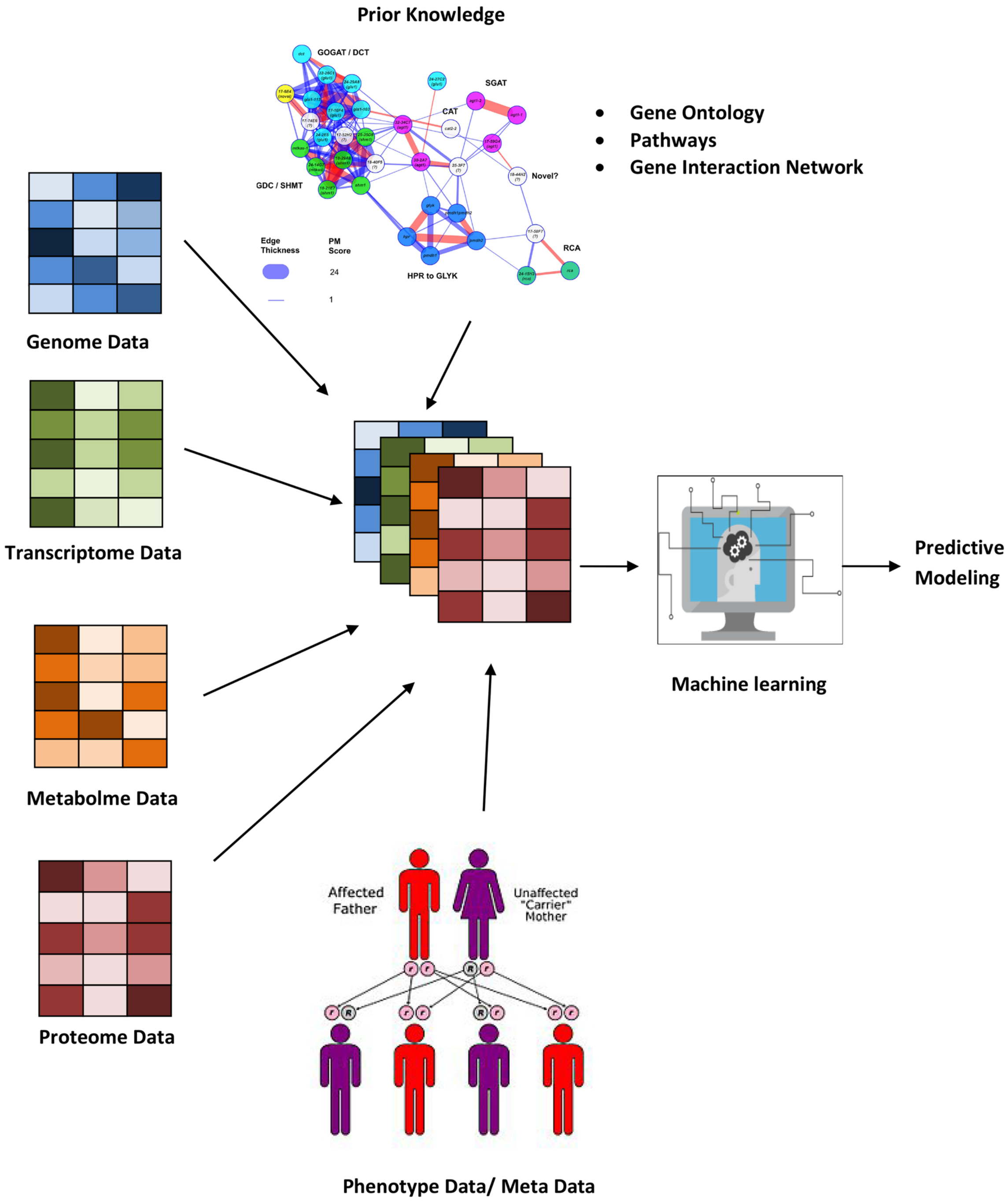

Fig. 2 Omic data used for Machine Learning Modeling

can help to find all the different cells that the pathologists manually classify on an image and do that automatically. AI can help us to match patients with the therapy that will maximize their chances of long-term survival.

1.2 Our contribution and organization of paper

Here in this paper, a review of the machine learning perspective in cancer research is presented. We have discussed various applications cancer using machine learning and their possible limitations, research issues etc. in detail. We have adressed various research questions and challenges corresponding to cancer research using machine learning. Further, we have also focus on machine learning techniques using microarray and NGS data.

The rest of the paper is organized as follows: Section 2 discusses the research methodology. This section describes the methods for selecting the literature. Section 3 presents a comparative summary of this survey with the already existing related surveys. Section 4 discusses challenges in using machine learning for cancer research. Section 5 provides a detailed discussion on Applications of Machine Learning in Cancer Research. Section 6 covers the future of cancer research using machine learning. Section 7 concludes the paper and discusses future directions. Figure 3 represents the complete layout of the manuscript.

2 Research Methodology

To conduct any kind of research or survey a research methodology has to be adopted. In this section, we have discussed the research methodology that helped us to conduct an extensive survey.

2.1 Research Questions asked by Researchers

The main motive of this review is to help young researchers in this field. There are many research questions that are addressed in this review paper. This review paper will help them to understand the basic terminology of cancer research using machine learning and to identify the key research problems in this area. These research questions are discussed in the Table 1.

2.2 Keywords for Searching Relevant Research Papers

Searching of papers is done based on the keywords related to cancer research using machine learning. Table 2 lists the keywords used for searching relevant papers. Initially, 4000 papers were shortlisted based on the searching and relevancy of this review. After that, further filtering is done to get insights from the most relevant papers. We have included papers from reputed journals and conferences. This search criterion helped a lot in this survey. Different keywords are used to find relevant research papers and articles.

3 Comparison with Existing Survey Papers

Various authors have attempted to review the literature on cancer research using machine learning. But most of the surveys are either focused on a single type of cancer or are not covering all the review questions mentioned in Table 1. Based on the review questions summarized in Table 1 we have compared the most relevant surveys with this survey. Table 3 summarizes the comparison of existing surveys on cancer research with our survey and highlights the prime difference of focus between them.

4 Challenges in Cancer Research using Machine learning

(a) *High dimensionality and imbalance class problem* Cancer data classification suffers from several issues like high dimensionality, imbalanced class problem. High dimensionality in data refers to the presence of an exceptionally large number of features as compared to samples. To deal with high dimensionality feature selection algorithms are designed. There are various methods and techniques [51–54] proposed in the literature for feature selection. However, still, no generic approach is developed which could handle all types of datasets and domains.

(b) *Model Biasedness* In class imbalance problem there is miss-match between the numbers of samples available for each class. It results in the biasedness of predictive models towards majority class samples. Various researchers have contributed solutions to this problem [55, 56]. But most of the existing work on cancer data classification is done using binary imbalanced classes; there is a need to address the imbalance problem in multi-class paradigm.

(c) *Heterogeneity in drug responses* Cancer patients showing heterogeneous response with the same cancer type has raised a major challenge of precision medication [57]. There is a need to develop a drug prediction model which could help in strengthening the present status of precision medication. There is no effective method to predict the drug responses of individual patients precisely and reliably. Genetic instability and variations among individuals are responsible for varied drug responses.

(d) *Efficient feature selection technique* Further, there is a need to propose a computationally efficient feature selection technique that could eliminate the need for the data cleaning procedures while generating high can-

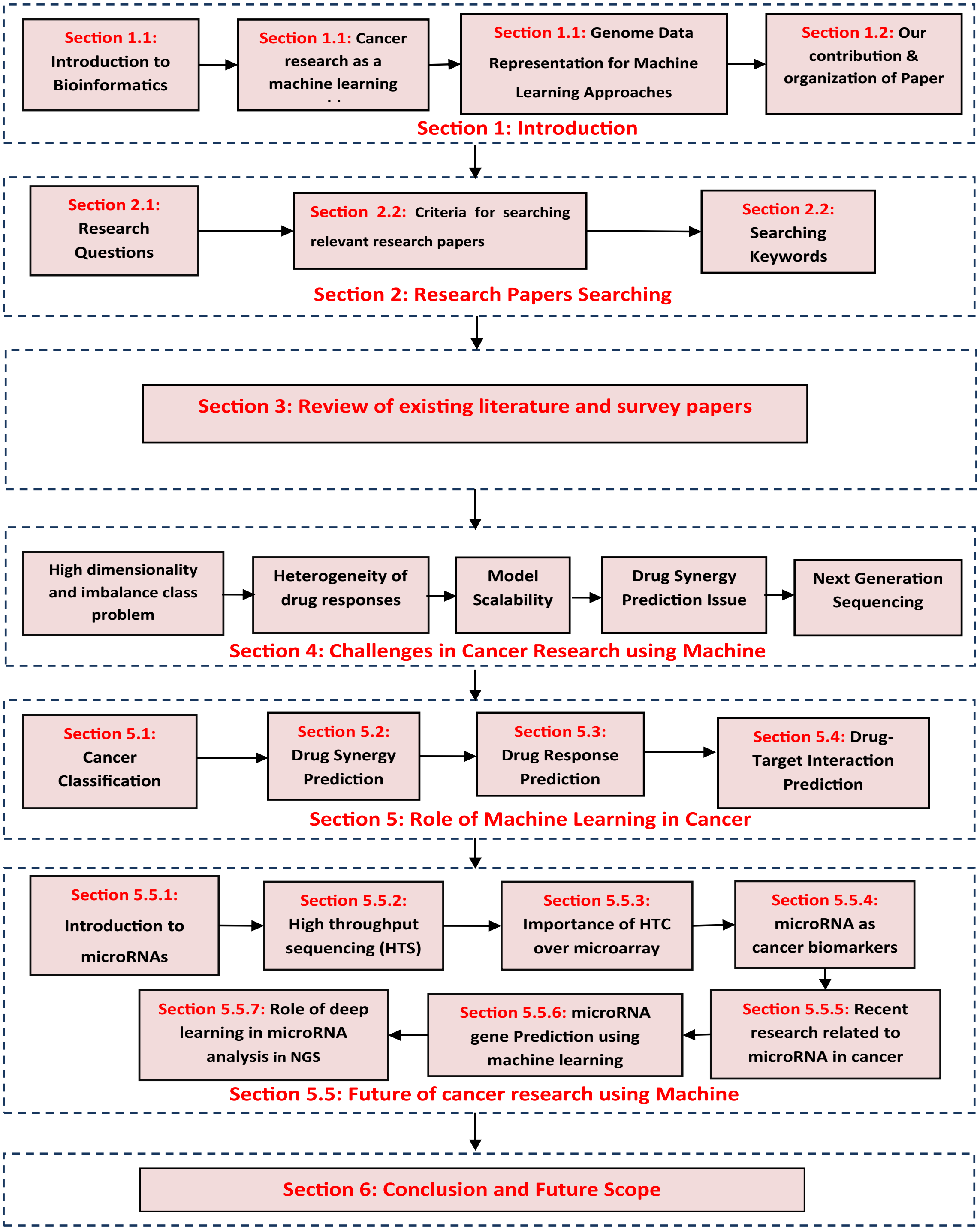
Section 1.1: Introduction to Bioinformatics
Section 1.1: Cancer research as a machine learning
Section 1.1: Genome Data Representation for Machine Learning Approaches
Section 1.2: Our contribution & organization of Paper
Section 1: Introduction
Section 2.1: Research Questions
Section 2.2: Criteria for searching relevant research papers
Section 2.2: Searching Keywords
Section 2: Research Papers Searching
Section 3: Review of existing literature and survey papers
High dimensionality and imbalance class problem
Heterogeneity of drug responses
Model Scalability
Drug Synergy Prediction Issue
Next Generation Sequencing
Section 4: Challenges in Cancer Research using Machine
Section 5.1: Cancer Classification
Section 5.2: Drug Synergy Prediction
Section 5.3: Drug Response Prediction
Section 5.4: Drug-Target Interaction Prediction
Section 5: Role of Machine Learning in Cancer
Section 5.5.1: Introduction to microRNAs
Section 5.5.2: High throughput sequencing (HTS)
Section 5.5.3: Importance of HTC over microarray
Section 5.5.4: microRNA as cancer biomarkers
Section 5.5.5: Recent research related to microRNA in cancer
Section 5.5.6: microRNA gene Prediction using machine learning
Section 5.5.7: Role of deep learning in microRNA analysis in NGS
Section 5.5: Future of cancer research using Machine
Section 6: Conclusion and Future Scope

Fig. 3 Layout of the manuscript

cer prediction accuracy with an optimal set of protein properties for drug design.

(e) *Model Scalability* Scalable feature selection technique is required which could consider maximum genetic aberrations simultaneously and efficiently [58]. There is a need to predict sensitive drugs for individual patients. As cancer is a complex disease and its complexity varies from patient to patient and one cannot rely on generalized medication and hence a scalable drug sensitivity criterion need to be taken into consideration.

(f) *Drug Synergy Prediction Issue* Machine learning potential for optimal drug synergy prediction are unexplored

Table 1 Summary of frequently asked research questions related to cancer and ML

S. No.	Research question
RQ1	What type of research is being done for cancer using machine learning?
RQ2	How cancer research using machine learning is different from traditional pathological studies?
RQ3	What is the classification of cancer using machine learning?
RQ4	What is drug response prediction using machine learning?
RQ5	What is drug repurposing?
RQ6	How is anti cancer drug target-interaction prediction done using machine learning?
RQ7	How is anti cancer drug synergy prediction done using machine learning?
RQ8	What is the status of research using machine learning on different types of cancers?
RQ9	What are the main performance evacuation parameters used for validating prediction results?
RQ10	What are the limitations of cancer using machine learning?
RQ11	What are the future directions in cancer research using machine learning?
RQ12	What is role of deep learning in cancer research?

Table 2 Keywords used for searching relevant papers

S. No.	Keywords
1	"Cancer classification using machine learning"
2	"Tumour classification using machine learning"
3	"Anti-Cancer drug response prediction using machine learning"
4	"Anti-Cancer drug synergy prediction using machine learning"
5	"Anti-Cancer drug target interaction prediction using machine learning"
6	"Drug repurposing using machine learning"
7	"Cancer gene selection using machine learning"
8	"Tumour gene selection using machine learning"
9	"Cancer and machine learning"
10	"Tumour and machine learning"

Table 3 Summary of survey papers on Cancer Research using Machine Learning

Related surveys considered	Reviewed up to	Research Questions											
		RQ1	RQ2	RQ3	RQ4	RQ5	RQ6	RQ7	RQ8	RQ9	RQ10	RQ11	RQ12
Nadeem et al. [105]	2020	✓	✓	✓						✓		✓	✓
Thakur et al. [106]	2020	✓	✓	✓								✓	✓
Sharif et al. [107]	2019	✓	✓	✓					✓	✓		✓	
Yassin et al. [108]	2018	✓	✓	✓						✓	✓	✓	
Chato et al. [109]	2017			✓						✓		✓	✓
Montazeri et al. [110]	2015	✓	✓	✓						✓			
Kourou et al. [111]	2015	✓	✓	✓					✓	✓	✓	✓	
Our review	2020	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

hence relevant machine learning models need to be developed for the proper diagnosis and treatment of stringent diseases like cancer**.** Drug synergism helps in designing novel drug combinations which could complement each other to suppress the progression of the disease. There is a need to extract potential drug combination features to understand drug-disease interaction in a holistic manner.

(g) *Next Generation Sequencing (NGS)* Analyzing NGS dataset using machine learning is also one of the biggest challenges that researchers are facing. Advancement in genome sequencing technology has empowered researchers to think beyond their imagination. Next-generation sequencing has helped to look into detailed intricacies of biological systems. It has provided an efficient and cost-effective approach with higher accuracy.

Advent of microRNAs also known as small non-coding genes has begun the paradigm shift in oncological research. We are now able to profile expression profiles of RNAs using RNA-seq data. microRNA profiling is helping in uncovering their relationship in various genetic and biological processes.

5 Applications of Machine Learning in Cancer

Microarray data analysis deals with gene classification, clustering using statistical approaches. Apart from statistical approaches**,** machine learning algorithms such as Decision Tree, Neural Networks, Support Vector Machine (SVM), and Random Forest are also used for microarray data analysis. Moreover we find literature evidences for various computational approaches using machine learning for drug synergy prediction, drug response prediction, and drug-target interaction prediction and biomarker identification. All these computational approaches help in identifying potential drug molecules for various diseases. Cancer is one of the most researched diseases which have gained huge attention from academia and pharmacy industries.

5.1 Cancer Classification

As we have already discussed that gene expression data has enormous potential in interpreting the significance of genes and their correlation with disease. To better understand the disease, the patient's gene expression data is collected in different biological environments. A comparison-based data analysis is performed to understand the disease state. The amount of mRNA produced by a gene tells about the active and inactiveness of genes. With the rapid advancement in computational biology research, there is a huge demand for microarray data. It is helping in developing predictive machine earning models which could help in cancer classification. Moreover, microarray data helps in the precise prediction of cancer types. Figure 4 describes the cancer classification steps using machine learning.

Many researchers have contributed different methods/techniques for tumor classification using microarray data [2]. These techniques varies from statistical methods to machine learning techniques for tumor classification. Microarray data suffers from the issue of high dimensionality of data. Feature selection algorithms are used to deal with this issue [3]. With the use of feature selection algorithms, model training time gets reduced as a result of the removal of irrelevant features. Scalability and generalization are two constraints that restrict the functioning of traditional feature selection algorithms. Deep Neural Networks (DNN) can be used in automatic feature extraction and develop generalized and scalable models.

Technological advancement in DNA-microarray has widely pushed the research in bioinformatics. Further, with the introduction of NGS (Next Generation Sequencing) we can sequence the whole genome structure of any individual. Scientists are performing parallel screening of gnomonic data to fetch the hidden patterns which could help in drug discovery. Such a parallel screening helps to identify gene–gene relationships, potential biomarkers for different genetic diseases, and genetic mutations/alterations. This parallel screening helps to early detect many rigorous diseases such as cancer. Over the last two decades, various bioinformaticians have collaborated to contribute to open-source tumor data sets [4–6] to boost cancer research. These datasets are generally microarray data of thousands of genes for different tissues (Patients). These are used as benchmark datasets to carry out data analysis/prediction for

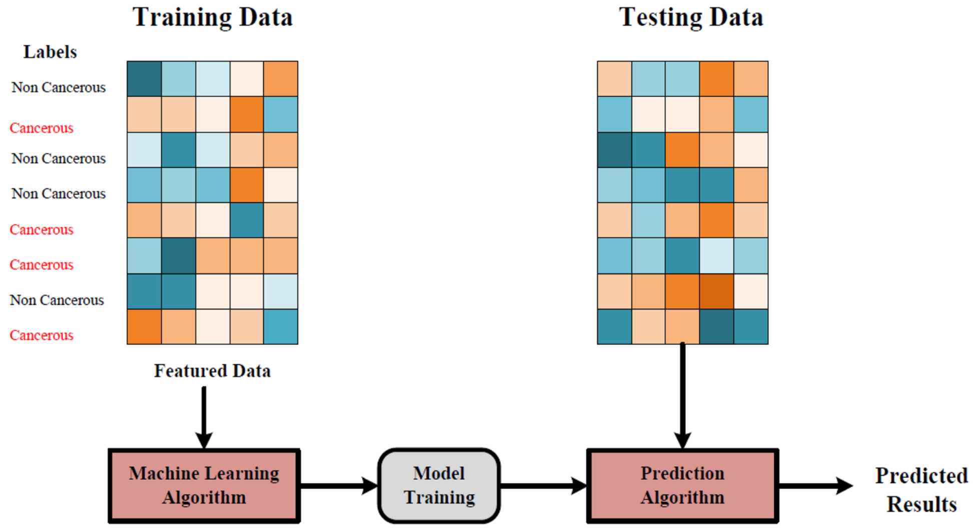

Fig. 4 Cancer classification using machine learning

Table 4 Datasets available for cancer classification

S. No.	Dataset	Link
1	UCI repository	https://mub.me/ONU
2	Breast Cancer Wisconsin (Diagnostic)	https://bit.ly/3fpVLbz
3	Leukemia	https://bit.ly/2Pgy8Yv
4	Lung Cancer	https://bit.ly/2XhBpeu
5	Lung Cancer	https://data.world/cancerdatahp/lung-cancer-data
6	Skin Cancer	https://bit.ly/3k0gE0B
7	All cancer types	https://www.cancerimagingarchive.net/collections/
8	Breast Cancer	https://datahub.io/machine-learning/breast-cancer
9	Lung Cancer	https://www.xenonstack.com/use-cases/lung-cancer-detection/
10	Head& Neck	https://bit.ly/319e2VD
11	Herlev database	http://fuzzy.iau.dtu.dk/download/smear2005

personalized medication and cancer classification. Machine learning is also used to exploit the potential of these datasets. Table 4 contains the datasets available for tumour classification. Various researchers have developed tumour classification techniques using machine learning [7–10]. Machine learning majorly focuses on identifying hidden patterns in data that could help to generalize the biological process/system. The key idea in cancer classification is to improve the classification model prediction accuracy and to find a minimum set of potential gene biomarkers.

Although all this seems to very interesting and easy the reality is that there are many key issues involved while designing the biological predictive modeling. Genes identification for tumor sub-type analysis is a tedious task as it depends on feature selection algorithms. These feature selection algorithms are dependent on optimization algorithms or statistical approaches that need to be defined very carefully for proper results. Broadly feature selection algorithms are classified as a wrapper, hybrid and filter methods. The filter method depends on the statistical background on data to identify the key genes which could serve as biomarkers [11]. Wrapper methods are based on a suitable learning approach to filter out the most relevant genes [11]. Wrapper methods have the benefit of delivering higher accuracy [12].

Microarray data has the issue of data high-dimensionality and this makes tumour classification a NP-hard problem. To solve such problems meta-heuristic algorithms are treated as an optimal choice [11]. Multi-objective functions are the real beauty of these algorithms as they help to find the global best solution. Conflicts between different objective functions have been resolved to fetch the optimal results. Many of these algorithms are bio-inspired optimization algorithms [13, 14, 17, 18]. Broadly they are classified as posterior-based [16] and prior-based [15] approaches. The concept of weighted multi-objective functions is used in prior approaches. Posterior approaches focus on the performance of the problem of finding an optimal solution. Table 5 contains a summary of selected cancer classification techniques using machine learning.

5.2 Drug Synergy Prediction

Targeted drug therapy is the most commonly used treatment given to cancer patients. These drugs are specially designed based on their targets which help to suppress cancer. These targets are known as anti-oncogene which is responsible for tumor suppression by suppressing mitosis (cell-division) [19]. Any alteration, changes in these genes lead to uncontrollable cell growth. Unlike these genes, there are oncogenes that promote tumor growth. Most of the targeted drug therapies are designed considering oncogenes as anti-oncogenes are hard to target. Various studies revealed the resistance of targeted drug therapies and hence results in nonresponsive drug behavior [20, 21]. This resistance may have occurred because of many reasons such as cell death inhibition, change in drug targets, etc.. Heterogeneous tumor microenvironment can also result in drug resistance [22]. Combination drug therapy can help to avoid drug resistance. It helps in overcoming the drug resistance by delaying tumor growth. It includes the usage of two or more drugs in fixed dose proportion and as a single dose formulation. Table 6 contains the datasets available for anti-cancer drug synergy prediction.

Combination therapy is showing excellent results in tumor suppression by reducing the chances of multiple mutations [23] and a single mutation [24] that can escape all the drugs.

Additionally, combination therapy helps in lowering drug dosage, side-effects [23]. A combination of two or more drugs is considered effective if the tumor suppression rate of combination is higher than individual drugs. Such a combination of drugs is known as synergistic drugs otherwise antagonistic. The proposition of dose also matters in drug synergy, we cannot mix them in any random

Table 5 Summary of selected cancer classification techniques using machine learning

References	Proposed technique	Contribution	Data sets	Performance parameters
Guyon et al. [7]	SVM technique based on Recursive Feature Elimination (RFE)	Gene Selection for Cancer Classification	Leukemia [4], Colon cancer [5]	leave-one-out success rate
Shen et al. [112]	Penalized Logistic Regression	Tumour Classification Using Microarray Data	Breast, Colon, Acute Leukemia lung, Ovarian, Prostate cancer, Central Nervous system	Classification Accuracy, Computational Time, Penalty Parameter
Wang et al. [113]	correlation-based feature selector, decision trees, naïve Bayes and SVM	Gene selection from microarray data	Leukemia [4]	CPU time (in seconds), Accuracy
Feng et al. [114]	Fuzzy Neural Network	Gene Selection and Cancer Classification	Lymphoma Data [115], SRBCT Data [116], Liver Cancer Data [117]	Number of genes, Accuracy
Wang et al. [118]	Gene Importance Ranking, Support Vector Machines (SVMs)	Finding the smallest set of genes	Lymphoma Data [115], SRBCT Data [116], Liver Cancer Data [117], GCM [119]	Number of genes, Accuracy
Cho & Won [120]	Ensemble of neural networks	Cancer classification	Leukemia, Colon, and Lymphoma data	Number of genes, Accuracy, Principal component analysis
Tan et al. [121]	Fuzzy neural network	Ovarian cancer diagnosis	Micro-array gene expression [122], Blood assays, Proteomic spectra [123]	Sensitivity, Specificity, Accuracy, Training time (s)
Glaab et al. [124]	Rule-Based Machine Learning	Gene Prioritization and Sample Classification	Prostate cancer [125], lymphoma [126], Breast cancer [127]	Average accuracy, Friedman test
Liu et al. [128]	Recursive Feature Addition, Supervised learning	Gene selection and classification	Six benchmark microarray gene expression data sets	Accuracy, Minimize the redundancy of the genes
Chen et al. [129]	Particle swarm optimization, Decision tree classifier	Cancer classification	10 datasets from GEMS, Taiwan Cancer Registry [130]	Accuracy, ANOVA, p-value
Margoosian & Abouei [131]	Ensemble-based Classifiers	Cancer Classification	BENCHMARK FOURTEEN CANCER DATA SET [132]	Classification accuracy
Zaher & Eldeib [133]	Deep Belief Networks	Cancer Classification	Wisconsin breast cancer	Confusion matrix, Misclassified sample rate
Dwivedi [134]	Artificial neural network (ANN)	Cancer classification	Leukemia [4]	Sensitivity, Specificity, Precision, Misclassification rate
Sevakula et al. [135]	Transfer Learning, Deep Neural Networks	Molecular Cancer Classification	(GEMLeR) [136]	Execution time, AUC, Statistical Tests
Ting et al. [137]	Convolutional Neural Network	Breast Cancer Classification	Mammographic Image Analysis Society dataset [138]	Sensitivity, Accuracy, AUC
Ghoneim et al. [139]	Convolutional neural networks & extreme learning machines	Cervical cancer classification	Herlev database (http://fuzzy.iau.dtu.dk/download/smear2005)	Accuracy, False negative, False positive, Confusion matrix
Yu et al. [140]	Deep Residual Networks	Automated Melanoma Recognition	Skin Lesion Analysis [141]	AUC, Sensitivity, Accuracy, Specificity
Albarqouni et al. [142]	Deep Learning From Crowds for Mitosis	Breast Cancer Detection	MICCAI-AMIDA13 challenge dataset [143]	ROC curves, Precision, Recall, and F1 score
Wang et al. [144]	Mean-Shift clustering algorithm and mathematical morphology	Classification of cervical Pap smear images	362 dataset images of cervical Pap smear	Accuracy, P-value
Zhang et al. [145]	Deep Convolutional Networks	Cervical Cell Classification	Cervical cytology images	Accuracy, AUC, Sensitivity, Specificity, F-measure, H-mean

Table 6 Datasets available for anti-cancer drug synergy prediction

S. No.	Dataset	Link
1	DrugComb	https://bit.ly/319f2Js
2	NCI-ALMANAC	https://bit.ly/3fmZp6i
3	DREAM Challenge Dataset	https://bit.ly/2XghYCH
4	Combination therapies dataset for melanoma	https://pubmed.ncbi.nlm.nih.gov/23239741/

proportions. Quantify drug synergy is a very complex task but still few researchers have given metrics to measure it. Some of the quantitative methods for drug synergy are the Bliss independence model [27], Dose equivalence, Isobolographic analysis [25], and Chou-Talalay [26]. Table 7 contains a summary of the anti-cancer drug synergy prediction approaches with respect to ML.

5.3 Drug Response Prediction

Abnormal mutations and changes in genes lead to cancer and also disrupt the normal functioning of cellular activities. Exposure of cells to an unfavourable environment promotes tumor growth. Understanding tumor microenvironment complexity is one of the challenging tasks. Even if patients have same type of cancer still they will response differently towards same type of therapy. Genetic differences among patients are the main reason for the difference in drug responses. Cancer patients can't be given medications based on their anatomical origin. An individual patient's genomic profile needs to be considered while making suitable prescriptions [29]. Treating cancer patients with better drugs and diagnosis is still a challenging task. Table 8 contains datasets available for drug sensitivity prediction.

Large-scale drug screening data is providing a helping hand in identifying the relationship between genes and drug responses. Datasets(Pharmacogenomics) are produced as a result of such large scale screenings. GDSC [30] and CCLE [31] are two such large databases which helps to promote oncological research.

Machine learning techniques are used in modelling cancerous research problems such as predicting drug responses, genomic biomarkers. Machine learning models such as random forest and elastic net regularization are the most frequently used in drug response prediction. Matrix factorization is one of the popularly used technique in drug response prediction [32]. Trust prorogation based technique is used by Jamali et al. [33] for predicting drug responses. Regularized factorization methods is also used in bioinformatics, brain activities prediction [34]. Table 9 contains a summary of selected drug sensitivity prediction techniques using machine learning.

5.4 Drug-Target Interaction Prediction

Drug discovery involves finding novel drugs and their novel potential targets. Identifying such drug target interaction out of pool of drugs and targets is a tedious task. Current research on drugs aims to repurpose an already existing drug for new diseases and targets. Drug repurposing for new diseases and target helps in saving time and money as the repurposed drugs are already approved. Drug-target interactions involve two sets of agents: Chemicals form the drug set and amino acid form a target set. This research problem has a vital role in discovering new drugs and to recognize new potential targets of it. They play a significant contribution to understand the operation of drugs and their side effects. However, there exist some key issues related to drug discovery such as toxicity towards patients, drug resistance, time-consuming clinical trials. Difference in drug effects on patients [35, 36] and mapping of drug effect with the drug interaction pathway [28] are the key issues discussed in the literature. Table 10 contains the datasets available for drug target interaction prediction.

We can predict drug target interactions using either of the two methods: clinical/experimental (in vivo) or with the help of computational (in silico) methods. These methods are classified as: Docking [38, 39], ligand-based [40], literature text mining [41], and pharmacogenomics [42, 43] methods. Clinical methods are inefficient, tiring, and even difficult to reproduce [44].

Clinical docking techniques are most widely used techniques but their time-consuming simulations and non availability of 3-D structure of proteins are major drawbacks. Using simulation techniques these methods predict about the target site for a given drug. There are some other similarity based techniques too that uses the similarity between targets (ligands) but no proper information about majority of the target ligands resulted in less popularity of these methods. One another method Literature text mining explores the literature to find out the relationship between the given drug and target. But they are also not so popular because of lack of information. Apart from these methods computational methods such as machine learning techniques and kernel-based are also used to find out potential drug-target interactions [45, 46]. Various online databases are available that provide access to the data related to compounds and target proteins [47–50]. These databases help to boost the research related to DTI. Various researches have used these databases in their studies to identify novel drug target interactions [42]. Table 11 is the summarization of drug target interaction prediction techniques with respect to ML.

Table 7 Summary of Anti-cancer Drug Synergy Prediction approaches with respect to ML

References	Proposed technique	Contribution	Data sets	Performance parameters
Kim et al. [146]	Deep neural networks	Anti-cancer Drug Synergy Prediction	Genetic data from multiple databases	Sensitivity, AUC, Accuracy
Jiang et al. [147]	Graph Convolutional Network (GCN) model	Prioritizing synergistic anticancer drug combinations	O'Neil et al.'s dataset [149]	AUC, AUPRC, Accuracy, Kappa
Ekşioğlu & Tan [148]	Ensemble Learning	Prediction of Drug Synergy	Large compound oncology dataset [149]	Mean Squared Error (MSE), Pearson correlation coefficient
Zhang et al. [150]	Deep Learning Model	Predicting Tumor Cell Response to Synergistic Drug Combinations	NCI ALMANAC database, Cancer cell line encyclopedia (CCLE) database, KEGG (Kyoto Encyclopedia of Genes and Genomes)	Pearson correlation coefficient
Kuru et al. [151]	Deep learning framework	Drug Synergy Prediction	DrugComb	Correlation, Mean squared error (MSE)
Preuer et al. [152]	Deep Learning	Predicting anti-cancer drug synergy	large-scale oncology screen [149]	MSE, P-value, RMSE, Pearson's r
Wildenhan et al. [153]	Random forest and Naive Bayesian learner	Prediction of Synergism from Chemical-Genetic Interactions	CGM dataset	AUC, ROC, Gini Index
Janizek et al. [154]	Extreme gradient boosted tree-based approach	Prediction of synergistic drug combinations	O'Neil et al.'s dataset [149]	Mean squared error (MSE), Five fold cross-validation
Mason et al. [155]	Machine Learning	Predict Synergistic Antimalarial Compound Combinations	1,540 antimalarial drug combinations	CV (cross-validation)
Chen et al. [156]	Deep belief network, ontology fingerprints	Predict effective drug combination	DREAM Challenge dataset	Precision, Recall, F1
Sharma and Rani [157]	Machine learning algorithms	Identification of effective and synergistic anti-cancer drug combinations	DREAM Challenge Dataset, Held et al. [158]	Accuracy, Specificity, Sensitivity

Table 8 Datasets available for Drug sensitivity prediction

S. No.	Dataset	Link
1	CCLE	https://bit.ly/2XlU8pc
2	GDSC	https://bit.ly/3flVJ4z
3	TCGA	https://bit.ly/3hXSvpF
4	NCI	https://bit.ly/3hXSvpF
5	CancerDR: Cancer Drug Resistance Database	https://bit.ly/39OQyJx
6	CancerMine	http://bionlp.bcgsc.ca/cancermine/
7	canSAR	https://cansarblack.icr.ac.uk/
8	Mutations and Drugs Portal (MDP)	https://ieeexplore.ieee.org/document/7545951
9	cBioPortal	https://www.cbioportal.org/
10	Liver Cancer	http://liverome.kobic.re.kr/

5.5 Cancer research using Next Generation Sequencing and Machine learning

In this section, we will discuss the future of anti-cancer drug prediction approaches in context to big data and Next Generation Sequencing (NGS). We will discuss the use of deep learning in anti-cancer drug prediction and how it will help to foster the research in this domain.

5.5.1 Introduction to microRNAs

microRNAs are small non-coding RNAs that bind to 3 UTR regions of their target mRNA. They are newly discovered types of RNAs, shorter in length as compared to other RNAs. Generally, mature microRNAs are single-stranded and 18–24 nucleotide long. They play an important role in controlling the post transcription regulation of coding genes, either by degrading them or inhibiting their translation. The translation is a post transcription cellular mechanism for protein synthesis with the help of ribosomes. Ribosomes decode the mRNA produced by DNA transcription. On the other hand, degradation is the process of ceasing mRNA translation. Their initial research gained momentum because of the keen interest of some researchers but later they were identified as a predominant component in the cellular mechanism. Each microRNA has been identified as a controller of a wide range of target genes [57]. These microRNAs regulate mRNAs but there is also a regulatory body known as polymerase-2 which regulates microRNAs [58, 59]. It is an enzyme used in the catalysis of DNA transcription during the synthesis of microRNA and other RNAs. Biological synthesis of microRNA is a seamlessly regulated process where multiple sub-processes are involved.

5.5.2 High Throughput Sequencing (HTS)

High throughput sequencing (HTS) is the use of modern technologies in the field of sequencing. It is also popularly known by another name, which is Next-generation sequencing (NGS). Advancement in computational capabilities has brought a radical change in the field of genome sequencing. These HTS technologies can generate a large amount of biological data at a much faster rate and in a cost-effective manner. We can perform deep sequencing and quantification of complete genome sequence transcriptomes. HTS has led to evolutionary insight into various biological processes and macromolecules identification.

These sequencing technologies are used to profile various genomic profiles to reveal the underlying biological aspects and interactions. It has provided the ability to look into interactions between proteomes, transcriptomes, and genomes. HTS technology has fostered research in characterizing small RNA transcriptomes. There are various platforms and technologies such as Illumina (Solexa), SOLID sequencing for HTS. RNA-seq is the most widely used RNA sequencing method using HTS, it allows wide-scale transcriptome analysis with higher resolution and lesser errors. Mostly RNA-seq experiments are based on a common protocol.

The basic principle for NGS is identical to Electrophoresis sequencing, the only difference lies in the incorporation of parallelization in the sequencing of DNA fragments by NGS. Illumina sequencing is the most preferred chemistry in academics and industry because of its accuracy. In RNA-seq experiments firstly complete RNA is extracted from the sample under consideration and further, it can be profiled to fetch individual microRNAs and mRNAs before library preparation. There are various steps involved in library preparation such as RNA fragmentation, reverse transcription, amplification, quantification, and quality control. Further, we can perform downstream analysis of fetched sequences to find differential expressions, novel transcripts [95]. Figure 5 contains the biogenesis of microRNA.

5.5.3 Importance of HTS over microarray

Microarray technology is serving the biological community from the last two decades; researchers have found their expertise in this technology hence most of the traditional sequencing was based on hybridization profiling. But with the advancement in Next-generation sequencing (NGS) technology and reduction in its cost have pushed researchers interest in it. NGS has the advantage of delivering more accurate profiling results as compared to microarray technology. Microarray has limitations in identifying novel

Table 9 Summary of selected Drug sensitivity prediction techniques using machine learning

References	Proposed technique	Contribution	Data sets	Performance parameters
Jang et al. [159]	110,000 different models, multifactorial experimental design testing	DRUG SENSITIVITY PREDICTION DRUG SENSITIVITY PREDICTION Drug sensitivity prediction	Cancer cell lines (CCLE), Sanger	IC50, AUC, ANOVA
Menden et al. [160]	Neural networks and Random forests	Prediction of Cancer Cell Sensitivity to Drugs using Genomic and Chemical Properties	GDSC	Root mean square error (RMSE), Coefficient of determination (R^2), Pearson correlation coefficient (R_p)
Turki et al. [161]	Transfer Learning	Drug Sensitivity Prediction in Multiple Myeloma Patients	(GEO) repository (http://www.ncbi.nlm.nih.gov/geo/)	P-values of t-test, AUC, Mean AUC (MAUC)
Wan & Pal [162]	Ensemble Learning	Drug Sensitivity Prediction	NCI-DREAM Challenge dataset, Cancer Cell Line Encyclopedia	Accuracy, Leave-one-out errors, Statistical significance
Dong et al. [163]	Support Vector Machine (SVM) and a recursive feature selection	Anticancer drug sensitivity prediction	CCLE, CGP	Accuracy, AUC
Rahman et al. [164]	Ensemble mode, Random Forests	Drug Sensitivity Prediction	CCLE, GDSC databases	MSE, AUC
Yuan et al. [165]	Multitask learning	Prediction of cancer drug sensitivity	CCLE, CTD2, NCI60	MSE, fivefold cross-validation
Ali and Aittokallio [166]	Machine learning, feature selection	Drug response prediction	NCI-DREAM Challenge	MSE, Accuracy
Haider et al. [167]	Multivariate Random Forests	Drug Sensitivity Prediction	GDSC and CCLE	Accuracy, AUC
He et al. [168]	Kernelized Rank Learning (KRL)	Personalized drug recommendation	Cancer cell lines, Clinical trials	Precision, Standard deviations
Matlock et al. [169]	Random Forests	Drug sensitivity prediction	Synthetic data, CCLE data	Prediction accuracy and AUC, MSE
Riddick et al. [170]	Random Forests, Ensemble approach,classification and regression trees	Predicting in vitro drug sensitivity	NCI-60, 19 Breast Cancer and 7 Glioma cell lines	R^2, correlation coefficients
Sharma and Rani et al. [171]	Ensemble and multi-task learning	Drug sensitivity prediction	NCI-Dream dataset, CCLE dataset	Wilcoxon ranksum test, CV std. error, MSD (Mean square deviation)
Sharma and Rani et al. [172]	Ensembled machine learning	Drug sensitivity prediction	GDSC, CCLE	MSE, paired t-test, Wilcoxon signed-rank test

Table 10 Datasets available for drug-target interaction prediction

S. No.	Dataset	Link of the dataset
1	CancerDR	https://bit.ly/39VoZhG
2	Drug Target Commons	https://bit.ly/3gmlgfr
3	DTI databases	https://bit.ly/2BR2ePb
4	DrugBank	https://www.drugbank.ca/
5	ChEMBL	https://bit.ly/3fpXbmp
6	ChemBank	https://bit.ly/2BTTKHa
7	STITCH	http://stitch.embl.de/
8	BindingDB	https://bit.ly/2Xn2jlo
9	TDR Targets	https://bit.ly/2Xn2kFY
10	SIDER	https://bit.ly/3fmXIFI

microRNAs and biasing in results can be introduced. Various comparative studies have identified NGS a better microRNA profiling approach as compared to the microarray. Results showed that many microRNAs would have remained undetected if they used microarray technology [96]. Profiling of microRNA expressions using NGS serves the potential in uncovering their relationship in many diseases and predicting their role in precision medication.

5.5.4 microRNA as Cancer Biomarkers

As cancer is a stringent disease with complex and inexpedient diagnosis procedures, so an inefficient diagnosis can lead to serious health impacts on patients. A cancer biomarker can be any variation in biological processes, tissue, or molecules that can predict cancer or its subtype significantly [61]. So there is an urgent need for good cancer markers which could strengthen the present state of diagnosis. Various studies revealed microRNAs as potential biomarkers for the diagnosis and prognosis of cancer [62]. microRNAs modulates their target genes [63] by suppressing their normal expression state [64, 67]. microRNAs can serve as biomarkers for different kinds of cancer. The presence of circulating microRNA (plasma and serum) in blood tissues of cancer patients has embossed it as an optimal diagnostic marker [68]. Other than these microRNA biomarkers, urinary microRNA is also considered potentially viable for prostatic cancer [69]. Recent advancement in high throughput

Table 11 Summarization of Drug-target interaction prediction techniques with respect to ML

References	Proposed technique	Contribution	Data Sets	Performance parameters
Ezzat et al. [173]	Ensemble learning, dimensionality reduction	Drug-target interaction prediction	drug-target interaction data [174], Second dataset [175]	Sensitivity Analysis, AUC
Chen et al. [176]	Machine Learning	Drug-Target Interaction Prediction	Review on databases such as DrugBank, KEGG, and STITCH	
Wen et al. [177]	Deep learning	Drug-target interaction prediction	'golden standard' dataset [41]	TPR, TNR, Accuracy, AUC
Yuan et al. [178]	Ensemble learning, k -nearest neighbor, Bipartite Local Model with support vector classification	Improving drug–target interaction prediction	DrugBank [179]	AUPR, precision, recall
Ezzat et al. [174]	Class imbalance-aware ensemble learning	Drug-target interaction	DrugBank database [48]	AUC
Zhang et al. [180]	A random projection ensemble approach	Drug-target interaction prediction	Dataset [181]	Precision, recall, Accuracy, F1-measure
Xie et al. [182]	Deep learning	Transcriptome data classification for drug-target interaction prediction	LINCS project, DTI database [183]	Accuracy, Predictive errors
Tian et al. [183]	Deep neural network (DNN)	Compound-protein interaction prediction	STITCH database [185], PubChem database [185], Pfam database [186]	Accuracy, Sensitivity, specificity, F1-measure
Feng et al. [187]	Deep Learning	Drug-Target Interaction Prediction	He et al. [188] Davis dataset [189], Metz dataset [190] and KIBA dataset [191]	R^2, RMSE
Xie et al. [192]	Deep-learning-based model	Drug–target interaction prediction	L1000 dataset	Accuracy, F-score, proportion of positive cases and predictive error
Sharma and Rani [193]	Dimensionality Reduction and Active Learning	Drug Target Interaction Prediction	Drug Bank [45], Ezzat et al. [174]	AUPR, AUC, sensitivity, specificity, accuracy

technologies has raised the possibility of easy identification of microRNA and their targets. Assays/Sequences generated through these technologies are platform-independent and provides better analytical and statistical inference. Table 12 summarizes the various microRNAs as potential cancer diagnostic biomarkers in blood.

5.5.5 Recent research related to microRNA in cancer

As we have already discussed the significant role of microRNA in various diseases and drug therapies, still identification of novel microRNAs and their targets is a challenging issue. Although various tools and pipelines have been developed the inconsistency between their results and no standardized approach has raised a serious issue among researchers in this field.

Still, researchers are trying their hard to relate these newly identified disease predictors with targeted drug therapies. Recently microRNA 374b is identified as a resistive agent in pancreatic cancer drug therapy [90]. The major goal of new generation drug prediction is to predict novel drugs that could be useful in a wide range of diseases. The Scripps Research Institute (TSRI) researchers have designed a drug that has shown tumor suppressor capabilities in breast cancer animal models [91]. Breast cancer is one of the most researched cancer types because of its intricacy and commonality among women, therefore deeper knowledge about its subtypes and drug therapies can help to fight against it. Circulating microRNAs have been identified as potential biomarkers for early detection of breast cancer [92]. Recently a study revealed the deregulation of microRNAs in the tumor environment and their role in cancer cell lines [93]. Laura Cantini, et al. have proposed a pipeline for subtype identification and analysis of colorectal cancer using an interaction network of mRNA-microRNA [94].

5.5.6 micoRNA Gene Prediction using machine learning

microRNAs are the essence and indeed need to present biomedical research. We have already discussed the importance and role of microRNAs in various biological systems. Many microRNAs have been identified but still many more to be discovered. Due to the limitation of biological experimental approaches microRNA identification suffers from serious bottleneck and hence efficient computational approaches are needed for the identification and prediction of novel

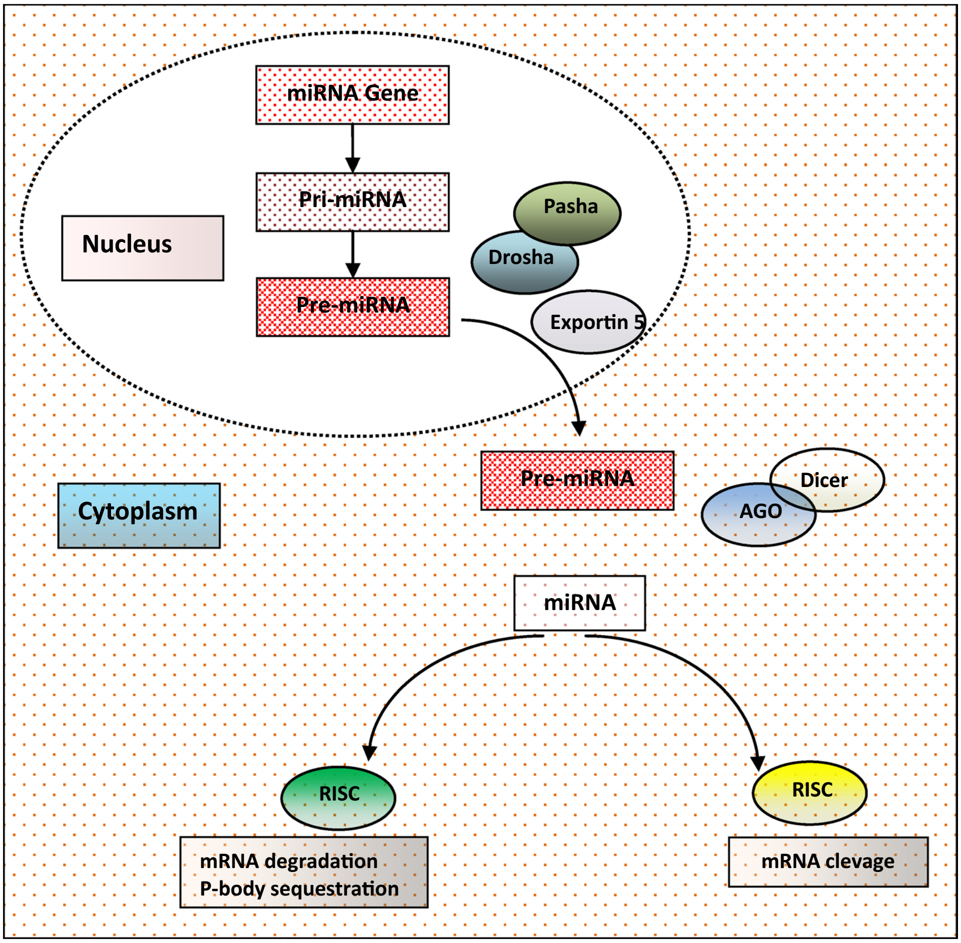

Fig. 5 Biogenesis of microRNA

microRNAs. Most of the real-world problems are complex in nature, which makes them difficult to model.

ML approaches can help in modeling such complex problems and to incorporate data-driven decision-making capabilities in resultant models. We can apply ML approaches on microRNA data for their identification, their target genes, and then further analysis of microRNA expression data. NGS has given a powerful platform for discovering new microRNAs and their targets. NGS platforms like Illumina/Solexa GA are popularly used platforms that give more accurate expression values as compared to hybridization-based technologies. As a result, significant improvement has been seen in microRNA identification and their targets. ML approaches can classify candidate target genes corresponding to identified microRNA. Classifiers such as Random Forest, SVM, and Decision Trees are used popularly. Figure 6 describes the generalized workflow for microRNA gene prediction using ML. The basic idea here is that ML will generalize the prediction rules based on the positive and negative data sets. A positive data set contains microRNA sequences that have been already identified and a negative dataset contains microRNA look-alike sequences, that are not microRNAs. Most of the available methods for microRNA gene prediction rely on structure similarity of the hairpin structure of pre-microRNA. They are based on the principle of homologous structure identification, if we could find microRNA in one genome, then there is a possibility of identifying it in another genome too. Homology modeling and ab-initio are presently available methods for microRNA prediction in bioinformatics. The homology technique is a simple method that predicts microRNAs based on existing information from already identified microRNAs. It is the sequence alignment technique, nothing new can be predicted regarding microRNAs.

There are various tools available based on the homology technique and ML such as ProMir [97] and MirFinder [98]. In contrast to homology-based methods, ab-initio methods are not similarity-based, they do not require any additional reference sequence for predicting microRNAs. Proper parameter selection can lead to the prediction of new microRNAs but if not selected properly can result in high false-positive predictions. Ab-initio methods also use ML capabilities and there is software such as MiPred [99], MiRenSVM [100], Triplet-SVM [101] based on it. The availability of a huge amount of biological data has raised the need for new data handling, prediction, and classification algorithms. Traditional methods are no more reliable enough to handle such an enormous growth of data. In such a scenario ML approaches are considered an optimal choice for better results. ML approaches are used in various fields of bioinformatics such as genomics, proteomics, transcriptomics, and system biology. ML algorithms for the prediction of microRNAs start with the training step to build an expert model. A model is designed based on the learning it gathers from sequence data, microRNA structure, and intensity data of microRNAs. Based on learning from these features it can classify unknown sequences as microRNA or not. But these ML algorithms suffer from serious class imbalance problem

Table 12 Summarization microRNAs as potential Cancer Diagnostic Biomarkers in Blood

S. No.	Cancer type	microRNAs	References
1	Pancreas	miR—2001, 200b, 210, 155, 18 a	Li et al. [68] Ho et al. [69] Wang et al. [70] Morimura et al. [71]
2	Prostate	miR—141	Mitchell et al. [72]
3	Breast	miR—21,155,195 and let-71	Zhu et al. [73] Heneghan et al. [74] Asaga [75] Zhao [76]
4	Lung	miR—21, 25, 126, 223, 155, 197 and 182	Chen et al. [77] Shen et al. [78] Zheng et al. [79]
5	Ovarian	miR—21, 141, 200c, 203, 205, 214, 200a, 200b, 92, 93, 126, 155, 127, 99b	Taylor et al. [80] Resnick et al. [81]
6	Gastric	miR—17-5p, 106a, 106b, 32, 182, 143 and 21	Tsujiura et al. [82] Li et al. [83]
7	Liver	miR—199, 195, 16, 500	Yamamoto et al. [84] Qu et al. [85]
8	Esophageal	miR—223, 133a, 127-3p, 22, 10a, 100 and 148b	Zhang et al. [86]
9	Squamous cell-Tongue	miR-184	Wong et al. [87]
10	Colorectal	miR-92, 29, 17-3p	Ng et al. [87] Huang et al. [89]

and most of the algorithms consider fixed loop size stems, reducing the overall prediction accuracy. Learning in ML models for microRNA predictions is based on positive and negative data. Majority of the time these datasets are derived from mirBase [104], although a little of pre-processing is needed before actually using it.

5.5.7 Role of deep learning in microRNA analysis in NGS

"Big Data" has been a buzz topic in the recent years; it has gained huge interest from academics as well as industry. The rate at which data is being produced has increased to many folds and so is the research in this field. Data related to bioinformatics has also evolved over many years. An increase in computational capabilities and the emergence of HTS technology has lead to a sudden outburst of biomedical data. This data serves a great potential in identifying disease biomarkers, discovering new drugs, but unfortunately, it is not effectively utilized. NGS technologies have created a serious need for new technologies and algorithms. In such a scenario deep learning using neural networks is considered an effective choice. Although ML approaches have been used for many years they have a limitation of processing raw data.

Deep learning is a new version of ML algorithms that incorporate artificial intelligence using multilayer neural networks. In contrast to traditional ML approaches, deep learning can extract features from data itself. In efforts to apply deep learning algorithms to microRNA prediction, researchers have proposed various deep learning algorithms. Seunghyun Park, et al. has proposed deepMiRGene [103] an algorithm to predict microRNA precursor. They used RNN, there is no need to input features manually, and the algorithm automatically identifies features from input data. This approach leads to the discovery of various new features too which can be used in future research. Similarly Cheng S., et al. developed MiRTDL [104] an algorithm for microRNA target prediction using CNN. It automatically extracts desired information from the data itself rather than relying on information fed manually. These algorithms have shown efficient results and have improved prediction results. The use of deep learning techniques in microRNA and their target prediction can help in novel microRNA predictions and one can investigate better knowledge about the underlying mechanism.

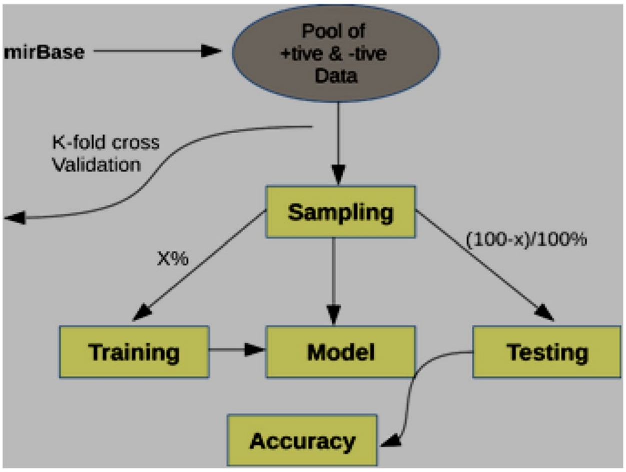

Fig. 6 Generalized Workflow for machine learning microRNA gene prediction

6 Conclusion and Future Directions

Although various researchers are working in the field of cancer but still there are various possible future directions which need to be addressed. Heterogeneous omic data can be considered to further improve the performance of cancer classification. Drug synergy data need to be extracted so as to foster the research in this field. The heterogeneous drug response of individuals need to be understand and considered while developing predictive models. Copy number variation, somatic mutation, and pathways can be further considered in predicting drug responses. Genomic data integration can be performed to further improve prediction results.

Further, apart from microarray data, we can use microRNAs which are small non-coding RNAs that bind to 3 UTR regions of their target mRNA. They play an important role in controlling the posttranslational regulation of coding genes, either by degrading them or inhibiting their translation. Various microRNA have been identified and many more to be discovered from a pool of genomic data. Various computational and statistical approaches are proposed to leverage the best results out of sequencing data. NGS technology is popularly used these days due to cost reduction, higher accuracy; as a result, we need efficient algorithms and pipelines which could cater to the present need. Machine learning and deep learning algorithms can prove useful in handling NGS data and develop biomedical applications. Using these technologies we can predict promising microRNA biomarkers which could later be used as drug targets for a variety of diseases. Hence microRNAs have paved the path for the precision medication in fighting against cancer. Identifying novel and tissue-specific microRNA can help to differentiate significantly between healthy and diseased cell states. This paper attempts to highlight the possible application areas of anticancer drug prediction using machine learning, NGS data using machine learning, and how microRNAs can help in better diagnosis and prognosis of cancer. This review paper is an attempt to summarize the various research directions for cancer using machine learning.

Funding None.

Compliance with ethical standards

Conflict of interest None.

References

1. Błaszczyński J, Stefanowski J (2015) Neighbourhood sampling in bagging for imbalanced data. Neurocomputing 150:529–542
2. Ying Lu, Han J (2003) Cancer classification using gene expression data. Inf Syst 28(4):243–268
3. Oleg O (2013) Survey of novel feature selection methods for cancer classification. Biological knowledge discovery handbook preprocessing mining, and postprocessing of biological data, pp 379–398
4. Golub Todd R, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H (1999) Molecular classification of cancer class discovery and class prediction by gene expression monitoring. Science 286(5439):531–537
5. Alon U, Barkai N, Notterman DA, Gish K, Ybarra S, Mack D, Levine AJ (1999) Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. Proc Natl Acad Sci 96(12):6745–6750
6. Van't Veer LJ, Dai H, Van De Vijver MJ, He YD, Hart AA, Mao M, Friend SH (2002) Gene expression profiling predicts clinical outcome of breast cancer. Nature 415(6871):530–536
7. Guyon I, Weston J, Barnhill S, Vapnik V (2002) Gene selection for cancer classification using support vector machines. Mach Learn 46(1–3):389–422
8. Shevade SK, Sathiya Keerthi S (2003) A simple and efficient algorithm for gene selection using sparse logistic regression. Bioinf 19(17):2246–2253
9. Furlanello C, Serafini M, Merler S, Jurman G (2003) Gene selection and classification by entropy-based recursive feature elimination. In: Proceedings of the international joint conference on neural networks, 4:3077–3082. IEEE
10. Chu W, Ghahramani Z, Falciani F, Wild DL (2005) Biomarker discovery in microarray gene expression data with Gaussian processes. Bioinf 21(16):3385–3393
11. Saeys Y, Inza I, Larrañaga P (2007) A review of feature selection techniques in bioinformatics. Bioinf 23(19):2507–2517
12. Inza I, Larrañaga P, Blanco R, Cerrolaza AJ (2004) Filter versus wrapper gene selection approaches in DNA microarray domains. Artif Intell Med 31(2):91–103
13. Shen Qi, Shi W-M, Kong W (2008) Hybrid particle swarm optimization and tabu search approach for selecting genes for tumor classification using gene expression data. Comput Biol Chem 32(1):53–60
14. Li S, Xixian Wu, Tan M (2008) Gene selection using hybrid particle swarm optimization and genetic algorithm. Soft Comput 12(11):1039–1048
15. Branke J, Deb K, Dierolf H, Osswald M (2004) Finding knees in multi-objective optimization. International conference on parallel problem solving from nature. Springer, Berlin, Heidelberg, pp 722–731
16. Marler RT, Arora JS (2004) Survey of multi-objective optimization methods for engineering. Struct Multidiscip Optim 26(6):369–395
17. BoussaïD I, Lepagnot J, Siarry P (2013) A survey on optimization metaheuristics. Inf Sci 237:82–117
18. Chakraborty A, Kar AK (2017) Swarm intelligence: a review of algorithms. In: Nature-inspired computing and optimization. Springer, pp 475–494
19. Weinberg RA (1991) Tumor suppressor genes. Science 254(5035):1138–1146
20. Knoechel B, Roderick JE, Williamson KE, Zhu J, Lohr JG, Cotton MJ, Gillespie SM (2014) An epigenetic mechanism of resistance to targeted therapy in T cell acute lymphoblastic leukemi. Nat Genet 46(4):364–370
21. Rini BI, Atkins MB (2009) Resistance to targeted therapy in renal-cell carcinoma. Lancet Oncol 10(10):992–1000
22. Housman G, Byler S, Heerboth S, Lapinska K, Longacre M, Snyder N, Sarkar S (2014) Drug resistance in cancer an overview. Cancers 6(3):1769–1792
23. Fitzgerald JB, Schoeberl B, Nielsen UB, Sorger PK (2006) Systems biology and combination therapy in the quest for clinical efficacy. Nat Chem Biol 2(9):458–466
24. Cokol M, Chua HN, Tasan M, Mutlu B, Weinstein ZB, Suzuki Yo, Nergiz ME (2011) Systematic exploration of synergistic drug pairs. Mol Syst Biol 7(1):544–553
25. Tallarida RJ (2011) Quantitative methods for assessing drug synergism. Genes Cancer 2(11):1003–1008
26. Ashton JC (2015) Drug combination studies and their synergy quantification using the Chou-Talalay method. Cancer Res 75(11):2400–2400
27. Foucquier J, Guedj M (2015) Analysis of drug combinations current methodological landscape. Pharmacol Res Perspect 3(3):00149
28. Kotelnikova E, Yuryev A, Mazo I, Daraselia N (2010) Computational approaches for drug repositioning and combination therapy design. J Bioinf Comput Biol 8(3):593–606
29. Xiao G, Ma S, Minna J, Xie Y (2014) Adaptive prediction model in prospective molecular signature-based clinical studies . Clin Cancer Res 20(3):531–539
30. Garnett MJ, Edelman EJ, Heidorn SJ, Greenman CD, Dastur A, Lau KW, Greninger P (2012) Systematic identification of genomic markers of drug sensitivity in cancer cells. Nature 483(7391):570–575
31. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, Wilson CJ (2012) The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature 483(7391):603–607
32. Yamada M, Lian W, Goyal A, Chen J, Wimalawarne K, Khan SA, Chang Y (2017) Convex factorization machine for toxicogenomics prediction. In: Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining, pp 1215–1224
33. Jamali M, Ester M (2010) A matrix factorization technique with trust propagation for recommendation in social networks. In: Proceedings of the fourth ACM conference on Recommender systems, pp 135–142
34. Wang L, Li X, Zhang L, Gao Q (2017) Improved anticancer drug response prediction in cell lines using matrix factorization with similarity regularization. BMC Cancer 17(1):513–524
35. Evans WE, McLeod HL (2003) Pharmacogenomics drug disposition, drug targets, and side effects. N Engl J Med 348(6):538–549
36. Wei D-Q, Wang J-F, Chen C, Li Y, Chou K-C (2008) Molecular modeling of two CYP2C19 SNPs and its implications for personalized drug design. Protein Pept Lett 15(1):27–32

37. Mizutani S, Pauwels E, Stoven V, Goto S, Yamanishi Y (2012) Relating drug–protein interaction network with drug side effects. Bioinformatics 28(18):i522–i528
38. Rarey M, Kramer B, Lengauer T, Klebe G (1996) A fast flexible docking method using an incremental construction algorithm. J Mol Biol 261(3):470–489
39. Xie Li, Evangelidis T, Xie L, Bourne PE (2011) Drug discovery using chemical systems biology weak inhibition of multiple kinases may contribute to the anti-cancer effect of nelfinavir. PLoS Comput Biol 7(4):e1002037
40. Jacob L, Vert J-P (2008) Protein-ligand interaction prediction an improved chemogenomics approach. Bioinformatics 24(19):2149–2156
41. Zhu S, Okuno Y, Tsujimoto G, Mamitsuka H (2005) A probabilistic model for mining implicit chemical compound–gene relations from literature. Bioinformatics 21(2):ii245–ii251
42. Yamanishi Y, Araki M, Gutteridge A, Honda W, Kanehisa M (2008) Prediction of drug–target interaction networks from the integration of chemical and genomic spaces. Bioinformatics 24(13):i232–i240
43. Wang Y-C, Zhang C-H, Deng N-Y, Wang Y (2011) Kernel-based data fusion improves the drug–protein interaction prediction. Comput Biol Chem 35(6):353–362
44. Fakhraei S, Huang B, Raschid L, Getoor L (2014) Network-based drug-target interaction prediction with probabilistic soft logic. IEEE/ACM Trans Comput Biol Bioinf (TCBB) 11(5):775–787
45. van Laarhoven T, Nabuurs SB, Marchiori E (2011) Gaussian interaction profile kernels for predicting drug–target interaction. Bioinformatics 27(21):3036–3043
46. Zheng X, Ding H, Mamitsuka H, Zhu S (2013) Collaborative matrix factorization with multiple similarities for predicting drug-target interactions. In: Proceedings of the 19th ACM SIGKDD international conference on knowledge discovery and data mining, pp 1025–1033
47. Bolton EE, Wang Y, Thiessen PA, Bryant SH (2008) PubChem integrated platform of small molecules and biological activities. Ann Rep Comput Chem 4:217–241
48. Knox C, Law V, Jewison T, Liu P, Ly S, Frolkis A, Pon A (2010) DrugBank 30 a comprehensive resource for ‘omics’ research on drugs. Nucleic Acids Res 39(1):D1035–D1041
49. Gaulton A, Bellis LJ, Patricia Bento A, Chambers J, Davies M, Hersey A, Light Y (2011) CHEMBL a large-scale bioactivity database for drug discovery. Nucleic Acids Res 40(D1):D1100–D1107
50. Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M (2011) KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res 40(D1):D109–D114
51. Hira ZM, Gillies DF (2015) A review of feature selection and feature extraction methods applied on microarray data. Adv Bioinf 2015(198363):1–13
52. Bolón-Canedo V, Sánchez-Maroño N, Alonso-Betanzos A (2016) Feature selection for high-dimensional data. Prog Artif Intell 5(2):65–75
53. Haixiang G, Yijing Li, Jennifer Shang Gu, Mingyun HY, Bing G (2017) Learning from class-imbalanced data: review of methods and applications. Expert Syst Appl 73:220–239
54. Krawczyk B, Galar M, Jeleń Ł (2016) Francisco Herrera Evolutionary undersampling boosting for imbalanced classification of breast cancer malignancy. Appl Soft Comput 38:714–726
55. Dagogo-Jack I, Shaw AT (2018) Tumour heterogeneity and resistance to cancer therapies. Nat Rev Clin Oncol 15(2):81–94
56. Tadist K, Najah S, Nikolov NS, Mrabti F, Zahi A (2019) Feature selection methods and genomic big data: a systematic review. J Big Data 6(79):1–24
57. Bartel DP (2009) Micrornas: target recognition and regulatory functions. Cell 136:215–233
58. Lee Y, Kim M, Han J, Yeom K-H, Lee S, Baek SH, Narry Kim V (2004) Microrna genes are transcribed by rna polymerase ii. EMBO J 23(20):4051–4060
59. Xie Z, Allen E, Fahlgren N, Calamar A, Givan SA, Carrington JC (2005) Expression of arabidopsis mirna genes. Plant Physiol 138(4):2145–2154
60. Richard Lu, Barca O (2012) Fine-tuning oligodendrocyte development by micrornas. Front Neurosci 6:13
61. Hayes DF, Bast RC, Desch CE, Fritsche H, Kemeny NE, Jessup JM, Locker GY, Macdonald JS, Mennel RG, Norton L et al (1996) Tumor marker utility grading system: a framework to evaluate clinical utility of tumor markers. J Natl Cancer Inst 88(20):1456–1466
62. Garzon R, Marcucci G, Croce CM (2010) Targeting micrornas in cancer: rationale, strategies and challenges. Nat Rev Drug Discovery 9(10):775–789
63. Ambros V (2003) Microrna pathways in flies and worms: growth, death, fat, stress, and timing. Cell 113(6):673–676
64. Doench JG, Sharp PA (2004) Specificity of microrna target selection in translational repression. Genes Dev 18(5):504–511
65. Zhang H, Kolb FA, Brondani V, Billy E, Filipowicz W (2002) Human dicer preferentially cleaves dsrnas at their termini without a requirement for atp. EMBO J 21(21):5875–5885
66. Kosaka N, Iguchi H, Ochiya T (2010) Circulating microrna in body uid: a new potential biomarker for cancer diagnosis and prognosis. Cancer Sci 101(10):2087–2092
67. Ploussard G, de la Taille A (2010) Urine biomarkers in prostate cancer. Nat Rev Urol 7(2):101–109
68. Li A, Omura N, Hong S-M, Vincent A, Walter K, Grith M, Borges M, Goggins M (2010) Pancreatic cancers epigenetically silence sip1 and hypomethylate and overexpress mir-200a/200b in association with elevated circulating mir-200a and mir-200b levels. Cancer Res 70(13):5226–5237
69. Ho AS, Huang X, Cao H, Christman-Skieller C, Bennewith K, Le Q-T, Koong AC (2010) Circulating mir-210 as a novel hypoxia marker in pancreatic cancer. Transl Oncol 3(2):109–113
70. Wang J, Chen J, Chang P, LeBlanc A, Li D, Abbruzzesse JL, Frazier ML, Killary AM, Sen S (2009) Micrornas in plasma of pancreatic ductal adenocarcinoma patients as novel blood-based biomarkers of disease. Cancer Prev Res 2(9):807–813
71. Morimura R, Komatsu S, Ichikawa D, Takeshita H, Tsujiura M, Nagata H, Konishi H, Shiozaki A, Ikoma H, Okamoto K et al (2011) Novel diagnostic value of circulating mir-18a in plasma of patients with pancreatic cancer. Br J Cancer 105(11):1733–1740
72. Mitchell PS, Parkin RK, Kroh EM, Fritz BR, Wyman SK, Pogosova- EL, Agadjanyan AP, Noteboom J, O’Briant KC, Allen A et al (2008) Circulating micrornas as stable blood-based markers for cancer detection. Proc Natl Acad Sci 105(30):10513–10518
73. Zhu W, Qin W, Atasoy U, Sauter ER (2009) Circulating micrornas in breast cancer and healthy subjects. BMC Res Notes 2(1):89
74. Heneghan HM, Miller N, Kelly R, Newell J, Kerin MJ (2010) Systemic mirna-195 differentiates breast cancer from other malignancies and is a potential biomarker for detecting noninvasive and early stage disease. Oncologist 15(7):673–682
75. Asaga S, Kuo C, Nguyen T, Terpenning M, Giuliano AE, Hoon DSB (2011) Direct serum assay for microrna-21 concentrations in early and advanced breast cancer. Clin Chem 57(1):84–91
76. Zhao H, Shen J, Medico L, Wang D, Ambrosone CB, Liu S (2010) A pilot study of circulating mirnas as potential biomarkers of early stage breast cancer. PLoS ONE 5(10):e13735
77. Chen Xi, Ba Yi, Ma L, Cai X, Yin Y, Wang K, Guo J, Zhang Y, Chen J, Guo X et al (2008) Characterization of micrornas in serum: a novel class of biomarkers for diagnosis of cancer and other diseases. Cell Res 18(10):997–1006

78. Shen J, Liu Z, Todd NW, Zhang H, Liao J, Lei Yu, Guarnera MA, Li R, Cai L, Zhan M et al (2011) Diagnosis of lung cancer in individuals with solitary pulmonary nodules by plasma microrna biomarkers. BMC Cancer 11(1):1
79. Zheng D, Haddadin S, Wang Y, Li-Qun Gu, Perry MC, Freter CE, Wang MX (2011) Plasma micrornas as novel biomarkers for early detection of lung cancer. Int J Clin Exp Pathol 4(6):575–586
80. Taylor DD, Gercel-Taylor C (2008) Microrna signatures of tumor-derived exosomes as diagnostic biomarkers of ovarian cancer. Gynecol Oncol 110(1):13–21
81. Resnick KE, Alder H, Hagan JP, Richardson DL, Croce CM, Cohn DE (2009) The detection of differentially expressed micrornas from the serum of ovarian cancer patients using a novel real-time pcr platform. Gynecol Oncol 112(1):55–59
82. Tsujiura M, Ichikawa D, Komatsu S, Shiozaki A, Takeshita H, Kosuga T, Konishi H, Morimura R, Deguchi K, Fujiwara H et al (2010) Circulating micrornas in plasma of patients with gastric cancers. Br J Cancer 102(7):1174–1179
83. Li X, Luo F, Li Q, Meihua Xu, Feng D, Zhang G, Wei Wu (2011) Identification of new aberrantly expressed mirnas in intestinal-type gastric cancer and its clinical significance. Oncol Rep 26(6):1431–1439
84. Yamamoto Y, Kosaka N, Tanaka M, Koizumi F, Kanai Y, Mizutani T, Murakami Y, Kuroda M, Miyajima A, Kato T et al (2009) Microrna-500 as a potential diagnostic marker for hepatocellular carcinoma. Biomarkers 14(7):529–538
85. Qu KZ, Zhang Ke, Li HaiRong, Afdhal NH, Albitar M (2011) Circulating micrornas as biomarkers for hepatocellular carcinoma. J Clin Gastroenterol 45(4):355–360
86. Zhang C, Wang C, Chen Xi, Yang C, Li Ke, Wang J, Dai J, Zhibin Hu, Zhou X, Chen L et al (2010) Expression profile of micrornas in serum: a fingerprint for esophageal squamous cell carcinoma. Clin Chem 56(12):1871–1879
87. Wong T-S, Liu X-B, Wong B-H, Ng R-M, Yuen A-W, Wei WI (2008) Mature mir-184 as potential oncogenic microrna of squamous cell carcinoma of tongue. Clin Cancer Res 14(9):2588–2592
88. Sung JJ, Chong WS, Jin H, Lam EK, Shin VY, Yu J, Poon TC, Ng SS, Ng EK (2009) 1070 Differential Expression of MicroRNAs in Plasma of Colorectal Cancer Patients: A Potential Marker for Colorectal Cancer Screening. Gastroenterol 136(5):A-165
89. Huang Z, Huang D, Ni S, Peng Z, Sheng W, Xiang Du (2010) Plasma micrornas are promising novel biomarkers for early detection of colorectal cancer. Int J Cancer 127(1):118–126
90. Schreiber R, Mezencev R, Matyunina LV, McDonald JF (2016) Evidence for the role of microRNA 374b in acquired cisplatin resistance in pancreatic cancer cells. Cancer Gene Ther 23(8):241–245
91. Velagapudi SP, Cameron MD, Haga CL, Rosenberg LH, Lafitte M, Duckett DR, Phinney DG, Disney MD (2016) Design of a small molecule against an oncogenic noncoding RNA. Proc Natl Acad Sci 113(21):5898–5903
92. Hamam R, Ali AM, Alsaleh KA, Kassem M, Alfayez M, Aldahmash A, Alajez NM (2016) microRNA expression profiling on individual breast cancer patients identifies novel panel of circulating microRNA for early detection. Sci Rep 6(1):1–8
93. Rupaimoole R, Calin GA, Lopez-Berestein G, Sood AK (2016) mirna deregulation in cancer cells and the tumor microenvironment. Cancer Discov 6(3):235–246
94. Cantini L, Isella C, Petti C, Picco G, Chiola S, Ficarra E, Caselle M, Medico E (2015) MicroRNA–mRNA interactions underlying colorectal cancer molecular subtypes. Nat Commun 6(1):1–2
95. Mortazavi A, Williams BA, McCue K, Schaefier L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by rna-seq. Nat Methods 5(7):621–628
96. Murakami Y, Tanahashi T, Okada R, Toyoda H, Kumada T, Enomoto M, Tamori A, Kawada N, Taguchi YH, Azuma T (2014) Comparison of hepatocellular carcinoma miRNA expression profiling as evaluated by next generation sequencing and microarray. PLoS ONE 9(9):e106314
97. Nam J-W, Shin K-R, Han J, Yoontae Lee V, Kim N, Zhang B-T (2005) Human microrna prediction through a probabilistic co-learning model of sequence and structure. Nucleic Acids Res 33(11):3570–3581
98. Huang T-H, Fan B, Rothschild MF, Zhi-Liang Hu, Li K, Zhao S-H (2007) Mirfinder: an improved approach and software implementation for genome-wide fast microrna precursor scans. BMC Bioinf 8(1):1
99. Ng KLS, Mishra SK (2007) De novo svm classification of precursor micrornas from genomic pseudo hairpins using global and intrinsic folding measures. Bioinformatics 23(11):1321–1330
100. Ding J, Zhou S, Guan J (2010) Mirensvm: towards better prediction of microrna precursors using an ensemble svm classifier with multi-loop features. BMC Bioinf 11(11):1
101. Xue C, Li F, He T, Liu G-P, Li Y, Zhang X (2005) Classification of real and pseudo microrna precursors using local structure-sequence features and support vector machine. BMC Bioinf 6(1):310
102. Ana Kozomara and Sam Griffiths-Jones (2014) mirbase: annotating high confidence micrornas using deep sequencing data. Nucleic Acids Res 42(D1):D68–D73
103. Seunghyun Park, Seonwoo Min, Hyunsoo Choi, and Sungroh Yoon (2016) deepmirgene: Deep neural network based precursor microrna prediction. arXiv preprint arXiv:1605.00017
104. Cheng S, Guo M, Wang C, Liu X, Liu Y, Xuejian Wu (2015) MiRTDL: a deep learning approach for miRNA target prediction. IEEE ACM Trans Comput Biol Bioinf 13(6):1161–1169
105. Nadeem MW, Ghamdi MA, Hussain M, Khan MA, Khan KM, Almotiri SH, Butt SA (2020) Brain tumor analysis empowered with deep learning: A review, taxonomy, and future challenges. Brain Sci 10(2):118
106. Thakur SK, Singh DP, Choudhary J (2020) Lung cancer identification: a review on detection and classification. Cancer Metastasis Rev
107. Sharif MI, Li JP, Naz J, Rashid I (2020) A comprehensive review on multi-organs tumor detection based on machine learning. Pattern Recognit Lett 131:30–37
108. Yassin NI, Omran S, El Houby EM, Allam H (2018) Machine learning techniques for breast cancer computer aided diagnosis using different image modalities: A systematic review. Comput Methods Progr Biomed 156:25–45
109. Chato L, Latifi S. (2017) Machine learning and deep learning techniques to predict overall survival of brain tumor patients using MRI images. In: 2017 IEEE 17th international conference on bioinformatics and bioengineering (BIBE), pp 9–14
110. Montazeri M, Montazeri M, Montazeri M, Beigzadeh A (2016) Machine learning models in breast cancer survival prediction. Technol Health Care 24(1):31–42
111. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI (2015) Machine learning applications in cancer prognosis and prediction. Comput Struct Biotechnol J 13:8–17
112. Shen L, Tan EC (2005) Dimension reduction-based penalized logistic regression for cancer classification using microarray data. IEEE/ACM Trans Comput Biol Bioinf 2(2):166–175
113. Wang Y, Tetko IV, Hall MA, Frank E, Facius A, Mayer KF, Mewes HW (2005) Gene selection from microarray data for cancer classification—a machine learning approach. Comput Biol Chem 29(1):37–46

114. Chu F, Xie W, Wang L (2004) Gene selection and cancer classification using a fuzzy neural network. IEEE Ann Meet Fuzzy Inf Process NAFIPS 2:555–559
115. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, Powell JI (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 403(6769):503–511
116. Khan J, Wei JS, Ringner M, Saal LH, Ladanyi M, Westermann F, Berthold F, Schwab M, Antonescu CR, Peterson C, Meltzer PS (2001) Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nature Med 7(6):673–679
117. Chen X, Cheung ST, So S, Fan ST, Barry C, Higgins J, Lai KM, Ji J, Dudoit S, Ng IO, Van De Rijn M (2002) Gene expression patterns in human liver cancers. Mol Biol Cell 13(6):1929–1939
118. Wang L, Chu F, Xie W (2007) Accurate cancer classification using expressions of very few genes. IEEE/ACM Trans Comput Biol Bioinf 4(1):40–53
119. Ramaswamy S, Tamayo P, Rifkin R, Mukherjee S, Yeang CH, Angelo M, Ladd C, Reich M, Latulippe E, Mesirov JP, Poggio T (2001) Multiclass cancer diagnosis using tumor gene expression signatures. Proc Natl Acad Sci 98(26):15149–15154
120. Cho SB, Won HH (2007) Cancer classification using ensemble of neural networks with multiple significant gene subsets. Appl Intell 26(3):243–250
121. Tan TZ, Quek C, Ng GS, Razvi K (2008) Ovarian cancer diagnosis with complementary learning fuzzy neural network. Artif Intell Med 43(3):207–222
122. Schummer M, Ng W, Bumgarner R, Nelson P, Schummer B, Bednarski D et al (1999) Comparative hybridization of an array of 21,500 ovarian cDNAs for the discovery genes overexpressed in ovarian carcinomas. Gene 238:375–385
123. Petricoin EF III, Ardekani AM, Hitt BA, Levine PJ, Fusaro VA, Steinberg SM et al (2002) Use of proteomic patterns in serum to identify ovarian cancer. Lancet 359(9306):572–577
124. Glaab E, Bacardit J, Garibaldi JM, Krasnogor N (2012) Using rule-based machine learning for candidate disease gene prioritization and sample classification of cancer gene expression data. PLoS ONE 7(7):e39932
125. Singh D, Febbo P, Ross K, Jackson D, Manola J et al (2002) Gene expression correlates of clinical prostate cancer behavior. Cancer Cell 1:203–209
126. Shipp M, Ross K, Tamayo P, Weng A, Kutok J et al (2002) Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. Nat Med 8:68–74
127. Chin S, Teschendorff A, Marioni J, Wang Y, Barbosa-Morais N et al (2007) High-resolution aCGH and expression profiling identifies a novel genomic subtype of ER negative breast cancer. Genome Biol 8:R215
128. Liu Q, Sung AH, Chen Z, Liu J, Chen L, Qiao M, Wang Z, Huang X, Deng Y (2011) Gene selection and classification for cancer microarray data based on machine learning and similarity measures. BMC Genom 12(S5):S1
129. Chen KH, Wang KJ, Wang KM, Angelia MA (2014) Applying particle swarm optimization-based decision tree classifier for cancer classification on gene expression data. Appl Soft Comput 24:773–780
130. Taiwan Cancer Registry, (2012), http://tcr.cph.ntu.edu.tw
131. Margoosian A, Abouei J (2013) Ensemble-based classifiers for cancer classification using human tumor microarray data. In: 2013 21st Iranian conference on electrical engineering (ICEE), IEEE, pp 1–6
132. Ramaswamy S et al (2002) Multiclass cancer diagnosis using tumor gene expression signatures. Proc Natl Acad Sci PNAS 98(26):15149–15154
133. Abdel-Zaher AM, Eldeib AM (2016) Breast cancer classification using deep belief networks. Expert Syst Appl 46:139–144
134. Dwivedi AK (2018) Artificial neural network model for effective cancer classification using microarray gene expression data. Neural Comput Appl 29(12):1545–1554
135. Sevakula RK, Singh V, Verma NK, Kumar C, Cui Y (2018) Transfer learning for molecular cancer classification using deep neural networks. IEEE ACM Trans Comput Biol Bioinf 16(6):2089–2100
136. Stiglic G, Kokol P (2010) Stability of ranked gene lists in large microarray analysis studies. J Biomed Biotechnol 2010:1–9
137. Ting FF, Tan YJ, Sim KS (2019) Convolutional neural network improvement for breast cancer classification. Expert Syst Appl 120:103–115
138. Mammographic Image Analysis Society (MIAS). (2018). http://www.mammoimage.org/databases/ Accessed: 25 January 2018
139. Ghoneim A, Muhammad G, Hossain MS (2020) Cervical cancer classification using convolutional neural networks and extreme learning machines. Future Gener Comput Syst 102:643–649
140. Yu L, Chen H, Dou Q, Qin J, Heng PA (2016) Automated melanoma recognition in dermoscopy images via very deep residual networks. IEEE Trans Med Imaging 36(4):994–1004
141. Gutman D, Codella NCF, Celebi E, Helba B, Marchetti M, Mishra N, Halpern A (2016) Skin lesion analysis toward melanoma detection: a challenge at the international symposium on biomedical imaging (ISBI) hosted by the International Skin Imaging Collaboration (ISIC), arXiv preprint arXiv:1605.01397
142. Albarqouni S, Baur C, Achilles F, Belagiannis V, Demirci S, Navab N (2016) Aggnet: deep learning from crowds for mitosis detection in breast cancer histology images. IEEE Trans Med Imaging 35(5):1313–1321
143. Von Ahn L (2006) Games with a purpose. Comput 39(6):92–94
144. Wang P, Wang L, Li Y, Song Q, Lv S, Hu X (2019) Automatic cell nuclei segmentation and classification of cervical Pap smear images. Biomed Signal Process Control 48:93–103
145. Zhang L, Lu L, Nogues I, Summers RM, Liu S, Yao J (2017) DeepPap: deep convolutional networks for cervical cell classification. IEEE J Biomed Health Inf 21(6):1633–1643
146. Kim Y, Zheng S, Tang J, Zheng WJ, Li Z, Jiang X. (2020) Anticancer Drug Synergy Prediction in Understudied Tissues using Transfer Learning. bioRxiv
147. Jiang P, Huang S, Fu Z, Sun Z, Lakowski TM, Hu P (2020) Deep graph embedding for prioritizing synergistic anticancer drug combinations. Comput Struct Biotechnol J 18:427–438
148. Ekşioğlu I, Tan M (2020) Prediction of Drug Synergy by Ensemble Learning. arXiv preprint arXiv:2001.01997
149. O'Neil J, Benita Y, Feldman I, Chenard M, Roberts B, Liu Y, Li J, Kral A, Lejnine S, Loboda A, Arthur W (2016) An unbiased oncology compound screen to identify novel combination strategies. Mol Cancer Ther 15(6):1155–1162
150. Zhang H, Feng J, Zeng A, Payne PR, Li F (2020) Predicting Tumor Cell Response to Synergistic Drug Combinations Using a Novel Simplified Deep Learning Model. bioRxiv
151. Kuru HI, Tastan O, Cicek AE (2020) MatchMaker: a deep learning framework for drug synergy prediction. bioRxiv
152. Preuer K, Lewis RP, Hochreiter S, Bender A, Bulusu KC, Klambauer G (2018) DeepSynergy: predicting anti-cancer drug synergy with Deep Learning. Bioinformatics 34(9):1538–1546
153. Wildenhain J, Spitzer M, Dolma S, Jarvik N, White R, Roy M, Griffiths E, Bellows DS, Wright GD, Tyers M (2015) Prediction of synergism from chemical-genetic interactions by machine learning. Cell Syst 1(6):383–395
154. Janizek JD, Celik S, Lee SI (2018) Explainable machine learning prediction of synergistic drug combinations for precision cancer medicine. bioRxiv, 1:331769
155. Mason DJ, Eastman RT, Lewis RP, Stott IP, Guha R, Bender A (2018) Using machine learning to predict synergistic antimalarial compound combinations with novel structures. Front Pharmacol 9:1096

156. Chen G, Tsoi A, Xu H, Zheng WJ (2018) Predict effective drug combination by deep belief network and ontology fingerprints. J Biomed Inf 85:149–154
157. Sharma A, Rani R (2018) An integrated framework for identification of effective and synergistic anti-cancer drug combinations. J Bioinf Comput Biol 16(05):1850017
158. Held MA, Langdon CG, Platt JT, Graham-Steed T, Liu Z, Chakraborty A, Bacchiocchi A, Koo A, Haskins JW, Bosenberg MW, Stern DF (2013) Genotype-selective combination therapies for melanoma identified by high throughput drug screening. Cancer Discov 3(1):52–67
159. Jang IS, Neto EC, Guinney J, Friend SH, Margolin AA (2014) Systematic assessment of analytical methods for drug sensitivity prediction from cancer cell line data. Biocomput 2014:63–74
160. Menden MP, Iorio F, Garnett M, McDermott U, Benes CH, Ballester PJ, Saez-Rodriguez J (2013) Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties. PLoS ONE 8(4):e61318
161. Turki T, Wei Z, Wang JT (2017) Transfer learning approaches to improve drug sensitivity prediction in multiple myeloma patients. IEEE Access 5:7381–7393
162. Wan Q, Pal R (2014) An ensemble based top performing approach for NCI-DREAM drug sensitivity prediction challenge. PLoS ONE 9(6):e101183
163. Dong Z, Zhang N, Li C, Wang H, Fang Y, Wang J, Zheng X (2015) Anticancer drug sensitivity prediction in cell lines from baseline gene expression through recursive feature selection. BMC Cancer 15(1):1–2
164. Rahman R, Matlock K, Ghosh S, Pal R (2017) Heterogeneity aware random forest for drug sensitivity prediction. Sci Rep 7(1):1–1
165. Yuan H, Paskov I, Paskov H, González AJ, Leslie CS (2016) Multitask learning improves prediction of cancer drug sensitivity. Sci Rep 6:31619
166. Ali M, Aittokallio T (2019) Machine learning and feature selection for drug response prediction in precision oncology applications. Biophys Rev 11(1):31–39
167. Haider S, Rahman R, Ghosh S, Pal R (2015) A copula based approach for design of multivariate random forests for drug sensitivity prediction. PLoS ONE 10(12):e0144490
168. He X, Folkman L, Borgwardt K (2018) Kernelized rank learning for personalized drug recommendation. Bioinformatics 34(16):2808–2816
169. Matlock K, De Niz C, Rahman R, Ghosh S, Pal R (2018) Investigation of model stacking for drug sensitivity prediction. BMC Bioinf 19(3):21–33
170. Riddick G, Song H, Ahn S, Walling J, Borges-Rivera D, Zhang W, Fine HA (2011) Predicting in vitro drug sensitivity using Random Forests. Bioinformatics 27(2):220–224
171. Sharma A, Rani R (2019) Drug sensitivity prediction framework using ensemble and multi-task learning. Int J Mach Learn Cybern 11:1231–1240
172. Sharma A, Rani R (2019) Ensembled machine learning framework for drug sensitivity prediction. IET Syst Biol 14(1):39–46
173. Ezzat A, Wu M, Li XL, Kwoh CK (2017) Drug-target interaction prediction using ensemble learning and dimensionality reduction. Methods 129:81–88
174. Ezzat A, Wu M, Li XL, Kwoh CK (2016) Drug-target interaction prediction via class imbalance-aware ensemble learning. BMC Bioinf 17(19):267–276
175. Tabei Y, Pauwels E, Stoven V, Takemoto K, Yamanishi Y (2012) Identification of chemogenomic features from drug–target interaction networks using interpretable classifiers. Bioinformatics 28(18):i487–i494
176. Chen R, Liu X, Jin S, Lin J, Liu J (2018) Machine learning for drug-target interaction prediction. Molecules 23(9):2208
177. Wen M, Zhang Z, Niu S, Sha H, Yang R, Yun Y, Lu H (2017) Deep-learning-based drug–target interaction prediction. J Proteome Res 16(4):1401–1409
178. Yuan Q, Gao J, Wu D, Zhang S, Mamitsuka H, Zhu S (2016) DrugE-Rank: improving drug–target interaction prediction of new candidate drugs or targets by ensemble learning to rank. Bioinformatics 32(12):i18-27
179. Law V, Knox C, Djoumbou Y, Jewison T. An Chi Guo, Yifeng Liu, Adam Maciejewski, David Arndt, Michael Wilson, Vanessa Neveu, and others (2014) DrugBank 4.0: shedding new light on drug metabolism. Nucleic Acids Res,42:D1
180. Zhang J, Zhu M, Chen P, Wang B (2017) Drugrpe: Random projection ensemble approach to drug-target interaction prediction. Neurocomp 228:256–262
181. He Z, Zhang J, Shi XH, Hu LL, Kong X, Cai YD, Chou KC (2010) Predicting drug-target interaction networks based on functional groups and biological features. PLoS ONE 5(3):e9603
182. Xie L, He S, Song X, Bo X, Zhang Z (2018) Deep learning-based transcriptome data classification for drug-target interaction prediction. BMC Genom 19(7):667
183. Tian K, Shao M, Wang Y, Guan J, Zhou S (2016) Boosting compound-protein interaction prediction by deep learning. Methods 110:64–72
184. Kuhn M, Szklarczyk D, Pletscher-Frankild S, Blicher TH, Von Mering C, Jensen LJ, Bork P (2014) STITCH 4: integration of protein–chemical interactions with user data. Nucleic Acids Res 42(D1):D401–D407
185. Wang J, Archambault B, Xu Y, Taleyarkhan RP (2010) Numerical simulation and experimental study on Resonant Acoustic Chambers—For novel, high-efficiency nuclear particle detectors. Nucl Eng Des 240(11):3716–3726
186. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, Sonnhammer EL (2014) Pfam: the protein families database. Nucl Eng Des 42(D1):D222–D230
187. Feng Q, Dueva E, Cherkasov A, Ester M (2018) Padme: a deep learning-based framework for drug-target interaction prediction. arXiv preprint arXiv:1807.09741
188. He T, Heidemeyer M, Ban F, Cherkasov A, Ester M (2017) Simboost: A readacross approach for predicting drug–target binding affinities using gradient boosting machines. J Cheminf 9(1):24
189. Davis MI, Hunt JP, Herrgard S, Ciceri P, Wodicka LM, Pallares G, Hocker M, Treiber DK, Zarrinkar PP (2011) Comprehensive analysis of kinase inhibitor selectivity. Nat Biotechnol 29(11):1046–1051
190. Metz JT, Johnson EF, Soni NB, Merta PJ, Kifle L, Hajduk PJ (2011) Navigating the kinome. Nat Chem Biol 7(4):200
191. Tang J, Szwajda A, Shakyawar S, Xu T, Hintsanen P, Wennerberg K, Aittokallio T (2014) Making sense of large-scale kinase inhibitor bioactivity data sets: A comparative and integrative analysis. J Chem Inf Model 54(3):735–743
192. Xie L, Zhang Z, He S, Bo X, Song X (2017) Drug—target interaction prediction with a deep-learning-based model. In: 2017 IEEE international conference on bioinformatics and biomedicine (BIBM), pp 469–476
193. Sharma A, Rani R (2018) BE-DTI': Ensemble framework for drug target interaction prediction using dimensionality reduction and active learning. Comput Methods Programs Biomed 165:151–162