frontiers | Frontiers in Medicine

TYPE Review
PUBLISHED 10 January 2023
DOI 10.3389/fmed.2022.1109365

OPEN ACCESS

EDITED BY
Liang Zhao,
Dalian University of Technology, China

REVIEWED BY
Huiyuan Lai,
University of Groningen, Netherlands
Jinyang Huang,
Hefei University of Technology, China

*CORRESPONDENCE
Zhengxia Wang
✉ 22120854000026@hainanu.edu.cn

SPECIALTY SECTION
This article was submitted to Precision Medicine, a section of the journal Frontiers in Medicine

RECEIVED 27 November 2022
ACCEPTED 28 December 2022
PUBLISHED 10 January 2023

CITATION
Wang S, Wang S and Wang Z (2023) A survey on multi-omics-based cancer diagnosis using machine learning with the potential application in gastrointestinal cancer.
Front. Med. 9:1109365.
doi: 10.3389/fmed.2022.1109365

A survey on multi-omics-based cancer diagnosis using machine learning with the potential application in gastrointestinal cancer

Suixue Wang[1], Shuling Wang[2] and Zhengxia Wang[3]*

[1]School of Information and Communication Engineering, Hainan University, Haikou, China, [2]Department of Neurology, Affiliated Haikou Hospital of Xiangya School of Medicine, Central South University, Haikou, China, [3]School of Computer Science and Technology, Hainan University, Haikou, China

Gastrointestinal cancer is becoming increasingly common, which leads to over 3 million deaths every year. No typical symptoms appear in the early stage of gastrointestinal cancer, posing a significant challenge in the diagnosis and treatment of patients with gastrointestinal cancer. Many patients are in the middle and late stages of gastrointestinal cancer when they feel uncomfortable, unfortunately, most of them will die of gastrointestinal cancer. Recently, various artificial intelligence techniques like machine learning based on multi-omics have been presented for cancer diagnosis and treatment in the era of precision medicine. This paper provides a survey on multi-omics-based cancer diagnosis using machine learning with potential application in gastrointestinal cancer. Particularly, we make a comprehensive summary and analysis from the perspective of multi-omics datasets, task types, and multi-omics-based integration methods. Furthermore, this paper points out the remaining challenges of multi-omics-based cancer diagnosis using machine learning and discusses future topics.

KEYWORDS

gastrointestinal cancer, multi-omics, machine learning, deep learning, integration

1. Introduction

Cancer is one of the leading causes of death worldwide (1), usually with few symptoms in the early stage. However, once a patient is diagnosed with cancer, it is in the advanced stage of cancer. Cancer has a high morbidity and mortality rate worldwide and has become a common human disease, therefore, it poses a great threat to human beings. According to statistics (1), in 2020, there were about 19.3 million new cancer patients globally, and nearly 10.0 million patients died of cancer. Specifically, The number of new cases of breast cancer in the world reaches 2.3 million a year, becoming the most common cancer type globally, while 1.8 million cases of lung cancer deaths a year, rank first in the global cancer death population. Breast, lung, colorectal, stomach, liver, and prostate cancers are the most general types of cancer. Among them, all cancers in the

digestive tract organs like colorectal cancer, gastric cancer, and liver cancer belong to gastrointestinal cancer, also becoming more and more common. In other words, the number of patients with these cancers is increasing every year. According to the institute for cancer research (1), some known risk factors such as smoking, unhealthy diet, being overweight and physical inactivity largely cause cancers, such as gastrointestinal cancer. It is extraordinarily difficult for doctors to diagnose and treat cancer patients, although surgery is one of the approaches to treatment for cancer patients, the recurrence rate is still high. Unfortunately, neither chemotherapy nor radiotherapy is ideal.

In order to improve the cancer treatment effect, as well as prolong the survival time for cancer patients, it is very essential to improve capabilities in precision medicine by using specific information about a patient's tumor to help make an accurate diagnosis, plan an effective treatment, find out how well treatment is working, or make a prognosis. In particular, accurate diagnosis can be used for the early diagnosis of cancer, many cancers can be cured if detected early and treated effectively (2). Even in the middle and advanced stages, being able to accurately diagnose cancer types or cancer molecular subtypes, also have a certain significance in enhancing treatment effects, improving the quality of life, and prolonging the life of patients.

However, accurate diagnosis of cancer is a scientific problem in the field of biomedicine. Fortunately, over the past few years, with the development of artificial intelligence technology, especially machine learning (ML) and deep learning (DL) (3), smart medicine has developed rapidly (4). Smart medicine combines artificial intelligence technologies such as ML with medical theories and then applies them to pathological reports of cancer patients for converting cancer diagnosis into problems of classification, regression, or clustering. Especially in recent years, cancer diagnosis based on artificial intelligence has made great progress. As illustrated in Figure 1, the number of publications demonstrated that the multi-omics-based integrative methods using ML have become increasing interest in the area of cancer diagnosis over the last decade. For example, Stanford computer scientists have created an AI diagnostic algorithm that diagnosed skin cancer as well as a board-certified dermatologist (5). It is particularly emphasized that ML technology based on multi-omics is playing an important role in cancer diagnosis like survival analysis, drug sensitivity response, etc. (6, 7), and they have achieved corresponding curative effects on various cancers.

This paper provides a comprehensive review of multi-omics-based ML models or artificial intelligence technologies in the field of cancer diagnosis, and then we highlight its prospects and applications in gastrointestinal cancer. Finally, we point out the difficulties in the current multi-omics-based ML integration methods and discuss some future research directions.

As shown in Figure 2, the rest of this paper is organized as follows. In Section 2, we detail the cancer task types based on multi-omics. We review some commonly used open-source

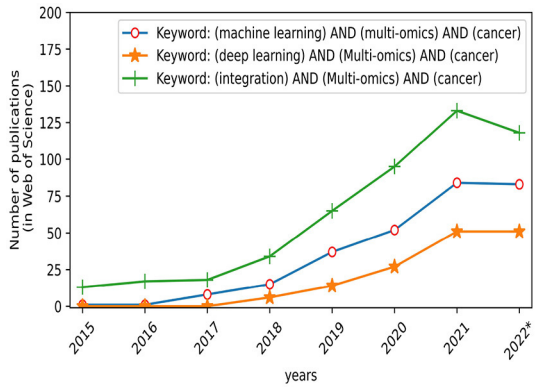

FIGURE 1
Number of publications published each year on different search keywords.

omics databases in Section 3. Section 4 summarizes and discusses the state-of-the-art multi-omics-based ML integration methods for cancer diagnosis. Finally, Section 5 concludes the work and points out challenges and future directions.

2. Multi-omics-based cancer task types

Typical types of cancer tasks based on multi-omics data integration methods are cancer molecular subtype classification, survival analysis, drug response prediction, and biomarker discovery. In addition, some tasks are not well studied in the literature, such as metastasis prediction, recurrence prediction, etc., they will not be discussed in this review.

2.1. Molecular subtype classification and discovery of new subtypes

To customize the optimal treatment strategy for patients and achieve the purpose of precision medicine, it is of great significance to improve the accuracy of a cancer diagnosis. Specifically, cancer is generally further divided into multiple molecular subtypes, and different molecular subtypes adopt different treatment strategies to achieve the best therapeutic effect (8). For example, breast cancer is subdivided into four molecular subtypes: HER2-enriched, Luminal A, Luminal B, and Basal-like (9). Each subtype is associated with a unique panel of mutated genes. Therefore, the task of cancer subtype classification is to automatically identify defined subtypes based on the multi-omics measurement results of patients (10–13). In addition, with the growth of multi-omics data, there are still some potential cancer subtypes that need to be mined (14–16).

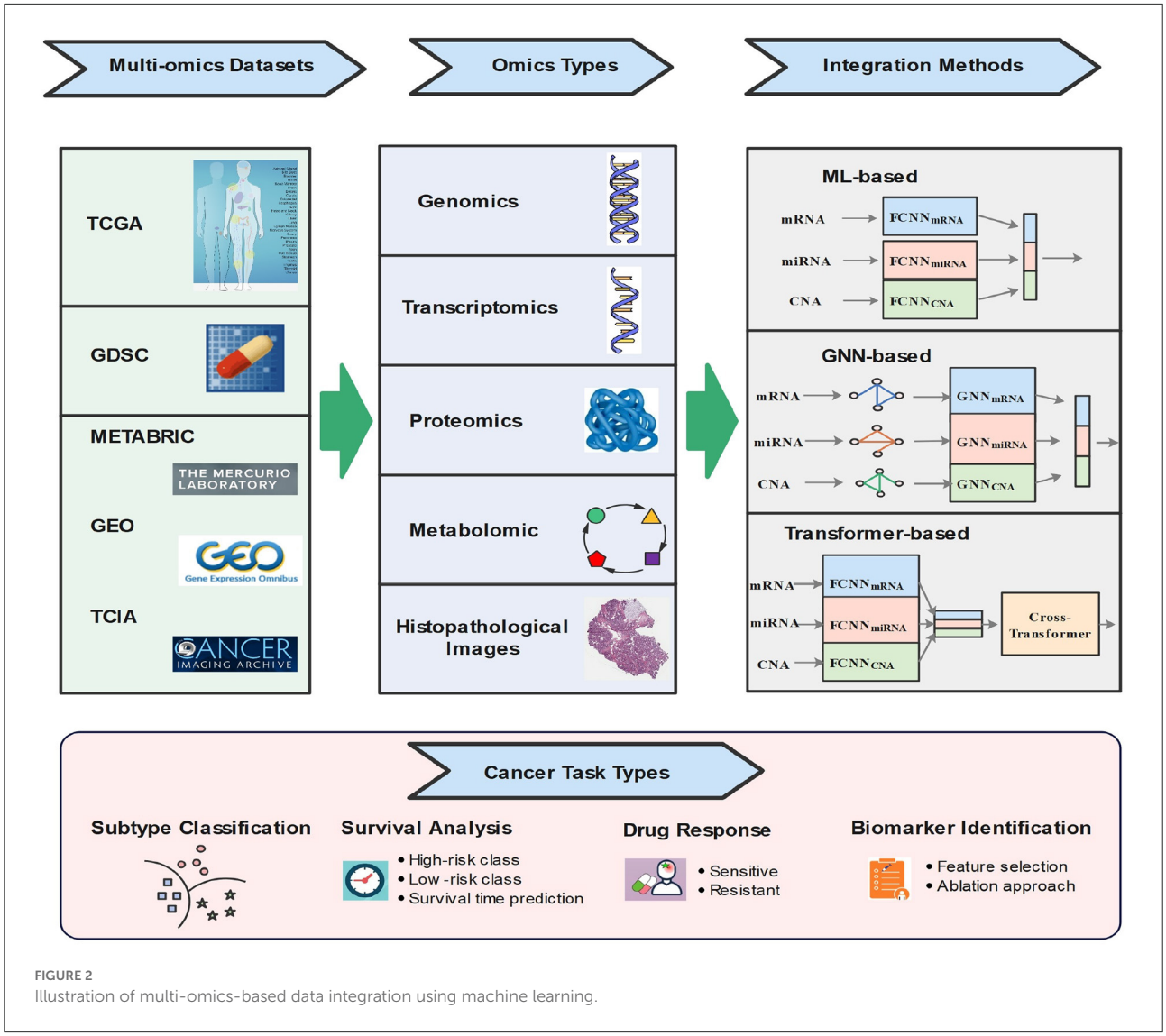

FIGURE 2
Illustration of multi-omics-based data integration using machine learning.

Therefore, the identification of cancer subtypes is usually treated as a supervised classification problem, and the discovery of new subtypes is generally treated as a clustering problem.

2.2. Survival analysis

To improve the survival rate of cancer patients, a large number of researchers studied and analyzed the factors affecting their survival times by collecting the survival times of cancer patients and using machine learning methods to discover possible survival rules (17, 18). Cancer survival analysis can be defined as a binary classification or a risk regression problem (19, 20). In a binary classification task, patients are divided into short-term and long-term survival groups, or low-risk and high-risk survival groups, according to a predefined survival time threshold (e.g., 5 years) (21). In risk regression tasks, the Cox proportional hazards model and its extensions are often used to calculate the risk score for each patient.

2.3. Drug response prediction

IC50 is widely used to assess the sensitivity of drug response, and it is the concentration of drug required to reduce the number of viable cells by half after drug administration (22). Drug response prediction tasks are the same as survival analysis tasks. The binary classification tasks and regression tasks are often used in drug response research (19, 23–26). In regression problems, drug response directly predicts IC50 values, while in binary classification tasks, a predefined threshold based on the distribution of IC50 values is used to predict drug response as sensitive or resistant.

2.4. Biomarker discovery

In this review, the goal of biomarker discovery is to find genes associated with cancer prognosis by combining multi-omics data, which can advance the understanding of molecular mechanisms of cancer and offer new ideas for clinical diagnosis and treatment (27, 28). For example, in clinical practice in gastrointestinal cancer, CEA is the most commonly used marker. Furthermore, some biomarkers have been used for cervical cancer including miR-215-5p, miR-192-5p, KAT2B, PCNA, and CD86 (29). Biomarkers are widely identified by using the methods of feature selection and feature importance ranking in traditional ML (10, 30, 31). When analyzing the contribution of each feature in multiple omics sources, the feature will be set to 0 in turn, and then the performance of the classification or regression model will be calculated, and it will be compared to performance using all features (10).

3. Multi-omics datasets

It has become increasingly apparent that many novel omics data sequencing technologies have emerged since the Human Genome Project was proposed and implemented (32–34), and the cost of sequencing like high-throughput, is gradually decreasing. Therefore, we can quickly obtain high-dimensional multi-omics data and provide data sources for research in the fields of biomedicine and bioinformatics.

3.1. Multi-omics datasets

In this section, we introduce the multi-omics cancer datasets that are widely used in the literature. The multi-omics datasets are shown in Table 1.

The Cancer Genome Atlas (TCGA) is a project jointly launched by the National Cancer Institute (NCI) and the National Human Genome Research Institute (NHGRI) in 2006 (35). It includes clinical information, histopathological images, and multiple omics data like genomics, transcriptomics, proteomics, and epigenomics. Especially, genomics and transcriptomics, are the most commonly used types of omics. For example, the data types of DNA methylation and copy number variation in genomics, as well as the data types of mRNA expression and miRNA expression in transcriptomics, appear most frequently in the literature on multi-omics-based ML integration methods. TCGA dataset currently includes a total of 33 types of cancer. In particular, all gastrointestinal cancer, including gastric cancer, colorectal cancer, liver cancer, etc., can be obtained in TCGA. TCGA is free and open, which greatly helps cancer researchers to improve the prevention, diagnosis, and treatment of cancer.

The Genomics of Drug Sensitivity in Cancer (GDSC) omics database was jointly developed by the Wellcome Trust Sanger Institute in the United Kingdom and the Massachusetts General Hospital Cancer Center in the United States (36). GDSC collects drug response data (IC50) of about 200 anticancer drugs in more than 1,000 human cancer cell lines. Variations in the cancer genome can affect the effectiveness of clinical treatment, and different targets have very different responses to drugs. Therefore, the GDSC data are really important for the discovery of potential tumor therapeutic targets, which have been widely used in anticancer drug screening.

In addition to the TCGA and GDSC datasets, other widely used databases also appear in the relevant literature, such as Molecular Taxonomy of Breast Cancer International Consortium (METABRIC), COSMIC Cell Lines, CPTAC, LinkedOmics, and the Cancer Imaging Archive (TCIA) (37–41).

3.2. Multi-omics challenges based on machine learning

Although multi-omics data can be used for cancer diagnosis using ML integration methods, there are still some problems with multi-omics data. We list some of the challenges that are quite general in the relevant literature (6, 42) as follows.

3.2.1. Small sample size

The first challenge is that almost all existing omics datasets suffer from the problem of a small number of observations in a specific class, with most classes having <100 observations. The features of omics usually have higher dimensionality, which is much larger than the number of observed samples, leading to the problem of the curse of dimensionality (43). In this case, it is crucial to use a reasonable evaluation method to estimate the classification error.

3.2.2. Missing values

There are many missing values in clinical information and omics sequencing results in multi-omics datasets. Some studies have proposed that (17, 19) when a feature has more than 20% missing values in omics data, this feature will be discarded. At the same time, if our experimental content is to integrate and analyze multiple omics data, when patients lack any kind of omics data, the observation sample of this patient will also be discarded.

3.2.3. Class imbalance

There is a problem of class distribution imbalance between different cancer types, as well as between different cancer

TABLE 1 Frequently used multi-omics datasets.

Dataset names	Data types	Supported task types	URL
TCGA (35)	• Genomics • Transcriptomics • Epigenomics • Proteomics • Slide Image	• Subtypes classification • Biomarker discovery • Survival analysis • Drug response	https://portal.gdc.cancer.gov/
GDSC (36)	• Genomics • Drug response	• Drug response • Biomarker discovery	https://www.cancerrxgene.org/
METABRIC (37)	• Genomics	• Subtypes classification • Biomarker discovery	https://www.cbioportal.org/study/ summary?id=brca-metabric
COSMIC Cell Lines (38)	• Genomics • Transcriptomics • Epigenomics • Drug response	• Subtypes classification • Biomarker discovery • Survival analysis • Drug response	https://cancer.sanger.ac.uk/cell-lines
CPTAC (39)	• Proteomics • Slide Image	• Subtypes classification • Biomarker discovery • Survival analysis	https://pdc.cancer.gov/pdc/
LinkedOmics (40)	• Genomics • Transcriptomics • Proteomics	• Subtypes classification • Biomarker discovery • Survival analysis	http://www.linkedomics.org/login.php
TCIA (41)	• Radiomics • Slide Image • Genomics	• Subtypes classification • Biomarker discovery • Survival analysis	https://www.cancerimagingarchive.net/

molecular subtypes, respectively. To solve this problem, up-sampling and down-sampling techniques are usually employed (44).

4. Data integration methods for multi-omics using machine learning

In recent years, with the increase in computing power and the decline in the cost of high-throughput sequencing, and the success of ML technology in various fields, ML has been widely employed in the fields of biomedical and bioinformatics computing (45). In particular, a variety of novel data integration models have been introduced from ML. There are some reviews for summarizing the data integration methods based on multi-omics (6, 46), for example, the data are fused according to three strategies of early fusion, intermediate fusion, and late fusion (7). In addition, some reviews classify data integration methods according to concatenation-based, transformed-based, and model-based approaches (42). In this section, we classify the newly proposed data integration models according to three types of groups: traditional ML-based, transformer-based, and graph neural network based.

4.1. Conventional machine learning technologies

Here we will briefly introduce three subgroups of models applied in data integration: traditional ML models, classical deep learning models, and auto-encoder models.

Logistic regression (LR), support vector machine (SVM), random forest (RF) and Xgboost are widely used traditional ML models (30, 31, 47), before feeding the data to them, it generally needs to reduce the dimensionality of high-dimensional features of multiple omics data based on feature extraction methods such as nearest component analysis (NCA) (19, 23) and principal component analysis (PCA) (21), and then concatenate the dimensionality-reduced features and finally feed the concatenated features to the model.

In contrast, it is not necessary for classical deep learning models like fully connected neural networks (FCNNs) and convolution neural networks (CNNs) to reduce the dimensionality of omics features to very low dimensions, due the models can automatically learn useful information from high-dimensional space (11, 48–51).

Auto-encoder is an unsupervised neural network model where the network can be replaced by FCNN, CNN, or other DNNs (16). Auto-encoder compresses the data to a lower

dimension, which is called encoding, and then reconstructs the original input data back, which is called decoding (11). Intuitively, auto-encoder can be used for dimensionality reduction which is similar to PCA, but its performance is stronger than PCA due to the neural network model can extract more effective new features (52). In addition to dimensionality reduction, new features learned by the auto-encoder can be fed into the supervised learning model for the tasks of classification or regression.

4.2. Graphic neural network technologies

In recent years, Graph Neural Networks (GNNs) have shown strong capabilities in handling non-Euclidean graph-structured data by naturally combining network topology structure and the information of node and link, GNNs have been employed to integrate multi-omics data since the last 2 years. Wang et al. (10) proposed MOGONET, for each omics data type, a weighted sample similarity network was constructed according to omics characteristics as the input of GNN and used for the identification of biomarkers. In Xing et al. (53), MLE-GAT is presented to explore the correlation information between genes contained in omics data. It assumes that genes usually interact rather than acting alone, so the weighted correlation network analysis (WGCNA) is firstly used to convert each patient's omics data into a co-expression map as the input of GAT, and then GAT outputs each node feature as a weighted combination of its neighbors and the current node.

4.3. Transformer technologies

The transformer model is widely used in different fields such as natural language processing and computer vision, it is becoming one of the most frequently used deep learning models (54–56). The success of the transformer architecture depends on the multi-head attention mechanism that calculates the attention between different positions in the input sequence multiple times. The Transformer is applied to multi-omics-based integration techniques since 2021, which is a relatively new data integration method. For example, in these two papers (57, 58), the transformers are used to calculate the cross-attention between multi-omics features and histopathological image features.

5. Conclusion

In this paper, we review the multi-omics-based integrative approaches using ML with potential applications in gastrointestinal cancer. Firstly, several cancer task types are elaborated on and discussed. Then we describe widely used cancer multi-omics datasets, and the challenges encountered in their use for integration based on ML. Finally, we analyze currently the state-of-the-art multi-omics-based data integration approaches in detail and divide them into three groups such as conventional ML technologies, graphic neural network technologies, and transformer technologies.

Although ML has performed excellently in the application of multi-omics data integration, there are still some challenges that require us to consider and explore deeply. Specifically, the existing methods for a missing value of multi-omics data are almost all treated as discards, rather than trying to fill them in. Therefore, to efficiently utilize the existing precious multi-omics data, it is necessary to further explore the method of filling in missing values in multi-omics. Additionally, since biomedical data are very precious and difficult to obtain, the patients who contain multiple omics sources and histopathology images simultaneously are particularly scarce. Hence, in the future, a pre-trained visual representation model may be transferred to histopathology images on a limited number of samples, which can be potentially solved by few-shot learning strategies. More importantly, more effective approaches for integrating multi-omics and histopathology images need further investigation for gastrointestinal cancer diagnosis and treatment, as a promising future research direction.

Author contributions

ShW contributed to the conception of the study and the verification of analytical methods. SuW performed the statistical analysis and wrote the manuscript. ZW revised the manuscript and approved the version to be published. All authors contributed to the article and approved the submitted version.

Acknowledgments

Thanks to the grants with Nos. 2021-016, ZDYF2021GXJS16 and 6216070295 for support.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin*. (2021) 71:209–49. doi: 10.3322/caac.21660

2. Pashayan N, Pharoah PD. The challenge of early detection in cancer. *Science*. (2020) 368:589–90. doi: 10.1126/science.aaz2078

3. Zhang X, Shams SP, Yu H, Wang Z, Zhang Q. A pairwise functional connectivity similarity measure method based on few-shot learning for early MCI detection. *Front Neurosci*. (2022) 16:1081788. doi: 10.3389/fnins.2022.1081788

4. Wang F, Casalino LP, Khullar D. Deep learning in medicine–promise, progress, and challenges. *JAMA Internal Med*. (2019) 179:293–294. doi: 10.1001/jamainternmed.2018.7117

5. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*. (2017) 542:115–8. doi: 10.1038/nature21056

6. Reel PS, Reel S, Pearson E, Trucco E, Jefferson E. Using machine learning approaches for multi-omics data analysis: a review. *Biotechnol Adv*. (2021) 49:107739. doi: 10.1016/j.biotechadv.2021.107739

7. Leng D, Zheng L, Wen Y, Zhang Y, Wu L, Wang J, et al. A benchmark study of deep learning-based multi-omics data fusion methods for cancer. *Genome Biol*. (2022) 23:1–32. doi: 10.1186/s13059-022-02739-2

8. Ogino S, Fuchs CS, Giovannucci E. How many molecular subtypes? Implications of the unique tumor principle in personalized medicine. *Expert Rev Mol Diagn*. (2012) 12:621–8. doi: 10.1586/erm.12.46

9. Hon JDC, Singh B, Sahin A, Du G, Wang J, Wang VY, et al. Breast cancer molecular subtypes: from TNBC to QNBC. *Am J Cancer Res*. (2016) 6:1864.

10. Wang T, Shao W, Huang Z, Tang H, Zhang J, Ding Z, et al. MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification. *Nat Commun*. (2021) 12:1–13. doi: 10.1038/s41467-021-23774-w

11. Islam MM, Huang S, Ajwad R, Chi C, Wang Y, Hu P. An integrative deep learning framework for classifying molecular subtypes of breast cancer. *Comput Struct Biotechnol J*. (2020) 18:2185–99. doi: 10.1016/j.csbj.2020.08.005

12. Ektefaie Y, Yuan W, Dillon DA, Lin NU, Golden JA, Kohane IS, et al. Integrative multiomics-histopathology analysis for breast cancer classification. *NPJ Breast Cancer*. (2021) 7:1–6. doi: 10.1038/s41523-021-00357-y

13. Yin C, Cao Y, Sun P, Zhang H, Li Z, Xu Y, et al. Molecular subtyping of cancer based on robust graph neural network and multi-omics data integration. *Front Genet*. (2022) 13:884028. doi: 10.3389/fgene.2022.884028

14. Hira MT, Razzaque M, Angione C, Scrivens J, Sawan S, Sarker M. Integrated multi-omics analysis of ovarian cancer using variational autoencoders. *Sci Rep*. (2021) 11:1–16. doi: 10.1038/s41598-021-85285-4

15. Liu J, Ge S, Cheng Y, Wang X. Multi-view spectral clustering based on multi-smooth representation fusion for cancer subtype prediction. *Front Genet*. (2021) 12:718915. doi: 10.3389/fgene.2021.718915

16. Franco EF, Rana P, Cruz A, Calderón VV, Azevedo V, Ramos RT, et al. Performance comparison of deep learning autoencoders for cancer subtype detection using multi-omics data. *Cancers*. (2021) 13:2013. doi: 10.3390/cancers13092013

17. Zhao L, Dong Q, Luo C, Wu Y, Bu D, Qi X, et al. DeepOmix: a scalable and interpretable multi-omics deep learning framework and application in cancer survival analysis. *Comput Struct Biotechnol J*. (2021) 19:2719–2725. doi: 10.1016/j.csbj.2021.04.067

18. He Z, Zhang J, Yuan X, Zhang Y. Integrating somatic mutations for breast cancer survival prediction using machine learning methods. *Front Genet*. (2021) 11:632901. doi: 10.3389/fgene.2020.632901

19. Malik V, Kalakoti Y, Sundar D. Deep learning assisted multi-omics integration for survival and drug-response prediction in breast cancer. *BMC Genomics*. (2021) 22:1–11. doi: 10.1186/s12864-021-07524-2

20. Xie G, Dong C, Kong Y, Zhong JF, Li M, Wang K. Group lasso regularized deep learning for cancer prognosis from multi-omics and clinical features. *Genes*. (2019) 10:240. doi: 10.3390/genes10030240

21. Tong L, Mitchel J, Chatlin K, Wang MD. Deep learning based feature-level integration of multi-omics data for breast cancer patients survival analysis. *BMC Med Inform Decis Mak*. (2020) 20:1–12. doi: 10.1186/s12911-020-01225-8

22. Stanfield Z, Coşkun M, Koyutürk M. Drug response prediction as a link prediction problem. *Sci Rep*. (2017) 7:1–13. doi: 10.1038/srep40321

23. Khan D, Shedole S. Leveraging deep learning techniques and integrated omics data for tailored treatment of breast cancer. *J Person Med*. (2022) 12:674. doi: 10.3390/jpm12050674

24. Clayton EA, Pujol TA, McDonald JF, Qiu P. Leveraging TCGA gene expression data to build predictive models for cancer drug response. *BMC Bioinform*. (2020) 21:1–11. doi: 10.1186/s12859-020-03690-4

25. Park S, Soh J, Lee H. Super. FELT: supervised feature extraction learning using triplet loss for drug response prediction with multi-omics data. *BMC Bioinform*. (2021) 22:1–23. doi: 10.1186/s12859-021-04146-z

26. Sharifi-Noghabi H, Zolotareva O, Collins CC, Ester M. MOLI: multi-omics late integration with deep neural networks for drug response prediction. *Bioinformatics*. (2019) 35:i501–9. doi: 10.1093/bioinformatics/btz318

27. Goossens N, Nakagawa S, Sun X, Hoshida Y. Cancer biomarker discovery and validation. *Transl Cancer Res*. (2015) 4:256.

28. Dhillon A, Singh A, Bhalla VK. A systematic review on biomarker identification for cancer diagnosis and prognosis in multi-omics: from computational needs to machine learning and deep learning. *Arch Comput Methods Eng*. (2022) 9:1–33. doi: 10.1007/s11831-022-09821-9

29. Kori M, Yalcin Arga K. Potential biomarkers and therapeutic targets in cervical cancer: insights from the meta-analysis of transcriptomics data within network biomedicine perspective. *PLoS ONE*. (2018) 13:e0200717. doi: 10.1371/journal.pone.0200717

30. Fan Y, Kao C, Yang F, Wang F, Yin G, Wang Y, et al. Integrated multi-omics analysis model to identify biomarkers associated with prognosis of breast cancer. *Front Oncol*. (2022) 12:899900. doi: 10.3389/fonc.2022.899900

31. Xu H, Lien T, Bergholtz H, Fleischer T, Djerroudi L, Vincent-Salomon A, et al. Multi-Omics marker analysis enables early prediction of breast tumor progression. *Front Genet*. (2021) 12:670749. doi: 10.3389/fgene.2021.670749

32. Watson JD. The human genome project: past, present, and future. *Science*. (1990) 248:44–49. doi: 10.1126/science.2181665

33. Collins FS, McKusick VA. Implications of the human genome project for medical science. *JAMA*. (2001) 285:540–4. doi: 10.1001/jama.285.5.540

34. Gibbs RA. The human genome project changed everything. *Nat Rev Genet*. (2020) 21:575–6. doi: 10.1038/s41576-020-0275-3

35. Tomczak K, Czerwińska P, Wiznerowicz M. Review the cancer genome atlas (TCGA): an immeasurable source of knowledge. *Contemporary Oncol/Współczesna Onkol*. (2015) 2015:68–77. doi: 10.5114/wo.2014.47136

36. Yang W, Soares J, Greninger P, Edelman EJ, Lightfoot H, Forbes S, et al. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res*. (2012) 41:D955–61. doi: 10.1093/nar/gks1111

37. Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, Dunning MJ, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*. (2012) 486:346–52. doi: 10.1038/nature10983

38. Iorio F, Knijnenburg TA, Vis DJ, Bignell GR, Menden MP, Schubert M, et al. A landscape of pharmacogenomic interactions in cancer. *Cell*. (2016) 166:740–54. doi: 10.1016/j.cell.2016.06.017

39. Edwards NJ, Oberti M, Thangudu RR, Cai S, McGarvey PB, Jacob S, et al. The CPTAC data portal: a resource for cancer proteomics research. *J Proteome Res*. (2015) 14:2707–13. doi: 10.1021/pr501254j

40. Vasaikar SV, Straub P, Wang J, Zhang B. LinkedOmics: analyzing multi-omics data within and across 32 cancer types. *Nucleic Acids Res*. (2018) 46:D956–63. doi: 10.1093/nar/gkx1090

41. Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, et al. The cancer imaging archive (TCIA): maintaining and operating a public information repository. *J Digit Imaging*. (2013) 26:1045–57. doi: 10.1007/s10278-013-9622-7

42. Momeni Z, Hassanzadeh E, Abadeh MS, Bellazzi R. A survey on single and multi omics data mining methods in cancer data classification. *J Biomed Inform*. (2020) 107:103466. doi: 10.1016/j.jbi.2020.103466

43. Verleysen M, François D. The curse of dimensionality in data mining and time series prediction. In: *International Work-conference on Artificial Neural Networks*. Berlin; Heidelberg: Springer (2005). p. 758–70.

44. Provost F. Machine learning from imbalanced data sets 101. In: *Proceedings of the AAAI'2000 Workshop on Imbalanced Data Sets. Vol. 68*. Austin, TX: AAAI Press (2000). p. 1–3.

45. Wang S, Wang S, Liu Z, Zhang Q. A role distinguishing Bert model for medical dialogue system in sustainable smart city. *Sustain Energy Technol Assess*. (2023) 55:102896. doi: 10.1016/j.seta.2022.102896

46. Nicora G, Vitali F, Dagliati A, Geifman N, Bellazzi R. Integrated multi-omics analyses in oncology: a review of machine learning methods and tools. *Front Oncol*. (2020) 10:1030. doi: 10.3389/fonc.2020.01030

47. Carrillo-Perez F, Morales JC, Castillo-Secilla D, Gevaert O, Rojas I, Herrera LJ. Machine-learning-based late fusion on multi-omics and multi-scale data for non-small-cell lung cancer diagnosis. *J Pers Med*. (2022) 12:601. doi: 10.3390/jpm12040601

48. Huang Z, Zhan X, Xiang S, Johnson TS, Helm B, Yu CY, et al. SALMON: survival analysis learning with multi-omics neural networks on breast cancer. *Front Genet*. (2019) 10:166. doi: 10.3389/fgene.2019.00166

49. Yu H, Yang LT, Zhang Q, Armstrong D, Deen MJ. Convolutional neural networks for medical image analysis: state-of-the-art, comparisons, improvement and perspectives. *Neurocomputing*. (2021) 444:92–110. doi: 10.1016/j.neucom.2020.04.157

50. Yu H, Yang LT, Fan X, Zhang Q. A deep residual computation model for heterogeneous data learning in smart Internet of Things. *Appl Soft Comput*. (2021) 107:107361. doi: 10.1016/j.asoc.2021.107361

51. Hu X, Ding X, Bai D, Zhang Q. A compressed model-agnostic meta-learning model based on pruning for disease diagnosis. *J Circ Syst Comput*. (2022) 2022:2350022. doi: 10.1142/S0218126623500226

52. Han K, Wang Y, Zhang C, Li C, Xu C. Autoencoder inspired unsupervised feature selection. In: 2018 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Calgary, AB: IEEE (2018). p. 2941–5.

53. Xing X, Yang F, Li H, Zhang J, Zhao Y, Gao M, et al. An interpretable multi-Level enhanced graph attention network for disease diagnosis with gene expression data. In: *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. Houston, TX: IEEE (2021). p. 556–61.

54. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: *Advances in Neural Information Processing Systems. Vol. 30*. Curran Associates, Inc. (2017). Available online at: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

55. Devlin J, Chang MW, Lee K, Toutanova K. Bert: pre-training of deep bidirectional transformers for language understanding. *arXiv preprint* arXiv:181004805 (2018). doi: 10.48550/arXiv.1810.04805

56. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint* arXiv:201011929. (2020). doi: 10.48550/arXiv.2010.11929

57. Lv Z, Lin Y, Yan R, Yang Z, Wang Y, Zhang F. PG-TFNet: Transformer-based fusion network integrating pathological images and genomic data for cancer survival analysis. *In: 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. Houston, TX: IEEE (2021). p. 491–6.

58. Lv Z, Lin Y, Yan R, Wang Y, Zhang F. TransSurv: transformer-based survival analysis model integrating histopathological images and genomic data for colorectal cancer. *IEEE/ACM Trans Comput Biol Bioinform*. (2022) 1–10. doi: 10.1109/TCBB.2022.3199244