Machine Learning Applications in Anti-cancer Drug Discovery

Aman Sharma and Rinkle Rani

1 Introduction

Traditional drug discovery pipelines are complex and inefficient. Worldwide various researchers are trying to shorten the drug discovery cycle to fight against many stringent diseases. Traditionally researchers were only using statistical and clinical methods for drug discovery [1]. Recently, researchers are using computational approaches to reduce drug discovery time [2, 3]. Machine learning approaches provide a powerful set of tools that can help in developing decision support systems. Such systems help in early prognosis and diagnosis of diseases such as cancer, diabetes, Parkinson's, etc. Their applications include drug response prediction, drug-target identification, biomarker identification, and drug synergy prediction. Cancer is considered a very stringent and complex disease worldwide. Scientists/researchers are trying their hard to find potential drugs or drug combinations that could help to fight against diseases such as cancer which is the leading cause of death worldwide [2, 3]. Figure 1 shows the central dogma of biology.

According to the study [4] the new drugs are produced at a constant rate during the past 60 years. Moreover, the Tufts Centre for the Study of Drug Development (CSDD) reported that the overall cost involved in developing a newly approved drug is about $2,558 million and time is about one decade. Such study focuses on the attention of researchers to develop computationally efficient drug discovery pipelines. Machine learning and deep learning methods came up as a breakthrough

A. Sharma (✉)
JUIT, Wakhnaghat, Solan, India

R. Rani
T.I.E.T, Patiala, India
e-mail: raggarwal@thapar.edu

S. Bhatia et al. (eds.), *Intelligent Healthcare*, EAI/Springer Innovations in Communication and Computing, https://doi.org/10.1007/978-3-030-67051-1_10

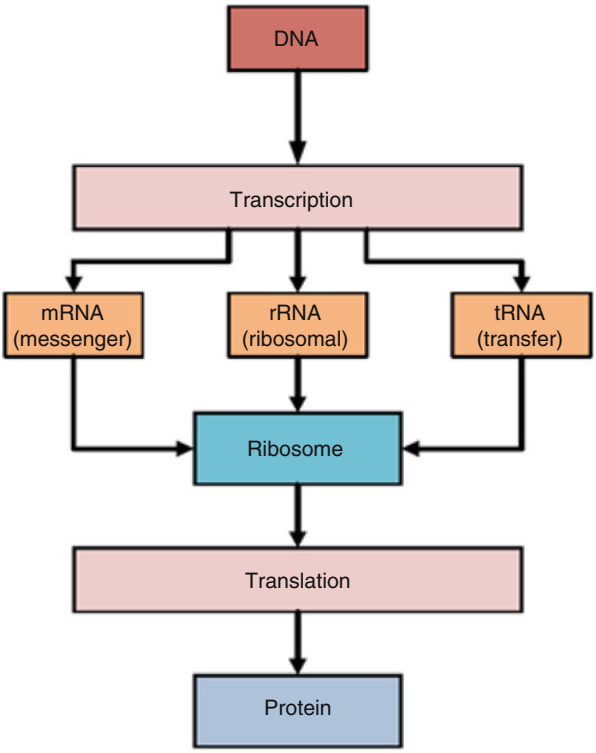

Fig. 1 Central dogma of biology [82]

in drug discovery research [5, 6]. Deep learning applications involve various hidden layers which provide data abstraction and help in data preprocessing and feature extraction [4].

All the machine learning approaches are data-driven and hence help in developing predictive modeling by utilizing the hidden correlations and patterns in data. Figure 2 shows the genome data representation for machine learning approaches. Machine learning approaches can be categorized as reinforcement, supervised, unsupervised, and semi-supervised learning. The major difference in these approaches is the quantity of information that is fed into the model which lays the basis for model training. Chemical researchers have extensively utilized machine learning capabilities especially supervised learning in anti-cancer research [7]. Supervised machine learning algorithms use target labels for training the input data and approximately predicting the output. Artificial neural networks (ANNs) and kernel methods are popular supervised learning algorithms used to transform the input space into a new feature space [8]. One of the most important features of ANNs is feature transformation using various input layers. On the other, hand kernel methods help in identifying non-linear relationships present within the data. Kernel methods utilize kernel function to perform non-linear data transformations

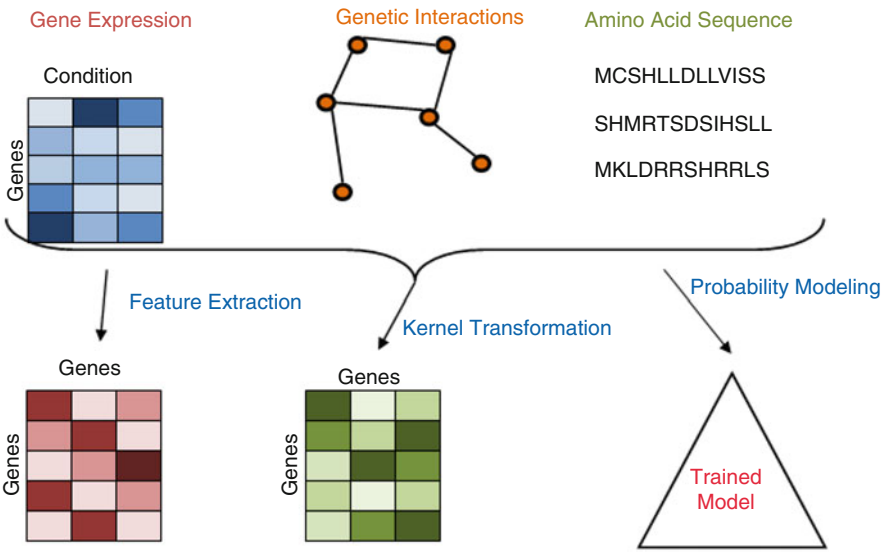

Fig. 2 Genome data representation for machine learning approaches

to use with linear algorithms. There are widespread applications of ANNs and deep learning (DL) algorithms in biomedical and drug discovery. Gene expression data and microarray data are used for developing anti-cancer drugs and biomarker prediction machine learning models.

The influence of genes in different types of cancer promoted the genetic data-driven research. Complex microenvironment results in difficult diagnosis and treatment of various cancer types. Even patients with similar type of tumor show varying responses toward the same type of drug treatment [9]. Although traditional machine learning algorithms [10] are quite helpful in developing biomedical computational models, recently we have seen a rise in deep learning algorithms. The major reason for such a sudden drift is because of the large availability of biomedical and pharmacogenomic datasets [11, 12] and high computing machines for parallel processing such as GPUs.

2 Background

In this section, we discussed the machine learning-based cancer applications. The essence of all the machine learning models is the high-quality data that we feed for training. From the last two decades, many researchers and consortiums have contributed to the field of drug discovery by providing high-quality chemical and biological data. PubChem [13] is one of the largest open-source chemical repositories. It provides the facility to search chemicals by name or structure. We

Table 1 Selected startup companies in the field of drug discovery

S. no.	Company name	Website
1.	Amplion	https://bit.ly/2PjOY94
2.	BioSymetrics	https://bit.ly/2Xk14U5
3.	Biorelate	https://bit.ly/3fpvdY2
4.	Causaly	https://bit.ly/2Xle1gf
5.	Data2Discovery	https://bit.ly/31aEkGX
6.	Data4Cure	https://bit.ly/3fgJ5Uw
7.	Elucidata Corporation	https://bit.ly/3gpqzdS
8.	Evid Science	https://bit.ly/3i0gCnC
9.	Genialis	https://bit.ly/39SnqB8
10.	HelixAI	https://bit.ly/3glrz2V
11.	Innoplexus	https://bit.ly/3fjG8CM
12.	Intellegens	https://bit.ly/3k48rZq
13.	InveniAI	https://bit.ly/30pX9XP
14.	Mozi	https://bit.ly/3190p8Z
15.	PatSnap	https://bit.ly/33dLEV4

can get the physical, biological, chemical, and toxicity data of various chemical compounds. ChEMBL [14] is a collection of bioactive drug-like small molecules.

It contains data corresponding to 2D structure and properties of bioactive drugs. The database is majorly curated and abstracted from the literature of modern drug discovery. The data for bioactivity of the drug molecule is provided in normalized form. Further, web links for the research studies corresponding to drugs are included in the database. The DrugBank [15] is the freely available database consisting of data about a wide range of drugs and their corresponding targets. The DrugBank combines data for two research domains: bioinformatics and chemoinformatics. It is like an encyclopedia for getting information and detailed description regarding various chemical compounds and their corresponding targets. It is a widely adopted resource by various pharmacists, physicians, researchers, and the drug industry. Table 1 contains selected startup companies in the field of drug discovery.

DrugCentral [16] is an online repository for information on various drugs. It contains information such as mode of action for drugs and active ingredients in chemical products. It also contains information regarding discontinued and approved drugs outside the USA. SuperDRUG2 [17] is one of the largest databases consisting of approved/marketed drugs and chemical ingredients. 2D and 3D structures, physicochemical properties, and pharmacokinetic data of drugs are also provided in the database. Along with this, it contains data for drug-drug and drug-target interactions. The GDSC [18] database is developed to improve cancer biomarker prediction and drug-target prediction. Informative data is provided corresponding to genomic variations when different cell lines are perturbed with drugs. The CCLE [12] database is a result of a collaborative effort by various drug discovery research labs. It contains 1870 RNA sequencing, 654 whole exome sequences, and 46 whole genome sequence files. Various researchers are trying their

Table 2 Selected publicly available online databases for drug discovery

S. no.	Database	Online access
1.	PubChem [13]	https://bit.ly/39MxdZc
2.	ChEMBL [14]	https://bit.ly/3k6kvcs
3.	DrugBank [15]	https://bit.ly/3hRKPFw
4.	DrugCentral [16]	https://bit.ly/2PfMTL0
5.	SuperDRUG2 [17]	https://bit.ly/2EJHdXV
6.	GDSC [18]	https://bit.ly/31bxIbp
7.	CCLE[12]	https://bit.ly/3fkagOf
8.	repoDB [19]	https://bit.ly/33oSHdL

hard to extract meaningful insights from CCLE using microarray data analysis. Table 2 contains selected publicly available online databases for drug discovery.

Drug Repurposing

The availability of freely downloadable healthcare datasets motivated the researchers to apply machine learning algorithms for predicting drug responses, biomarkers, signaling pathways, drug synergism, etc. Figure 3 describes the Gaussian kernel and multi-task learning used for anti-cancer drug response prediction. The Bayesian model has been used by Sean Ekins et al. [20] for compound selection. In the proposed technique, they have used bioinformatics as well as chemoinformatics data. Various researchers have also used machine learning models for ligand-based virtual screening (LBVS) [21]. Naive Bayes algorithm is also prominently used for predicting toxicity and biological pathways for anti-cancer drug prediction [22]. Kuang Z et al. [23] have presented a regularization-based technique for drug repurposing. Their statistical analysis suggests various drugs that can be repurposed in varying biological situations. Patrick MT et al. [24] have implemented an approach for summarizing drug information from 20 million research articles. They trained their model on various stringent diseases such as psoriasis, alopecia areata, and immune-mediated diseases to obtain the drug repurposing opportunities. Zeng X et al. [25] have proposed a deep learning approach for computational drug repurposing. Their proposed approach exploits data from various networks such as drug-disease, drug-target, and drug-drug networks. The proposed model was trained in Alzheimer's and Parkinson's disease. Kim E et al. [26] have developed the machine learning-based approach for predicting the hidden pharmacological benefits of herbal compounds. The common assumption that all the researchers assume while developing computational approaches for drug repurposing is that similar diseases can be treated with similar drugs. However, similarity can be defined in terms of drug-drug, tissue-tissue, and disease-disease similarity. Table 3 contains selected publicly available online databases for drug repurposing.

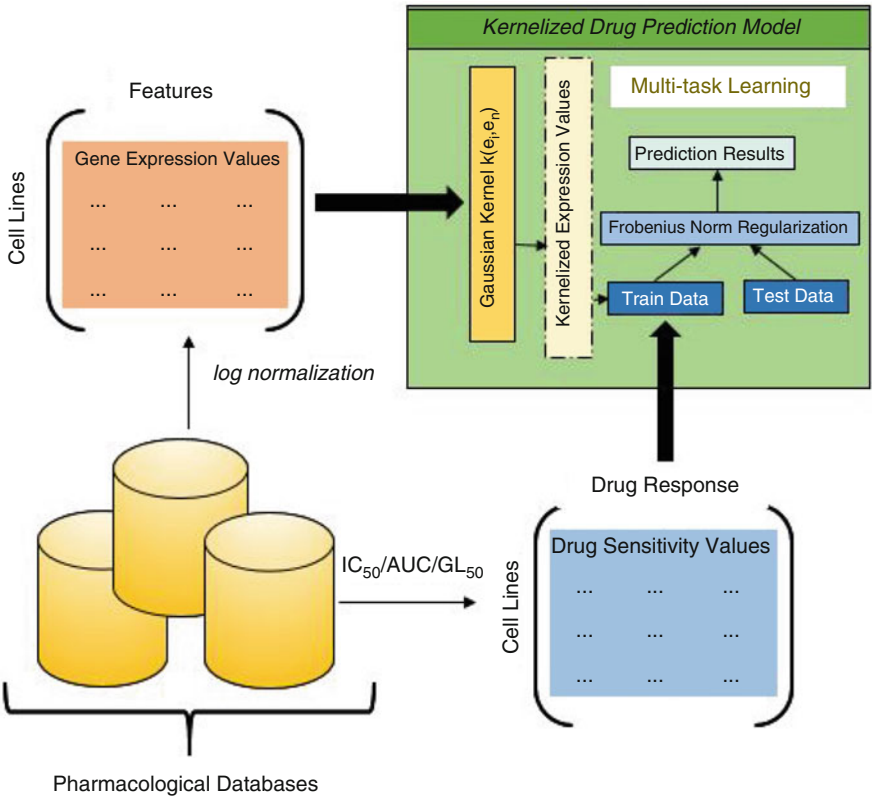

Fig. 3 Gaussian kernel and multi-task learning used for anti-cancer drug response prediction

Table 3 Selected publicly available online databases for drug repurposing

S. no.	Database	Online access
1.	NCI-DREAM 7	https://bit.ly/2Pfj4Kz
2.	NCI-60 [27]	https://bit.ly/33lCxBE
3.	TCGA [28]	https://bit.ly/39QPapq
4.	TCPA [29]	https://bit.ly/33hI4cs
5.	GDSC[18]	https://bit.ly/33jtyAU
6.	CCLE [12]	https://bit.ly/2DrT1NI

Cancer Classification

Gene selection is a challenging process in microarray data analysis. Although a lot of research has been done on identifying genomic biomarkers for different types of cancer, still no generic pipeline has been designed for cancer classification. Various algorithms/approaches have been proposed in the literature to identify relevant and potential genomic biomarkers. These algorithms can be broadly classified as wrapper, filter, and hybrid methods for gene selection [30, 31]. The filter method

is defined by statistical analysis and properties of the dataset for obtaining the best optimal gene subset. Ranking of genes is performed using different types of statistical methods [32, 33]. Genes that score relatively higher rank are considered for further analysis. The methods included in this category are T-test [33], max-min correntropy [34], and information gain [35]. For detailed information on such methods, one can consider a survey on filter methods for gene selection [36]. The wrapper method relies on some kind of evolutionary technique to optimally search the relevant feature subset. In wrapper methods, random initialization of the population is done consisting of the subset of features. The fitness of each subset is obtained using an appropriate fitness evaluator. Iteratively the whole process is repeated several times to fetch the optimal solution. Such methods include the use of a genetic algorithm [37], artificial bee colony algorithm [38], bat algorithm [39], and swarm optimization [40] for gene selection. Table 4 contains selected publicly available online datasets for cancer classification (Fig. 4).

Hybrid methods are also evolutionary-based methods, but they use filter methods in the initial phase for the screening of the most promising genes from the

Table 4 Selected publicly available online datasets for cancer classification

S. no.	Datasets	Online access
1.	SRBCT cancer	Khan et al. [41]
2.	Leukemia cancer	Golub et al. [42]
3.	Prostate cancer	Singh et al. [43]
4.	Breast cancer	Hedenfalk et al. [44]
5.	Breast cancer	https://bit.ly/2XmvQLM
6.	Central nervous system cancer	https://bit.ly/33lCPse
7.	GSE25136	https://bit.ly/2PjA2HL

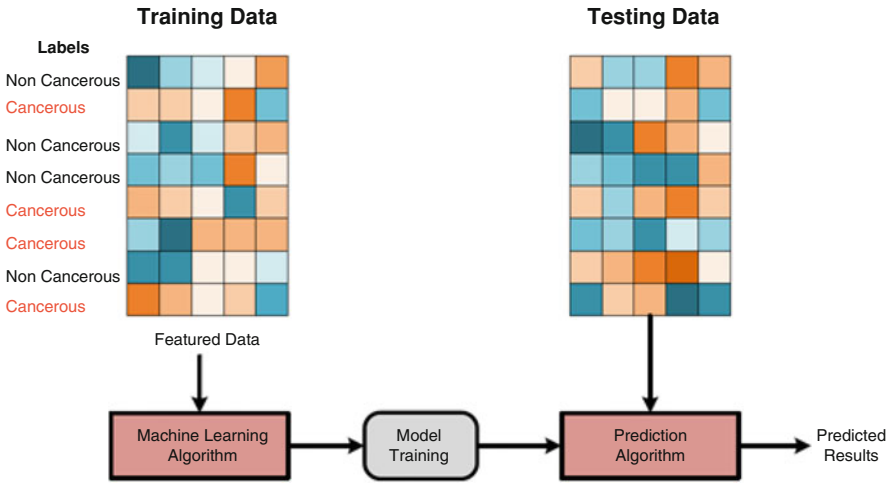

Fig. 4 Example of cancer classification using machine learning application [82]

microarray dataset. These methods use filter methods to reduce the running time of an algorithm. Some of these approaches include the chi-square test with GA [45], mRMR with GA [46], and similarity scheme with ABC [47]. Apart from these methods, some approaches integrate the feature selection task along with classification. It selects the feature subset, builds a classifier, and then checks the classifier accuracy. If performance is not appropriate, then it removes the poor genes and builds the classifier again iteratively. Such methods are known as embedded methods [48]. Most of these methods are difficult to replicate and are computationally expensive. Most of these methods are not able to optimally utilize high-dimensional gene expression data and suffer from overfitting.

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) [49]. Any alteration and changes in these genes lead to uncontrollable cell growth. Unlike these genes, some oncogenes 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, which hence results in nonresponsive drug behavior [50, 51]. 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 [42]. Combination drug therapy is a good option to avoid drug resistance. It helps in overcoming the drug resistance by delaying tumor growth. It includes the usage of multiple drugs in fixed dose proportion and as a single-dose formulation. Combination drug therapy is showing excellent results in tumor suppression by reducing the chances of multiple mutations [52] and a single mutation [53] that can escape all the drugs. Additionally, combination therapy helps in lowering drug dosage and side effects [52]. 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 proportions.

Combination drug therapy is widely used in treatment of various diseases such as HIV and cancer and many more diseases [54, 55]. Combination drug therapy becomes more essential for complex diseases such as cancer because of the involvement of multiple growth pathways in such diseases [54]. However, there is risk of toxicity with combination therapies, which can be handled with appropriate quantity of dosage. Many combination drugs are already approved by FDA for treating stringent diseases. For example, aspirin and dipyridamole are used in combination to reduce the risk of stroke [56], and sabarubicin and cisplatin are used in lung cancer [57]. The Drug Combination Database (DCDB) provides

information about 330 FDA-approved and 1033 investigational drug combinations [58]. Drug combination works on the principle of synergy effect, which means that the overall effect of combination drug therapy is more as compared to individual drugs. Although it is observed that there is benefit of drug combination therapy, quantification of drug synergy effect is still one of the challenging tasks. Many researchers have proposed different methods such as Chou-Talalay method [59], Loewe additivity [60], and Bliss independence [61] to calculate dose-response effect of combination drugs. These techniques are based on comparison between expected and observed combination drug responses [62, 63]. Till now, for most of the diseases, drug combinations are identified based on clinical trials. But clinical trials using "trial and error" is a labor- and cost-intensive and time-consuming task [64]. Another disadvantage of clinical trials is unwanted exposure of harmful chemicals to patients [65]. Apart from clinical trials, high-throughput screening (HTS) is also used in identifying potential drug combinations [66]. In such screenings, different concentrations of drugs are used to identify potential drug combinations, but still they are not accurate enough to capture the real microenvironment [67].

Although all these methods play a crucial role in quantifying drug synergy, still there are many issues with these methods such as no method can quantify drug synergism in different feasible situations [62] and different dose-response methods can even produce different results [68]. Moreover, these methods are based on screening of all the possible drug combinations, which is impractical and time-consuming. In such a situation, identification of effective drug combinations is a challenging task. Many researchers are using machine learning and computational models to predict potential drug combinations for various diseases [69, 70].

3 Research Gaps in Computational Drug Discovery

The last two decades has seen a tremendous awareness and growth toward cancer research. Many researchers, clinicians, and academicians are trying their hard to fight against cancer. We have discussed the various already proposed computational drug discovery approaches and applications in Sect. 2. But still, there are many issues/research gaps left that need to be worked upon. Most of the techniques are crudely based on statistics which limit the utility of the applications only to statisticians. Hence, there is a need to develop user-friendly applications, which are provided by machine learning algorithms. The following research gaps still exist:

1. High dimensionality of data is one of the major issues while developing applications from genomic data. Moreover, there is an issue of imbalance class; there is a majority of one class due to lack of samples of other class. Although various techniques have been developed, still no approach has been developed which covers a wide range of applications. Some techniques are good in one situation and others in different situations. So it is more or like hit and trial method.

2. Existing approaches are developed using binary imbalanced datasets. There is a need to test those applications on multi-class imbalanced datasets. With the increase in the severity of the genetic disease, their subtypes also increase. Hence multi-class classification is required to predict the correct subtype of the disease.
3. Heterogeneity in the genetic structure of cancer patients results in heterogeneity in their drug responses. Earlier drugs were discovered based on the anatomical region of the disease, but cancer is a genetic disease, so any anti-cancer drug discovery needs to consider genetic influence while developing new drugs. Moreover, if any computational method is proposed for drug discovery or anti-cancer drug prediction, then it should strictly consider the genetic variations that are responsible for cancer.
4. Feature selection is one of the primary tasks in cancer classification approaches. But existing feature selection approaches are not scalable enough to handle maximum genetic aberrations simultaneously.
5. Cancer is a complex disease; we cannot generalize the drug therapies for different patients. There is a need to provide personalized therapies corresponding to an individual patient's drug sensitivity.
6. Machine learning capabilities for predicting drug synergism are unexplored. Predicting drug synergy will boost the anti-cancer drug discovery process. There is a need to extract better features for predicting drug synergy.

4 Future of Computational Drug Discovery

Deep Learning for Drug Discovery

"Artificial intelligence" as the name itself states that it is a kind of intelligence which is incorporated artificially in a system. There are various definitions of artificial intelligence, but broadly they are categorized as thinking humanly or rationally and acting humanly or rationally. It is an interdisciplinary branch of science which can be applied into molecular biology and genomics and in various other disciplines. Researchers are actively using artificial intelligence in bioinformatics for analyzing large amount of data and DNA sequencing [71]. From the last two decades, there is an enormous increase in healthcare data from researchers, academicians, and industry. This data holds enough potential to explore and fetch meaningful insights. There is a huge possibility of exploring hidden patterns and knowledge from this healthcare data using computational approaches. Although many researchers are using statistical approaches and machine learning algorithms to get meaningful insights from data, it is impossible to process healthcare datasets using conventional machine learning algorithms because of their huge volume, velocity, and variety. In such a scenario, deep learning algorithms play a very important role. Nowadays, researchers are using deep learning algorithms such as CNN [72] and RNN [73] for dealing with healthcare big datasets.

Role of Deep Learning in Cancer Classification

Existing literature on cancer classification has used traditional machine learning algorithms. However, very few approaches have been proposed using deep learning. Wang et al. [74] developed a deep learning-based technique to identify metastatic breast cancer using an image dataset. Ahmed et al. [75] have proposed a deep belief network-based for breast cancer classification. Skin cancer is very common nowadays, and it's hard to diagnose and predict their targets. To classify and identify the most promising biomarkers for skin cancer, Haofu et al. [76] proposed a classification approach using deep learning. Fakoor et al. [77] have developed an unsupervised feature selection technique for identifying and diagnosing cancer types. Arunkumar and Ramakrishnan [78] have developed the hybrid approach for feature selection. All these techniques have focused on reducing the dimensions of input dataset using feature selection approaches.

Role of Deep Learning in MicroRNA Analysis in NGS

"Big Data" has been a buzz topic in recent years, and 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 have led to the sudden outburst of biomedical data. This data serves a great potential in identifying disease biomarkers and discovering new drugs, but unfortunately, it is not effectively utilized. NGS technologies have created a serious need for new technologies and algorithms. Figure 5 shows biogenesis of microRNA. 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 the 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. [79] have proposed deepMiRGene, an algorithm used to predict microRNA precursor. They used RNN, because 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 [80], 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

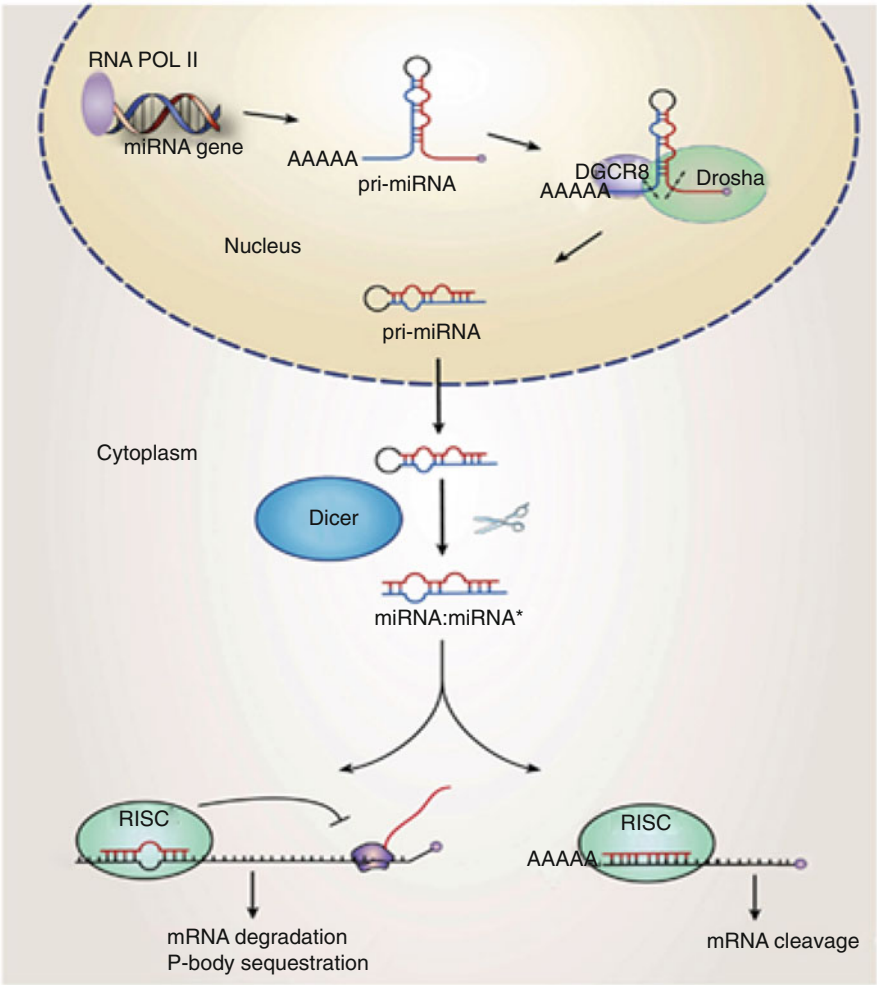

Fig. 5 Biogenesis of microRNA (Image from [81])

knowledge about the underlying mechanism. Table 5 contains selected microRNAs as potential cancer diagnostic biomarkers in blood.

5 Conclusion and Future Directions

We have seen in previous sections that various machine learning applications have been developed in the literature for anti-cancer drug discovery. But still it is a challenge to predict drugs using computational techniques which are also clinically

Table 5 Selected microRNAs as potential cancer diagnostic biomarkers in blood

S. no.	Tool/pipeline	Features	URL
1.	BioVLAB-MMIA-NGS	To find DE microRNAs and their target genes (DEGs)	https://bit.ly/3gzgN9r
2.	CAP-miRSeq	Supports sequential and parallel processing of deep sequencing microRNA data	https://mayocl.in/2DcM3fF
3.	iMir	Provides automated pipeline for microRNA data analysis	https://bit.ly/30iOHJM
4.	CPSS	Standalone tool with single data submission	https://bit.ly/30lOG7H
5.	MAGI	MicroRNA-Seq analysis using GPU technology	https://bit.ly/39LByfm
6.	miRSeqNovel	R/bioconductor pipeline package to predict novel microRNA for plant and animal microRNA	https://bit.ly/2DbmQ5s
7.	mirTools 2.0	Performs comparative analysis of experimental samples and identifies the DE microRNAs among experimental group	https://bit.ly/3flotKV
8.	MMIA	Integrates microRNA and mRNA expression data for detailed analysis	https://bit.ly/39VcJh1

efficient. Cancer is a very stringent and complex disease which needs multi-focused approach for treatment. We can't treat a patient by focusing on a single aspect of genetic behavior of individual. Multiple pathways need to be considered while developing potential treatment drugs. Drug resistance can also be targeted while designing drugs. In such cases, multiple drug combinations can be selected as a treatment option. Enormous increase in oncological datasets is also a boom to cancer research. But this data can't be mined using traditional/conventional machine learning algorithms. Deep learning algorithms should be extensively studied to utilize such big healthcare datasets.

References

1. S. Wold, L. Eriksson, S. Clementi, Statistical validation of QSAR results. Chemometric Methods in Molecular Design (Wiley-VCH, Weinheim, 1995), pp. 309–338
2. A. Sharma, R. Rani, Drug sensitivity prediction framework using ensemble and multi-task learning. Int. J. Mach. Learn. Cybern. **11**, 1231–1240 (2019)
3. A. Sharma, R. Rani, An optimized framework for cancer classification using deep learning and genetic algorithm. J. Med. Imaging Health Inform. **7**(8), 1851–1856 (2017)

4. B. Munos, Lessons from 60 years of pharmaceutical innovation. Nat. Rev. Drug Discov. **8**(12), 959–968 (2009)
5. Y. LeCun, Y. Bengio, G. Hinton, Deep learning. nature, 521(7553), 436–444 (2015)
6. N. Stephenson, E. Shane, J. Chase, J. Rowland, D. Ries, N. Justice, ... R. Cao, Survey of machine learning techniques in drug discovery. Curr. Drug Metab. **20**(3), 185–193 (2019)
7. K. Alberi, M.B. Nardelli, A. Zakutayev, L. Mitas, S. Curtarolo, A. Jain, ... M. Kanatzidis, The 2019 materials by design roadmap. J. Phys. D. Appl. Phys. **52**(1), 013001 (2018)
8. M. Rupp, Von O.A. Lilienfeld, K. Burke, Guest editorial: Special topic on data-enabled theoretical chemistry (2018)
9. J. Sheng, F. Li, S.T. Wong, Optimal drug prediction from personal genomics profiles. IEEE J. Biomed. Health Inform. **19**(4), 1264–1270 (2015)
10. M.W. Libbrecht, W.S. Noble, Machine learning applications in genetics and genomics. Nat. Rev. Genet. **16**(6), 321–332 (2015)
11. W. Yang, J. Soares, P. Greninger, E.J. Edelman, H. Lightfoot, S. Forbes, ... S. Ramaswamy, Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. Nucleic Acids Res. **41**(D1), D955D961 (2012)
12. J. Barretina, G. Caponigro, N. Stransky, K. Venkatesan, A.A. Margolin, S. Kim, ... A. Reddy, The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature **483**(7391), 603–607 (2012)
13. E.E. Bolton, Y. Wang, P.A. Thiessen, S.H. Bryant, PubChem: integrated platform of small molecules and biological activities. In Annual Reports In Computational Chemistry, vol. 4 (Elsevier, San Diego, 2008), pp. 217–241
14. A. Gaulton, L.J. Bellis, A.P. Bento, J. Chambers, M. Davies, A. Hersey, ... J.P. Overington, ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res. **40**(D1), D1100–D1107 (2012)
15. D.S. Wishart, C. Knox, A.C. Guo, D. Cheng, S. Shrivastava, D. Tzur, ... M. Hassanali, DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic Acids Res. **36**(suppl_1), D901–D906 (2008)
16. O. Ursu, J. Holmes, J. Knockel, C.G. Bologa, J.J. Yang, S.L. Mathias, ... T.I. Oprea, DrugCentral: online drug compendium. Nucleic Acids Res. **45**, D932–D939 (2016)
17. A. Goede, M. Dunkel, N. Mester, C. Frommel, R. Preissner, SuperDrug: a conformational drug database. Bioinformatics **21**(9), 1751–1753 (2005)
18. M.J. Garnett, E.J. Edelman, S.J. Heidorn, C.D. Greenman, A. Dastur, K.W. Lau, ... Q. Liu, Systematic identification of genomic markers of drug sensitivity in cancer cells. Nature **483**(7391), 570–575 (2012)
19. A.S. Brown, C.J. Patel, A standard database for drug repositioning. Sci. Data **4**(1), 1–7 (2017)
20. S. Ekins, P.B. Madrid, M. Sarker, S.G. Li, N. Mittal, P. Kumar, ... P. Bourbon, Combining metabolite-based pharmacophores with bayesian machine learning models for Mycobacterium tuberculosis drug discovery. PloS one **10**(10), e0141076 (2015)
21. A. Lavecchia, Machine-learning approaches in drug discovery: methods and applications. Drug Discov. Today **20**(3), 318–331 (2015)
22. H. Sun, A naive Bayes classifier for prediction of multidrug resistance reversal activity on the basis of atom typing. J. Med. Chem. **48**(12), 4031–4039 (2005)
23. Z. Kuang, Y. Bao, J. Thomson, M. Caldwell, P. Peissig, R. Stewart, ... D. Page, A machine-learning-based drug repurposing approach using baseline regularization, in *Computational Methods for Drug Repurposing* (Humana Press, New York, 2019), pp. 255–267
24. M.T. Patrick, K. Raja, K. Miller, J. Sotzen, J.E. Gudjonsson, J.T. Elder, L.C. Tsoi, Drug repurposing prediction for immune-mediated cutaneous diseases using a word-embedding–based machine learning approach. J. Investig. Dermatol. **139**(3), 683–691 (2019)
25. X. Zeng, S. Zhu, X. Liu, Y. Zhou, R. Nussinov, F. Cheng, deepDR: a network-based deep learning approach to in silico drug repositioning. Bioinformatics **35**(24), 5191–5198 (2019)
26. E. Kim, A.S. Choi, H. Nam, Drug repositioning of herbal compounds via a machine-learning approach. BMC Bioinf. **20**(10), 33–43 (2019)

27. R.H. Shoemaker, The NCI60 human tumour cell line anticancer drug screen. Nat. Rev. Cancer **6**(10), 813–823 (2006)
28. K. Tomczak, P. Czerwińska, M. Wiznerowicz, The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemp. Oncol. **19**(1A), A68 (2015)
29. J. Li, Y. Lu, R. Akbani, Z. Ju, P.L. Roebuck, W. Liu, ... C. Wakefield, TCPA: a resource for cancer functional proteomics data. Nat. Methods **10**(11), 1046–1047 (2013)
30. M.W. Khan, M. Alam, A survey of application: Genomics and genetic programming, a new frontier. Genomics **100**(2), 65–71 (2012)
31. Y. Saeys, I. Inza, P. Larrañaga, A review of feature selection techniques in bioinformatics. Bioinformatics **23**(19), 2507–2517 (2007)
32. V. Pihur, S. Datta, S. Datta, Finding common genes in multiple cancer types through meta–analysis of microarray experiments: a rank aggregation approach. Genomics **92**(6), 400–403 (2008)
33. Y. Qi, H. Sun, Q. Sun, L. Pan, Ranking analysis for identifying differentially expressed genes. Genomics **97**(5), 326–329 (2011)
34. N. Zhou, L. Wang, A modified T-test feature selection method and its application on the HapMap genotype data. Genomics Proteomics Bioinf. **5**(3–4), 242–249 (2007)
35. M. Mohammadi, H.S. Noghabi, G.A. Hodtani, H.R. Mashhadi, Robust and stable gene selection via maximum–minimum correntropy criterion. Genomics **107**(2–3), 83–87 (2016)
36. C. Lazar, J. Taminau, S. Meganck, D. Steenhoff, A. Coletta, C. Molter, ... A. Nowe, A survey on filter techniques for feature selection in gene expression microarray analysis. IEEE/ACM Trans. Comput. Biol. Bioinform. **9**(4), 1106–1119 (2012)
37. Y. Wang, X. Chen, W. Jiang, L. Li, W. Li, L. Yang, ... S. Wang, Predicting human microRNA precursors based on an optimized feature subset generated by GA–SVM. Genomics **98**(2), 73–78 (2011)
38. B.A. Garro, K. Rodríguez, R.A. Vázquez, Classification of DNA microarrays using artificial neural networks and ABC algorithm. Appl. Soft Comput. **38**, 548–560 (2016)
39. M. Dashtban, M. Balafar, P. Suravajhala, Gene selection for tumor classification using a novel bio-inspired multi-objective approach. Genomics **110**(1), 10–17 (2018)
40. S. Kar, K.D. Sharma, M. Maitra, Gene selection from microarray gene expression data for classification of cancer subgroups employing PSO and adaptive K-nearest neighborhood technique. Expert Syst. Appl. **42**(1), 612–627 (2015)
41. J. Khan, J.S. Wei, M. Ringner, L.H. Saal, M. Ladanyi, F. Westermann, ... P.S. Meltzer, Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat. Med. **7**(6), 673–679 (2001)
42. T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, ... C.D. Bloomfield, Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science **286**(5439), 531–537 (1999)
43. D. Singh, P.G. Febbo, K. Ross, D.G. Jackson, J. Manola, C. Ladd, ... E.S. Lander, Gene expression correlates of clinical prostate cancer behavior. Cancer Cell **1**(2), 203–209 (2002)
44. I. Hedenfalk, D. Duggan, Y. Chen, M. Radmacher, M. Bittner, R. Simon, ... Z. Yakhini, Gene-expression profiles in hereditary breast cancer. N. Engl. J. Med. **344**(8), 539–548 (2001)
45. C.P. Lee, Y. Leu, A novel hybrid feature selection method for microarray data analysis. Appl. Soft Comput. **11**(1), 208–213 (2011)
46. El A. Akadi, A. Amine, A. El Ouardighi, D. Aboutajdine, A two-stage gene selection scheme utilizing MRMR filter and GA wrapper. Knowl. Inf. Syst. **26**(3), 487–500 (2011)
47. E. Hancer, B. Xue, D. Karaboga, M. Zhang, A binary ABC algorithm based on advanced similarity scheme for feature selection. Appl. Soft Comput. **36**, 334–348 (2015)
48. X. Chen, H. Ishwaran, Random forests for genomic data analysis. Genomics **99**(6), 323–329 (2012)
49. Y. Lu, J. Han, Cancer classification using gene expression data. Inf. Syst. **28**(4), 243–268 (2003)

50. O. Okun, Survey of novel feature selection methods for cancer classification, in *Biological Knowledge Discovery Handbook: Preprocessing, Mining, and Postprocessing of Biological Data*, vol. 379, ed. by M. Elloumi, A.Y. Zomaya (Wiley-Blackwell, Chichester, 2013)
51. Y. Bengio, *Learning Deep Architectures for AI* (Now Publishers Inc, 2009)
52. U. Alon, N. Barkai, D.A. Notterman, K. Gish, S. Ybarra, D. Mack, A.J. Levine, 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 (1999)
53. L.J. Van't Veer, H. Dai, M.J. Van De Vijver, Y.D. He, A.A. Hart, M. Mao, ... G.J. Schreiber, Gene expression profiling predicts clinical outcome of breast cancer. Nature **415**(6871), 530–536 (2002)
54. A.W. Tolcher, L.D. Mayer, Improving combination cancer therapy: the CombiPlex® development platform. Future Oncol. **14**(13), 1317–1332 (2018)
55. T.A. Yap, A. Omlin, J.S. De Bono, Development of therapeutic combinations targeting major cancer signaling pathways. J. Clin. Oncol. **31**(12), 1592–1605 (2013)
56. T.L. Lenz, D.E. Hilleman, Aggrenox: a fixed-dose combination of aspirin and dipyridamole. Ann. Pharmacother. **34**(11), 1283–1290 (2000)
57. M. Bigioni, A. Benzo, C. Irrissuto, G. Lopez, B. Curatella, C.A. Maggi, ... M. Binaschi, Antitumour effect of combination treatment with Sabarubicin (MEN 10755) and cis-platin (DDP) in human lung tumour xenograft. Cancer Chemother. Pharmacol. **62**(4), 621–629 (2008)
58. Y. Liu, B. Hu, C. Fu, X. Chen, DCDB: drug combination database. Bioinformatics **26**(4), 587–588 (2010)
59. J.C. Ashton, Drug combination studies and their synergy quantification using the Chou–Talalay method. Cancer Res. **75**(11), 2400–2400 (2015)
60. S. Loewe, Die mischarznei. Klinische Wochenschrift **6**(23), 1077–1085 (1927)
61. C.I. Bliss, The toxicity of poisons applied jointly. Ann. Appl. Biol. **26**, 585–615 (1939)
62. N.J. Sucher, Searching for synergy in silico, in vitro and in vivo. Synergy **1**(1), 30–43 (2014)
63. T.C. Chou, Theoretical basis, experimental design, and computerized simulation of synergism and antagonism in drug combination studies. Pharmacol. Rev. **58**(3), 621–681 (2006)
64. D. Day, L.L. Siu, Approaches to modernize the combination drug development paradigm. Genome Med **8**(1), 115 (2016)
65. K. Pang, Y.W. Wan, W.T. Choi, L.A. Donehower, J. Sun, D. Pant, Z. Liu, Combinatorial therapy discovery using mixed integer linear programming. Bioinformatics **30**(10), 1456–1463 (2014)
66. L. He, E. Kulesskiy, J. Saarela, L. Turunen, K. Wennerberg, T. Aittokallio, J. Tang, Methods for high-throughput drug combination screening and synergy scoring, in *Cancer Systems Biology* (Humana Press, New York, 2018), pp. 351–398
67. D. Ferreira, F. Adega, R. Chaves, The importance of cancer cell lines as in vitro models in cancer methylome analysis and anticancer drugs testing. Oncogenomics and cancer proteomics-novel approaches in biomarkers discovery and therapeutic targets in cancer (2013), pp. 139–166
68. C.D. Doern, When does 2 plus 2 equal 5 A review of antimicrobial synergy testing. J. Clin. Microbiol. **52**(12), 4124–4128 (2014)
69. J. Wildenhain, M. Spitzer, S. Dolma, N. Jarvik, R. White, M. Roy, ... M. Tyers, Prediction of synergism from chemical-genetic interactions by machine learning. Cell Syst. **1**(6), 383–395 (2015)
70. J.D. Janizek, S. Celik, S.I. Lee, Explainable machine learning prediction of synergistic drug combinations for precision cancer medicine. bioRxiv, 331769 (2018)
71. P.J. Brézillon, P. Zaraté, F. Saci, Artificial intelligence approach in analysis of DNA sequences. Biochimie **75**(5), 337–345 (1993)
72. H.C. Shin, H.R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, ... R.M. Summers, Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE Trans. Med. Imaging **35**(5), 1285–1298 (2016)
73. S. Kombrink, T. Mikolov, M. Karafiát, L. Burget, Recurrent neural network based language modeling in meeting recognition, in *Twelfth Annual Conference of the International Speech Communication Association* (2011)

74. D. Wang, A. Khosla, R. Gargeya, H. Irshad, A.H. Beck, Deep learning for identifying metastatic breast cancer. arXiv preprint arXiv:1606.05718 (2016)
75. A.M. Abdel-Zaher, A.M. Eldeib, Breast cancer classification using deep belief networks. Expert Syst. Appl. **46**, 139–144 (2016)
76. H. Liao, A deep learning approach to universal skin disease classification. University of Rochester Department of Computer Science, CSC (2016)
77. R. Fakoor, F. Ladhak, A. Nazi, M. Huber, Using deep learning to enhance cancer diagnosis and classification, in *Proceedings of the international conference on machine learning*, vol. 28 (ACM, New York, 2013)
78. E.M. Mashhour, M.F. El Enas, K.T. Wassif, A.I. Salah, Feature selection approach based on firefly algorithm and chi-square. Int. J. Elect. Comput. Eng. **8**(4), 2338 (2018)
79. S. Park, S. Min, H. Choi, S. Yoon, deepMiRGene: Deep neural network based precursor microrna prediction. arXiv preprint arXiv:1605.00017 (2016)
80. S. Cheng, M. Guo, C. Wang, X. Liu, Y. Liu, X. Wu, MiRTDL: a deep learning approach for miRNA target prediction. IEEE/ACM Trans. Comput. Biol. Bioinform. **13**(6), 1161–1169 (2015)
81. O. Barca-Mayo, Q.R. Lu, Fine-tuning oligodendrocyte development by microRNAs. Front. Neurosci. **6**, 13 (2012)
82. A. Sharma, R. Rani, A Systematic Review of Applications of Machine Learning in Cancer Prediction and Diagnosis. Arch. Comput. Methods Eng., 1–22 (2021). https://doi.org/10.1007/s11831-021-09556-z