Computational Techniques and Tools for Omics Data Analysis: State-of-the-Art, Challenges, and Future Directions

Article *in* Archives of Computational Methods in Engineering · February 2021

DOI: 10.1007/s11831-021-09547-0

CITATIONS
58

READS
3,595

3 authors, including:

Parampreet Kaur
Thapar University
6 PUBLICATIONS **153** CITATIONS
SEE PROFILE

Ashima Singh
Thapar Institute of Engineering and Technology Patiala
59 PUBLICATIONS **882** CITATIONS
SEE PROFILE

Archives of Computational Methods in Engineering
https://doi.org/10.1007/s11831-021-09547-0

ORIGINAL PAPER

Computational Techniques and Tools for Omics Data Analysis: State-of-the-Art, Challenges, and Future Directions

Parampreet Kaur[1] · **Ashima Singh**[1] · **Inderveer Chana**[1]

Received: 20 June 2020 / Accepted: 10 January 2021

Abstract
The heterogeneous and high-dimensional nature of omics data presents various challenges in gaining insights while analysis. In the era of big data, omics data is available as genome, proteome, transcriptome, and metabolome. Apart from the single omics data type, integrative omics known as multi-omics, and omics imaging data known as radiomics approaching to big data are being used for predictive analysis. The various computational approaches such as data mining, machine learning, deep learning, statistical methods, metaheuristic techniques have gained attention to process, normalize, integrate, analyse omics data. This paper presents the critical review of state-of-the-art techniques for omics, multi-omics, radiomics data analysis themed at disease prediction, disease recurrence, survival analysis, and biomarker discovery. The paper investigates, compares and categorizes various existing tools and technologies based on common characteristics for integration and analysis of omics data. In addition, the significant research challenges and directions are discussed for futuristic omics research. This survey would guide researchers to understand the use of computationally intelligent approaches for efficient omics data analysis.

1 Introduction

Rapid advances in high throughput technologies of high-throughput omics (genomic, transcriptomic, proteomic, and metabolomic) have led to the fast accumulation of patient data. Currently, biomedical research has abundant data but still starves for knowledge. Computational bioinformatics plays a crucial role and helps in knowledge discovery by dealing with storage, retrieval, and optimal use of omics data. Also, in precision medicine, several aspects of patient data such as molecular traits, lifestyle, and environment are taken to ensure that the right patient is getting the right treatment at the right time [1]. In cancer research, the integration of multi-omics is highly required because of omics data production in projects such as The Cancer Genome Atlas (TCGA), International Cancer Genome Consortium (ICGC), and Therapeutically Applicable Research to Generate Effective Treatments (TARGET) [2]. So there exist unprecedented opportunities like integration and data-intensive analysis of omics data for complex diseases such as cancer [3].

For transforming health care, gaining knowledge from omics data having features like high dimensionality, complexity, and heterogeneity is a significant task. To exploit the available data, feature engineering plays its role to obtain more robust and useful features. It involves numerous intelligent techniques and technologies such as machine learning, deep learning, data mining, and statistical learning approaches [2, 4–6]. Further, models for clustering and prediction are built. In both steps, challenges like complicated data scenarios and lack of sufficient domain knowledge are there. To obtain learning models from complex data, various learning technologies are used. Machine learning and multi-omics technologies transformed the method of acquiring and processing data. Machine Learning (ML) algorithms learn from data without the need for processes and model knowledge. So, ML algorithms' strength lies in the quality and size of data used. Nowadays, because of sequencing and molecular technologies, a massive amount of inexpensive and high-quality omics data is generated. These datasets majorly are not able to be used as training data for machine

✉ Parampreet Kaur
pkaur20_phd17@thapar.edu

Ashima Singh
ashima@thapar.edu

Inderveer Chana
inderveer@thapar.edu

[1] Computer Science and Engineering Department, Thapar Institute of Engineering and Technology, Patiala, Punjab, India

Published online: 01 February 2021

learning models. So, models are needed for processing, normalizing, integrating, and transforming the heterogenous multi-omics data to make it useful as a training set required for learning and analysis [7]. In bioinformatics, it is a major challenge to develop models using multi-omics data to predict clinical outcomes for improved diagnostics, prognostics, and therapeutics. Figure 1 represents the process of omics data analysis.

1.1 Motivation and Our Contribution

In healthcare research, there are tremendous opportunities and challenges due to the high availability of complex and unstructured omics data. So, a need is felt to study the literature of omics data broadly. The contribution and novelty of this review paper are also discussed.

- A comprehensive review has been conducted to discuss the role of existing techniques for improvement in omics data analysis.
- The research of existing techniques is categorised as omics data analysis, multi-omics data analysis, radiomics data analysis, metaheuristic algorithms for omics data analysis.
- The existing tools available to users for omics data analysis are reviewed.
- The existing techniques and tools are compared based on common characteristics. The complete review helps in the identification of future research directions for academic researchers and bioinformaticians.

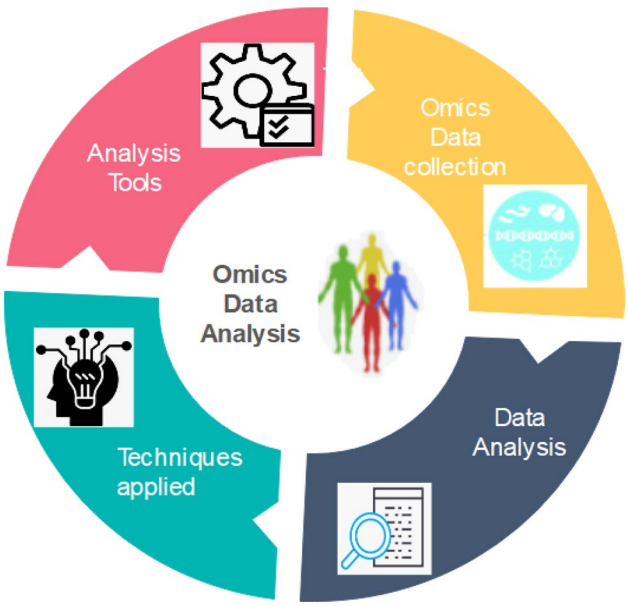

Fig. 1 Omics data analysis process

1.2 Related Surveys and Our Work

Some authors have done surveys discussing omics data in recent years. These surveys are summarized and are compared with our survey.

Antonelli et al. [8] present a survey of omics and image data along with integrated data analysis and databases used. Zhang et al. [9] offers a study of deep learning use in solving omics problems. The authors discussed the models of deep learning, the combined use of deep learning and omics in various research areas, and its existing opportunities and challenges. Li et al. [10] present a survey on integration techniques of omics and clinical data for analysis using machine learning methods. Rappoport and Shamir [11] review the clustering algorithms developed in machine learning for multi-omics data. Wei [12] presents a review of statistical methods used for integration and survival analysis of cancer data. Wu et al. [13] mainly focus on variable selection methods used for the integration of multi-omics data. Also presented a review of integrative analysis using supervised, semi-supervised, and unsupervised methods.

After analysis of the existing surveys, it has been noticed that mostly disease prediction and survival analysis for omics and multi-omics data have been discussed. There is a necessity to summarize the existing machine learning, deep learning, and metaheuristic techniques for omics, multi-omics, and radiomics data analysis which is further categorised based on disease prediction, disease recurrence, survival analysis, and biomarker identification. This survey integrates the existing research of techniques, tools and is an enhancement of existing surveys. Table 1 summarizes the comparative study of the proposed survey with existing surveys on omics data analysis.

1.3 Structure of Survey Paper

The survey paper has been organized into seven sections. Section 2 describes background with an overview of omics data analysis, integrative analysis, machine learning for predictive analysis. Section 3 discusses various research questions and review methods. Section 4 presents a systematic review of existing techniques for omics, multi-omics, and radiomics data analysis. The metaheuristic techniques used for omics data analysis are also reviewed and discussed. The techniques are compared and categorized based on common characteristics. Section 5 presents and compares various existing tools for omics data analysis. In Sect. 6, the review is summarized with open challenges and future research directions. Section 7 concludes the review and provide recommendations for future research.

Table 1 Comparison of our survey with existing surveys

Author [Ref.]	1	2	3	4	5	6	7	8	9	10	11	12
Antonelli et al. [8]	✓	✓	✓	✓	×	✓	✓	×	✓	✓	×	×
Zhang et al. [9]	✓	✓	×	✓	×	✓	×	×	×	✓	×	✓
Li et al. [10]	✓	✓	×	✓	×	✓	✓	×	✓	✓	×	✓
Rappoport and Shamir [11]	✓	✓	×	✓	×	✓	×	✓	✓	✓	×	×
Wei [12]	✓	✓	×	✓	×	✓	×	×	✓	×	×	×
Wu et al. [13]	✓	✓	×	✓	×	✓	×	✓	✓	×	×	×
Our survey	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

1—Omics data, 2—Integrative data(multi-omics), 3—Radiomics, 4—Disease prediction, 5—Disease recurrence, 6—Survival analysis,7—Biomarker identification, 8—Metaheuristic/optimization, 9—Machine Learning, 10—Deep Learning, 11—Ensemble techniques, 12—Tools

2 Background-Overview of Omics Data

The omics study started with genomics, which helps in studying genetic variants related to a complex and mendelian disease. The genomics focuses on whole-genome, which makes it different from genetics which studies single genes or individual variants. The omics field is getting huge importance as advancement in technology has made an efficient analysis of biological molecules possible. Omics data has four major types, i.e., genome, proteome, transcriptome, and metabolome [7]. In a cell, for the representation of molecules of DNA, a genome is used, for protein proteome is used. For the representation of molecules of RNA, transcriptome is used, and for the metabolite, the metabolome is used. Figure 2 shows the different omics data types of biological levels.

Genome: A genome provides complete information of an organism's DNA. In medical research, the main focus of genomics is to identify genetic variants related to disease, treatment response, and prognosis of a patient. Genome-Wide Association Study (GWAS) is a popular method used for the identification of genetic variants of complex disease in human populations [14]. Whole-genome sequencing (DNA-seq) is also a technique used for interrogation of DNA information, and it is used to assemble and discover genetic variants of a re-sequenced organism [7]. Sequence Read Archive (SRA) [15] and Gene Expression Omnibus (GEO) [16] are two databases having publicly available genomic data.

Proteome: In a cell, proteome describes the complete universe of proteins. The platform used for proteomic profiling is Mass-spectrometry. The MS-based methods are used for analysis and quantification of proteins and also used to analyse thousands of proteins in body fluids or cells [14]. To detect interactions among proteins, classic unbiased methods like yeast two-hybrid assays, phage display is used. ProteomeXchange [17], PRIDE [18], and ProteomicsDB [19] are the databases where proteome profiling is stored.

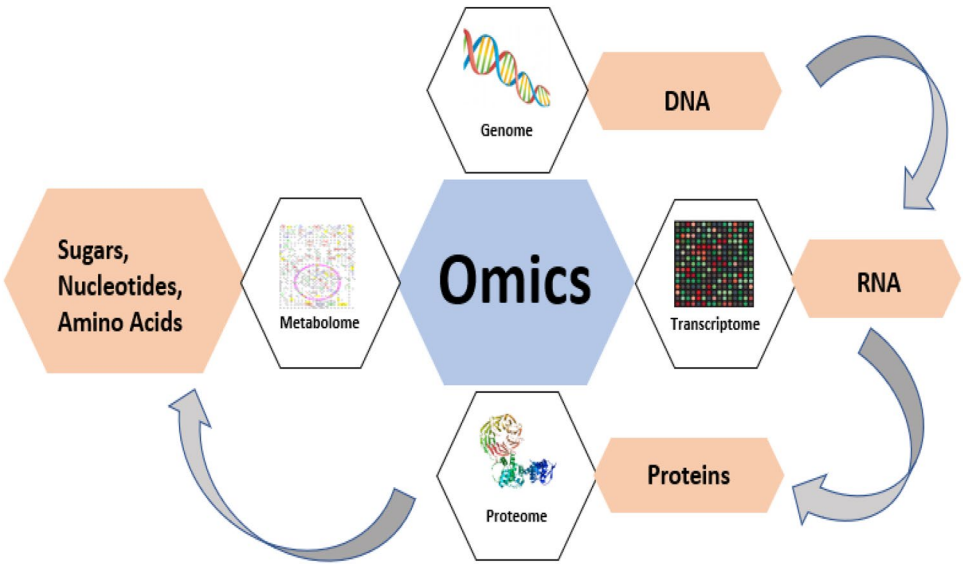

Fig. 2 Multi-omics data types of different biological levels

Transcriptome: In a cell, the transcriptome is the set of complete RNA molecules. In transcriptomics, both qualitative (presence of transcripts, novel splice sites identification) and quantitative (expression value of each transcript) examination of RNA levels genome-wide is done. RNA is an intermediate molecule between DNA and proteins, which is known as a primary functional readout of DNA [14]. RNA-seq and microarrays is a technique used for transcriptional profiling. Quantification of mRNA, identification of novel transcript, and novel splicing sites discovery is done with the help of raw data. GEO database [16], and SRA [15] are the databases having transcriptional profiling available publicly.

Metabolome: In an organism, the metabolome is a complete set of small-molecules types, like amino acids, carbohydrates, and fatty acids. Similar to proteome, quantification of metabolites is discovered with Mass-Spectrometry (MS) technology. Metabolic function is reflected by metabolite levels, and relative ratios and perturbations out of normal range indicate the disease. Quantitative measures of metabolite levels are used to discover novel genetic loci that regulate small molecules or relative ratios in plasma and other tissues [14]. The databases having metabolome data are very limited. MetaboLights [20] is one of the available datasets having metabolome experiments.

To effectively integrate the omics data types, the integrative method is required, which improves the analysis of omics data.

2.1 Integrative Analysis of Omics Data

The integrative analysis uses different data sources to understand the system in a better way. Single source omics data is used in many studies, but causes of complex traits are not explained in it. Figure 3 shows the difference between single level analysis and integrative analysis. It is found by researchers that analysis of single type data is not sufficient to understand the biological system as many levels are used for the regulation of the system [21, 22]. To get a better understanding of the biology of disease, the behaviour of molecules, and the interaction of various biological levels is needed. In recent years, with the increase in datasets of multi-omics data, many computation models and applications are developed to integrate multiple levels of data [23].

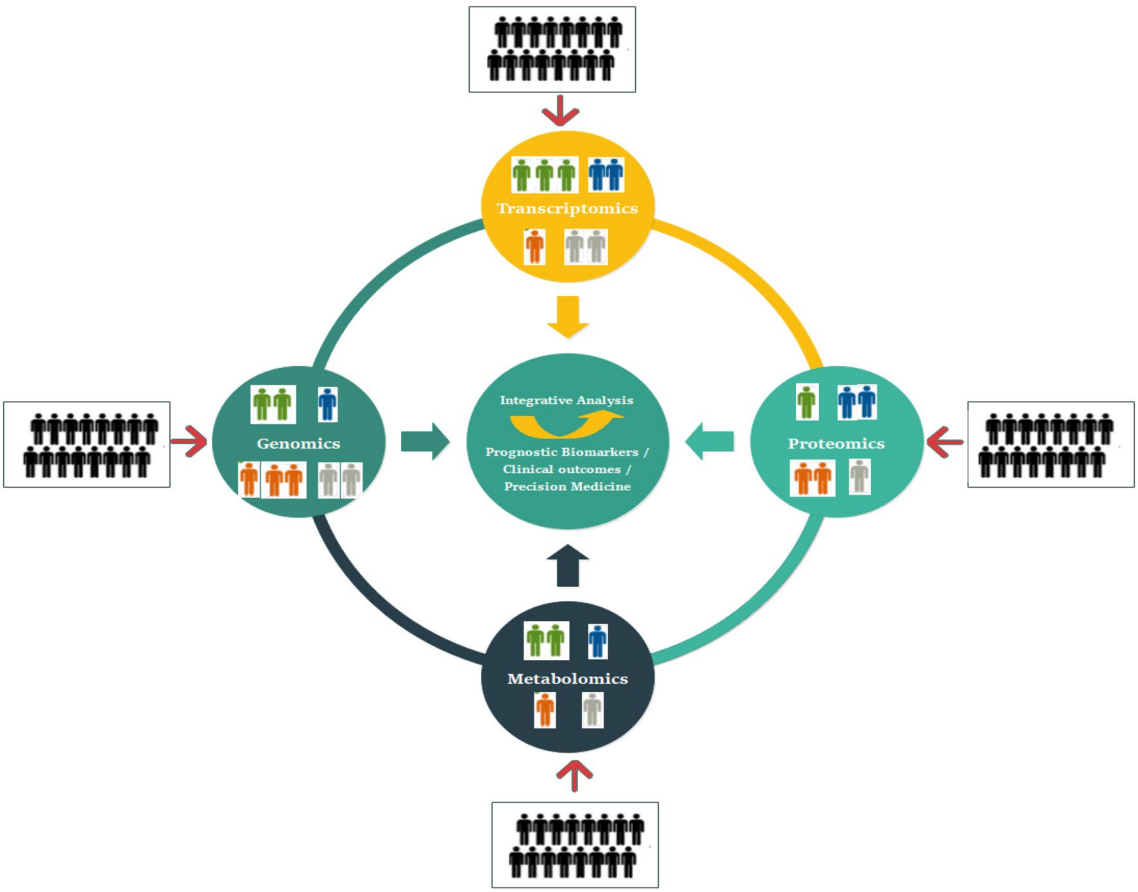

Fig. 3 Single level analysis and integrative analysis

In computational biology, integrative analysis plays a crucial role.

In the integration of omics data, genomic data provides information about mutation, expression, and regulation aspects of biology. Also, image features are integrated with genomic data by some researchers for predictive analysis [24]. The integrated data sources provide better performance than single data sources [25]. In an integrative analysis for simultaneous analysis, many methods are used to combine various data types. The methods for integration are divided into three main categories [26].

Concatenation based integration: In this method, multiple types of omics data are combined, and then the combined matrix of the dataset is used for analysis. In this method for the combined matrix, existing analysis methods for single omics data works appropriately.

Transformation based integration: In this method, firstly, the data type is transformed into a matrix of graph or kernel type. This intermediate form of data is merged to get integrative representation. This method is more robust compared to concatenation based integration because many different types of data such as categorical or continuous or sequence data can be integrated.

Model based integration: In this method, the data sets are analysed separately, and then the results are combined to get results. This model is very flexible as different models can be applied for different data types for analysis. In the field of bioinformatics, this model based integration is used extensively. This model approach is categorised into supervised and unsupervised depending on the models applied to individual data types. In supervised category, different types of data are used as training sets to generate multiple models. Then bagging or voting is used to combine these generated models. In unsupervised category, from different data types, the clustering results are obtained. Then the results are aggregated to conduct integration based on some optimization criteria.

Machine learning consists of various algorithms required for analysis, which leads to an effective prediction of omics data.

2.2 Machine Learning for Predictive Modeling and Analysis of Omics Data

Machine learning analytics is used in biology to deal with complex omics data and its integration. The pipeline for machine learning analytics for omics data is shown in Fig. 4. It consists of data preprocessing, modeling, and active learning [7].

Data pre-processing: To handle data effectively for analysis, firstly, normalization is done on multi-omics data. In the second step, to select the subset of features

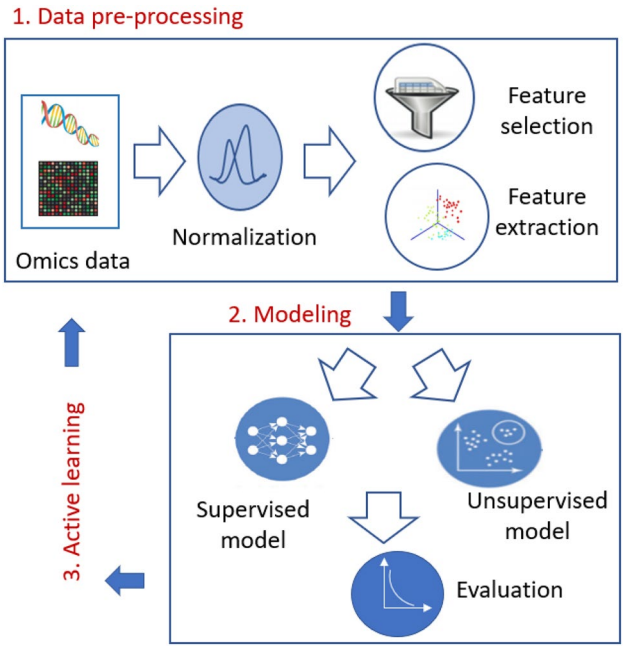

Fig. 4 Predictive analysis of omics data using machine learning

for modeling, feature selection is used. Pearson correlation coefficient and mutual information are supervised approaches, and Principal Component Analysis (PCA) is the unsupervised approach used for feature selection.

Modeling: In modeling, a model is built from training data with supervised or unsupervised learning, and then various criteria are used to evaluate the performance of the model. Supervised learning is a machine learning method that uses labeled data to infer a function. For prediction problems, several machine learning methods are used. Mainly regression based methods are used like Naïve Bayes, KNN, SVM, Neural network, and Ensemble method [27]. Amongst all methods, ensemble method is mostly used [7]. It occurs as multiple models perform better than a single model. Unsupervised learning is a machine learning method in which inference from data is drawn without the requirement of class labels [28]. Clustering is the most commonly used unsupervised learning method in various identification problems.

Active learning: After construction and evaluation of the model, uncertainty in the model has to be minimized. For this, active learning is applied, which guides about performing the next experiments. Active learning was used firstly for supervised settings, but recently it is being applied in unsupervised settings also [7]. Active learning is mainly used to get more accuracy for a machine learning algorithm having few labeled training instances. The complete taxonomy of omics data analysis is given in Fig. 5.

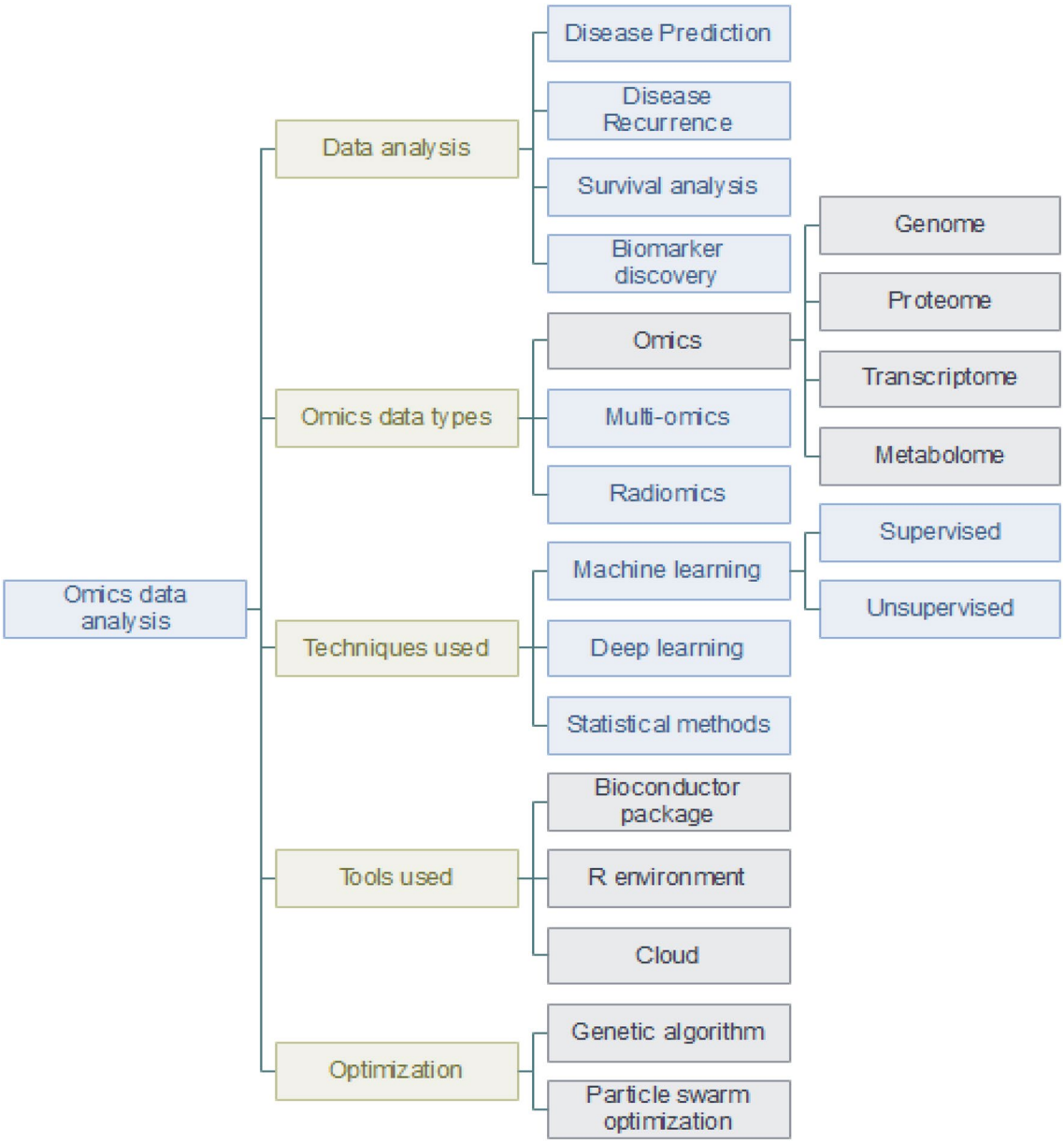

Fig. 5 Taxonomy of omics data analysis

2.3 Evolution of Research in Omics Data Analysis

In the last quarter of the nineteenth century, a variant of -oma originated as a -ome suffix [29]. The origin of -ome suffix firstly appeared in rhizome, sclerome terms derived from Greek words. The origin of genomics started with the introduction of the gene concept in 1909 by Johannsen. Also, genotype and phenotype terms coined by him. The term Genome was proposed in 1920 by Hans Winkler [30]. The Genome was previously coined in German as Genom. For an organism, the Genome provided the complete genetic makeup. In the late 1960s and early 1970s, using bacteriophage RNAs [31], the first sequencing results were attained on protein-coding genes. In 1970, the gene-cloning era began. In 1990, the era of genomic began with the sequencing of microorganisms [30]. The complete evolution of genomics is shown in Fig. 6. To show the trend of omics data analysis, Fig. 7 shows the number of publications per year. The research contribution of omics, multi-omics, radiomics using machine learning, deep learning, along with optimization, is shown in Fig. 8.

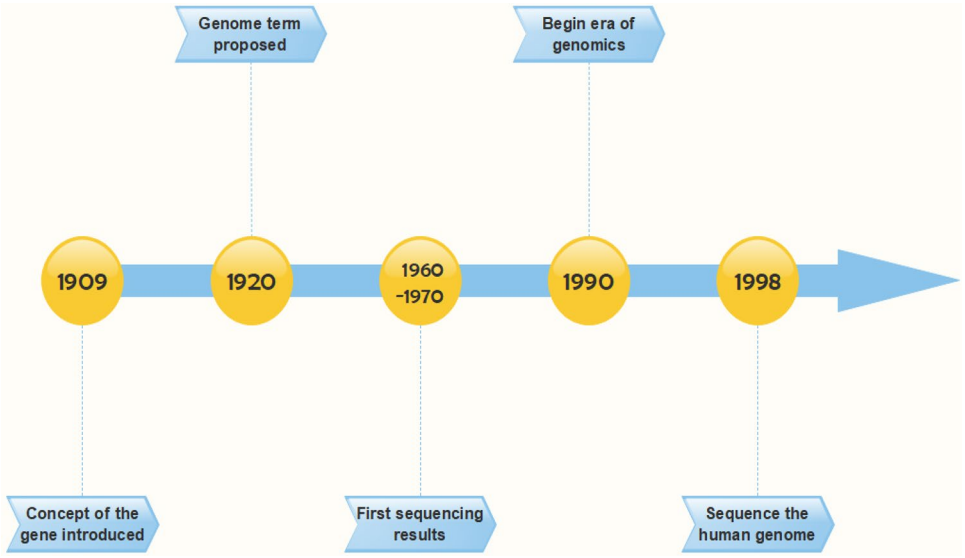

Fig. 6 Evolution of genomics

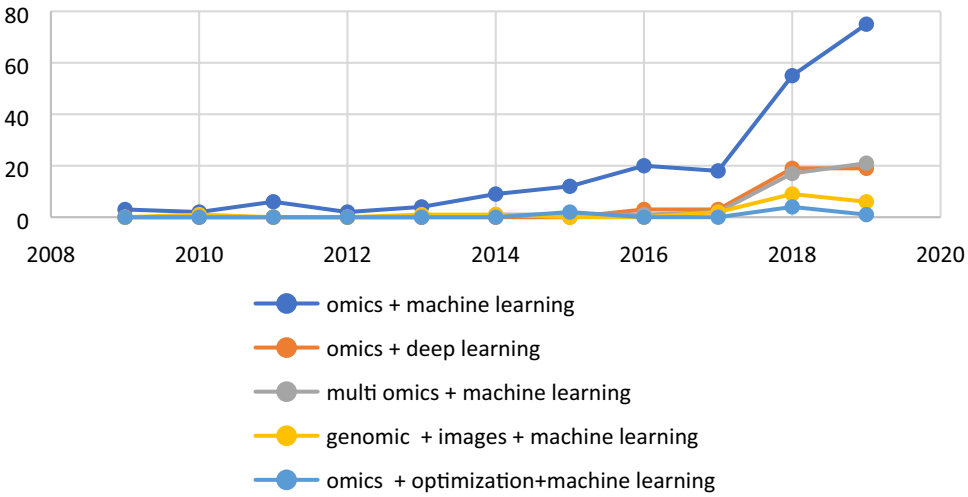

Fig. 7 Distribution of articles by the publication year in omics data analysis

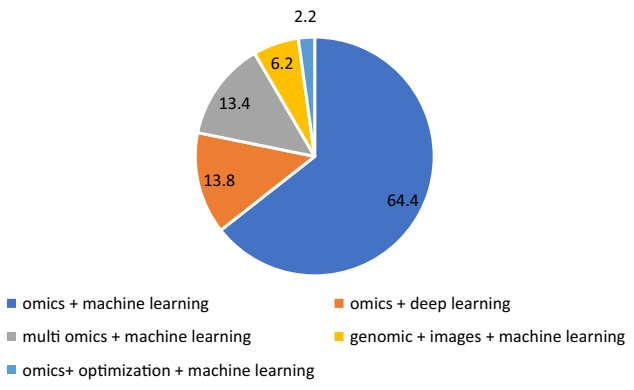

Fig. 8 Research contribution of publications in omics/multi-omics/radiomics/optimization

3 Review Method

A systematic review of techniques and tools used for omics data analysis is done by following the methodology of Kitchenham et al. [32] to summarize the existing work and to highlight the research gaps. The steps needed to conduct the review start with a list of research questions to be addressed, as given in Sect. 3.1. The main motive of this review is to discuss the latest developed techniques and tools for the analysis of omics data by answering the research questions. The search keywords are used in different databases to search the existing literature either electronically or manually exhaustively. Finally, a particular exclusion criterion is applied to refine the search process of data analysis.

3.1 Research Questions

This review provided the latest research work in the identification of techniques and tools by answering the research questions which are designed as given below.

RQ1: What are omics data and its types?
RQ2: What is an integrative analysis (multi-omics data)? Why is it needed?
RQ3: What are the techniques developed for omics, multi-omics, and radiomics data analysis?
RQ4: What is the significance of metaheuristic techniques in omics data analysis?
RQ5: What are the tools developed for omics data analysis?
RQ6: What are the open challenges and opportunities in the field of omics data analysis?

3.2 Sources of Information

To search articles from different publications, the various online electronic databases selected are Google Scholar, Springer, Science Direct, IEEE explore, Elsevier, Wiley online library, Web of Science. The above sources contain documents in several types, such as reviews, articles, book chapters, proceeding papers, and editorial material.

3.3 Search Criteria

The search criteria start with strings in the title "omics data", "omics data analysis", "techniques for omics data", "tools for omics data". Based on keywords over a time period (2004–2020), many search strings are formed as given below:

- "Omics data"
- "Omics data" + "analysis"
- "Omics data" + "analysis" + "technique"
- "Omics data" + "analysis" + "tool"
- "Omics data" + "analysis" + "machine learning"
- "Omics data" + "analysis" + "deep learning"
- "Omics data" + "analysis" + "optimization technique"
- "Omics data" + "analysis" + "cloud"
- "Genomic" + "images" + "machine learning"
- "Radiomic" + "analysis"

In the search, the research papers from various articles, journals, and conferences are included. The searching is done manually with keywords, as some articles do not have a search string in its abstract or title.

3.4 Data Inclusion and Exclusion Process

An extensive search is done to ensure the coherence of our search with the selection procedure explained in Fig. 9. In this review, 104 studies have been selected through the process of data inclusion and exclusion. The selection process starts with a search string, which also results in articles not relevant to the study. From Fig. 9, the process begins with 1158 research papers returned with search strings. The count is reduced to 607 with exclusion based on the title. Based on the relevant abstract, the count is reduced to 560. Then 230 articles were left after considering the full text. At final, for the literature review, 104 research articles are selected.

To analyse omics data is a complex task which can be tackled with various data mining and machine learning techniques as discussed in Sect. 4.

4 Omics Data Analysis Techniques Review

The use of machine learning for analysis of omics data in biomedical research is recent and ensure help in precision medicine. In the following sections, recent literature related to the use of machine learning, deep learning, metaheuristic techniques for omics data, multi-omics data, radiomics data is discussed. Also, tools developed for analysis of omics data are discussed.

4.1 Omics Data Analysis Using Machine Learning and Deep Learning

The medical research community is having a large amount of omics data available. But still, it is challenging for medical researchers to accurately predict the disease outcome. The analysis of omics data is done by researchers mainly for disease prediction, disease recurrence prediction, survival analysis or biomarker discovery. In the medical field, machine learning is becoming a popular method of analysis because of its capability to detect key features from complex datasets. For the development of prediction models, various machine learning techniques have been used. A multiple kernel method is used by Tao et al. [33] to classify subtypes of breast cancer among progesterone receptor, estrogen receptor, and HER2. The authors used TCGA dataset with cancer data of types methylation, mRNA, and CNV. From the results, it is concluded that multiple kernel performs better on three types of omics data compared to the use of single omics data. In the method, sequential minimization optimization technique is used to perform classification task. SVM is used by Liu [34] to develop a method of active learning for classification of cancer. The author used dataset of lung cancer, colon cancer, prostate cancer having gene expression data. From the results, it is concluded

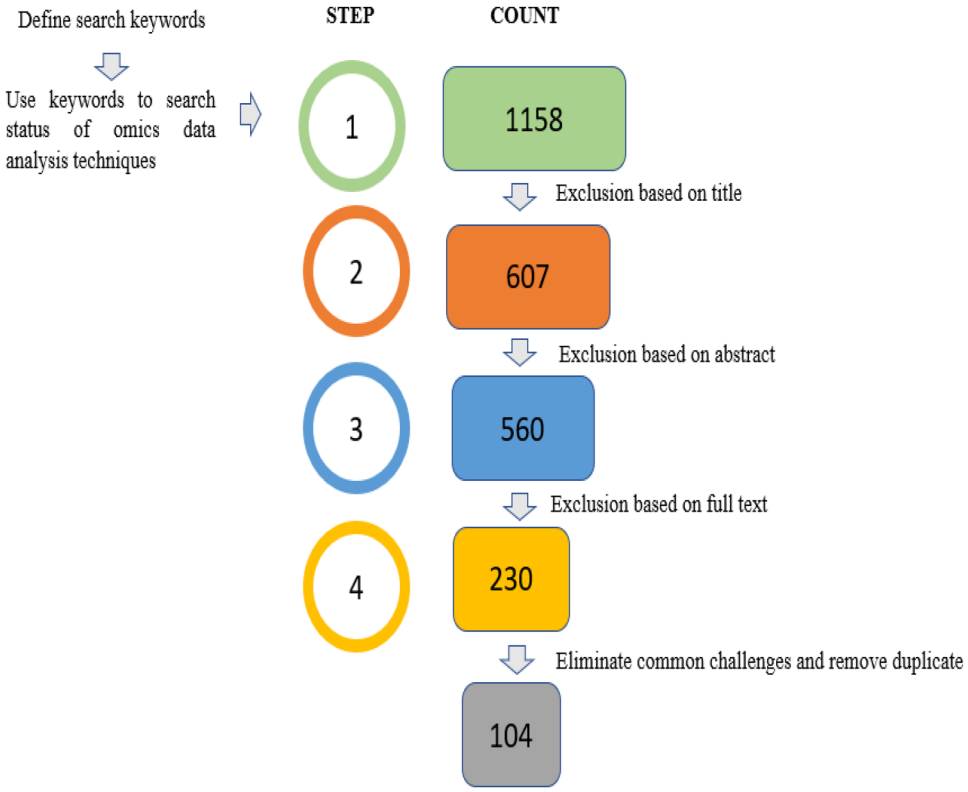

Fig. 9 Systematic review technique

that more accuracy is achieved by active learning approach compared to passive learning. The AUC value obtained by active learning is 0.81, which is high from 0.50 obtained by passive learning. In the proposed method, labelled instances of training data needed are required in very less number. Similarly, Xu et al. [35] predicted breast cancer prognosis with SVM-RFE method. The authors developed 50 gene signature on gene expression data by using leave-one-out evaluation technique. The 50 gene signature performed better in combination with SVM compared to previously developed 70 gene signature. To perform multiclass classification, SVM is used by Chen et al. [36] to identify information of somatic mutation from cancer patients. The authors used a database having tumor of lung, skin, liver, large intestine, and pancreas patients. The authors performed classification using predictors of type somatic mutation pattern, gene symbol, and chromosome. The results using database COSMIC and KEGG gives an accuracy of 0.70. For early prediction, Manogaran et al. [4] used a hidden Markov model with gaussian mixture cluster to diagnose cancer with a change in DNA copy number. The proposed model predicted an early stage of cancer and showed improved performance over existing segment neighborhood, binary segmentation, and PELT methods. Despite using single learning method, some studies develop ensemble methods which use multiple learning methods for better prediction. Anaissi et al. [37] used an ensemble method using a random forest method to develop ESVM-RFE (Ensemble SVM-Recursive Feature Elimination) approach to perform classification on leukaemia dataset. The proposed technique results in better classification accuracy on five microarray datasets over SVM-RFE existing method. Similarly, Cai et al. [38] used an ensemble method to classify lung cancer types, SCLC, SQCLC, and LADC. The authors used a machine learning method random forest to calculate ROC and mRMR using dataset. In feature selection, an ensemble method is used by following an incremental approach. For the dataset, for training, 28, 134, and 126 samples are taken for SCLC, SQCLC, and LADC, respectively, and for testing, 359 and 452 samples are taken for SQCLC and LADC, respectively. In the dataset, DNA methylation features are used for the classification purpose, and the proposed approach gives 86.54% accuracy. The recurrence prediction is done using network based algorithm developed by Ruan et al. [39]. The authors used random walk method which combines features of EC and DM genes. The algorithm used classifier SVM to predict the recurrence in cancer using DNA methylation data, and it gives high accuracy of prediction.

The biomarker discovery is identification of features that cause alteration in biological system. In machine learning, many feature selection methods are used for discovery of biomarkers. The feature selection using AUCRF, Vita, and Boruta method is done by Long et al. [40]. The author discovered novel signatures in colorectal cancer with a method

developed using combination of statistical method and transcriptomics data. The performance is evaluated using naïve Bayes, logistic regression, RF, and kNN model. The gene expression data in the RNA seq form is taken from TCGA and GTEx. The results show better performance of RF based on sensitivity, specificity, and accuracy parameters. Similarly, Bravo-Merodio et al. [41] used Elastic Net and LASSO methods for feature selection. A pipeline for identification of biomarkers related to datasets having omics and clinical data of high dimension is developed. The workflow firstly performs pre-processing of input omics data, and features which are highly related to biomarker candidates are identified. The authors performed tenfold cross-validation and evaluated the performance using AUC and ROC parameters for datasets having lipidomic and transcriptomic data. The results show identification of new biomarkers with proposed pipeline, which is beneficial for translational medicine. Moon and Nakai [42] used unsupervised method for feature selection and performed normalization using Box-Cox transformation. The authors combined gene expression and DNA methylation data to develop method for integrative analysis. The dataset of renal cell carcinoma is used for integration and identification of cancer candidate biomarkers. The results show improved performance for integrated dataset compared to datasets used alone. Hamzeh and Rueda [43] used filter-based feature selection for reduction of features count and wrapper-based feature selection for collection of genes having high accuracy in classification. The authors proposed method using machine learning for identification of cancer biomarkers. The proposed framework consists of DisGeNET database defining relations between genes and disease. The dataset of prostate cancer having 104 patients with RNA-seq data from NCBI is used for analysis. The proposed pipeline of machine learning helps in identification of biomarkers by using knowledge of literature. An Iterative Feature Elimination method, RGIFE, is proposed by Swan et al. [44] which is based on rule guided machine learning. RGIFE is applied to proteome and transcriptome data to find biomarkers for Osteoarthritis disease. The results show that RGIFE performs better on dataset compared to other feature selection methods and compared to dataset without use of any feature selection. Zuo et al. [45] proposed INDEED, which used partial correlation for building of sparse differential network. The biomarker discovery is done with integration of differential expression and differential network analysis. The hepatocellular carcinoma (HCC) having proteomic and glycomic data is used to discover biomarkers. The results show high prediction accuracy of HCC cases with biomarkers discovered using INDEED. Using breast cancer transcriptomic data with biomarker candidates selected by INDEED, the survival time prediction is highly accurate. Ramroach et al. [46] optimized cancer classification by using five algorithms for performance comparison on same dataset. Gene expression data of TCGA is used in the study for multi-class classification. Random forest, SVM, GBM, neural network, and kNN are the algorithms used. The results show that GBM, Random Forest, Neural network gives better classification accuracy out of which Random forest outperforms. Also, the biomarkers and potential drug targets are identified from 40 highest ranked genes of training set.

The survival prediction is the overall survival or specific survival outcome after diagnosis, treatment of disease. There are various machine learning methods that are used for survival prediction. For example, Artificial neural network is used by Ching et al. [47] to propose Cox-nnet framework. The model is implemented using Theano package in python. By using Cox-nnet in high-throughput transcriptomics data, prognosis prediction of a patient is done more accurately and efficiently. Cox-nnet is applied on TCGA dataset, and when compared with CoxBoost, Random Forests Survival, Cox-proportional hazards regression gives better survival prediction. An ensemble model is used by Roadknight et al. [48] to predict five-year survival rate with machine learning and anti-learning methods. The authors used SVM, J48 trees, Logistic regression, classification, and regression tree algorithms to create an ensemble model. For dealing with complex data, the authors used anti-learning method to select attributes using inverse ranking method. The authors worked on patients of TNM stage 2 and 3 and from results, it is concluded that ensemble learning method shows excellent improvement in survival prediction. A supervised method of dimension reduction is proposed by Spirko-Burns and Devarajan [49] to analyse omics data and to find outcomes of survivability. In proposed model, Continuum Power Regression (CPR) is integrated with AFT (Accelerated Failure Time) model. In an integrated framework, for dimension reduction, OLS, PLS, and PCR are used, and to handle survival data AFT is used. The proposed method, ACPR-AFT, gives better predictive performance compared to CPR-AFT when used on cancer genomic datasets, i.e., HNSCC, GBM, Ovarian cancer, and Oral cancer.

Deep learning is used for high-dimensional and larger datasets (e.g., RNA measurements, DNA sequencing) to capture internal structure in high-throughput biology [1]. Using deep learning models, the performance is improved over traditional models as high-level features are discovered with deep models. Deep learning provides additional understanding of biological data structure and increases the interpretability. Huang et al. [50] proposed SALMON (Survival Analysis Learning with Multi-Omics Network), an algorithm based on deep learning to predict survival in breast cancer. The proposed model implemented Cox proportional hazards regression networks and as a network input gene co-expression network analysis is used. In the proposed model, multi-omics data is used, and the model is compared with Cox-nnet, GLMNET, DeepSurv, and RSF. From the

results, it is seen that SALMON gives better concordance index and enhanced the survival prognosis accuracy. Lee et al. [51] proposed DeepHit, an approach for survival analysis by using deep neural network. The network used in DeepHit is trained with loss function, having both relative risks and survival times. The authors demonstrated that the performance of DeepHit is better when applied on UNOS, METABRIC, SEER and SYNTHETIC datasets compared with survival models such as Cox, RSF, MP-LogitR, MP-RForest, DeepSurv, MP-AdaBoost and Threshold Regression. Yousefi et al. [52] used risk propagation method and provided a framework SurvivalNet to interpret deep survival models. The authors also compared machine learning methods and deep survival models, which are Bayesian optimized for analysis of survivability. TCGAIntegrator is used to create datasets from TCGA for analysis. It is concluded that deep survival models give more accurate predictions of cancer outcomes. Chaudhary et al. [6] developed deep learning based model for prediction of survival in liver cancer. The model used methylation data, miRNA sequencing, RNA sequencing from TCGA dataset. The authors applied two steps for prediction. In first step, from the TCGA dataset, the labels of survival risk classes are obtained. Then in second step, SVM model is applied on the dataset. The accuracy of proposed model is evaluated using five different datasets. The proposed model identified features for survival of patients having liver cancer. The comparative analysis of the existing work done in omics data analysis is given below in Table 2.

Despite of single omics data type many researchers used multiple omics data types for analysis as discussed in Sect. 4.2.

4.2 Multi-omics Data Analysis Using Machine Learning and Deep Learning

Multi-omics or multiomics is an analysis approach in biology with datasets having many "omes". It consists of genome, transcriptome, proteome, metabolome, epigenome, and microbiome. Multi-omics is also termed as integrative omics as it integrates diverse omics data to study life in a concerted way. To fully understand the flow of information to relevant interactions or functional consequences from original disease cause (genetic, development, environment), multi-omics provide more opportunities compared to single omics interrogations. In literature, many researchers worked on multi-omics data as discussed.

For predictive analysis of multi-omics data, the researchers used machine learning models to work on integrated omics data types. For unsupervised integration of multi-omics data, Argelaguet et al. [53] proposed a computational method, Multi-Omics Factor Analysis (MOFA). For multi-omics data, MOFA is viewed as a generalization of principal component analysis (PCA). The proposed method is applied to patients having chronic lymphocytic leukaemia having data of DNA methylation and RNA expression. The major dimensions of disease heterogeneity are identified with this proposed method. Yan et al. [3] compared data integration algorithms, i.e., graph and kernel based to classify complex traits. In this model, seven different algorithms, i.e., composite association network, graph sharpening integration, semisupervised learning, Bayesian network, relevance vector machine, semi definite programming support vector machine, ada boost relevance vector machine are applied on two cancer datasets. In the results, it is concluded that kernel based algorithms perform better than graph based but use more computation time. A graph diffusion based method, i.e., NetICS is developed by Dimitrakopoulos et al. [54] which integrates different data types to prioritize cancer genes. The method is developed using network based integration of multi-omics data. In NetICS, integrated data is of different types, i.e., copy number variations, somatic mutations, miRNA expression, and methylation. The cancer genes are predicted by proposed method using five cancer types from TCGA dataset. Sun et al. [55] developed a method GPMKL to predict the outcome of breast cancer by integration of genomic data and pathological data. The method uses multiple kernel learning. The authors used heterogeneous data, i.e. gene expression, gene methylation, protein expression, and copy number alteration for analysis. From the results, it is found that for predicting patients with breast cancer, along with genomic data, pathological data also play an efficient role. A semi-supervised algorithm based on graph is presented by Torshizi and Petzold [56] for classification. The method follows detection of condition-responsive genes from biological pathways to construct graph. The proposed method is applied on ovarian cancer dataset having both DNA methylation and gene expression data. From the experimental results, it is concluded that proposed approach performs classification better on integrated data when compared with SVM and ANN. Bayesian model is used by Fang et al. [57], Gevaert et al. [58], Subhani et al. [59]. Fang et al. [57] proposed a full Bayesian model to work with multi-omics data having missing samples. A self-learning cross-validation (CV) decision scheme is also proposed to decide how the samples with missingness are incorporated. From the simulation results, on child asthma dataset shows improvement in accuracy prediction and feature selection. Similarly, Gevaert et al. [58] predicted breast cancer prognosis using Bayesian networks. The authors used clinical and microarray data in an integrated form for analysis. The data is integrated using partial integration, decision integration, full integration. This integrated data is used for classification of breast cancer patients. The results show partial integration method as best compared to other methods. Similarly, Subhani et al. [59] developed a model using a

Table 2 Comparison of work done in omics data analysis (A: disease prediction, B: disease recurrence, C: survival analysis, D: biomarker discovery, E: ensemble technique, ML: machine learning, DL: deep learning)

Author [Ref.]	ML/DL	Method	Dataset used	Type of analysis					Performance parameters	Evaluation/contribution	Future directions
				A	B	C	D	E			
Tao et al. [33]	ML	Random Forest, Neural Network, SMO-MKL (Sequential Minimal Optimization -Multiple Kernel Learning)	Breast cancer patients having methylation, CNV, mRNA data from TCGA dataset	✓	×	×	×	×	Accuracy, AUC	SMO-MKL performed better than Random Forest and Neural network	Can perform analysis using combination of pathway and omics data
Liu [34]	ML	SVM, Active learning	Prostate cancer, colon cancer, lung cancer from http://sdmc.lit.org.sg/, GEDataset (gene expression data)	✓	×	×	×	×	Accuracy, AUC	Active learning using SVM outperformed passive learning	–
Xu et al. [35]	ML	SVM, KNN	Gene expression signature of 295 breast cancer patients from the Netherlands Cancer Institute (fresh-frozen–tissue bank)	✓	×	×	✓	×	Accuracy, specificity, sensitivity, AUC	SVM-RFE model identified a 50-gene signature that gives 35%,3%,48% high accuracy, specificity, sensitivity compared to 70 gene signature	Clinical treatment and carcinogenesis can be understood by further analysis of 50 gene signature
Chen et al. [36]	ML	SVM	Somatic mutation data of cancer from COSMIC database (Lung, liver, pancreas, skin, large intestine cancer)	✓	×	×	×	×	Accuracy, precision, recall, F-measure	Performed multiclass cancer classification with accuracy 0.62, Precision 0.75, Recall 0.60, F measure 0.60	Can perform cancer classification using deep learning or other machine learning methods
Manogaran et al. [4]	ML	Bayesian Hidden Markov Model (HMM), PELT, binary segmentation, segment neighborhood	DNA data from sample PA.C.Dan.G	✓	×	×	×	×	Accuracy, error, correctly predicted, miss predicted	Proposed HMM using GM clustering outperforms PELT, segment neighborhood, binary segmentation	–
Anaissi et al. [37]	ML	ESVM-RFE (Ensemble SVM-Recursive Feature Elimination, Random Forest	Microarray data of Colon cancer, breast cancer, NCI dataset, childhood leukaemia dataset (Children's Hospital at Westmead)	✓	×	×	×	✓	AUC accuracy	Using proposed ESVM, accuracy increase is 9%, 5% compared to SVM-RFE, and random forest, respectively	Can collect more genomic data relevant to an individual patient for analysis Can thoroughly evaluate clinical patterns using data analysis
Cai et al. [38]	ML	Random forest	Lung cancer with DNA methylation from cancer genome atlas and gene expression omnibus	✓	×	×	×	✓	Accuracy, precision, recall, F-score	Incremental feature selection of ensemble performs better for classification and accuracy of 86.54 and recall of 84.37 is achieved	–

Table 2 (continued)

Author [Ref.]	ML/DL	Method	Dataset used	Type of analysis					Performance parameters	Evaluation/contribution	Future directions
				A	B	C	D	E			
Ruan et al. [39]	ML	SVM	Endometrial cancer specimens collected from ongoing work	×	✓	×	×	×	AUC, kappa, specificity, and accuracy	Proposed network based approach and SVM gives better accuracy	Curated pathways can be integrated with computationally derived subnetworks for analysis A classification model without any dependency on DM genes can be developed
Long et al. [40]	ML	Random Forest, naïve Bayes, logistic regression, kNN	Colorectal cancer patients of GSE83889, GSE41258, GSE44861 dataset taken from GEO	×	×	×	✓	×	Accuracy, specificity, sensitivity	Random Forest gives 0.998 accuracy, 0.999 specificity, and 0.998 sensitivity	The proposed signature can be validated for clinical settings
Bravo-Merodio et al. [41]	ML	LASSO and Elastic Net methods of feature selection	Microarray expression data of pancreatic cancer and leukaemia, lipidomics data from Cambridge Baby Growth Study	×	×	×	✓	×	ROC AUC	The proposed pipeline helps in identification of relevant biomarkers to be used in translational medicine	Future translational projects can be guided by using single cell omics in proposed pipeline
Moon and Nakai [42]	ML	Naïve Bayes, SVM, random forest, logistic regression	Gene expression and DNA methylation data from KIPAN dataset (TCGA-Broad GDAC Firehose)	×	×	×	✓	×	F1 score, accuracy, MCC, AUC	Proposed method of integrative analysis and feature extraction by unsupervised method performs better than individual dataset	Analysis of multi-omics dataset can be done by using proposed method
Hamzeh and Rueda [43]	ML	SVM-RBF(Support Vector Machine with Radial BasisFunction Kernel), Random Forest, Naïve Bayes, kNN	Prostate cancer patients (104) having RNA-Seq data from NCBI database	×	×	×	✓	×	Accuracy, Sensitivity, ROC	The proposed model with filter and wrapper based feature selection identified related genes having good classification accuracy	At feature selection phase deep learning can be used
Swan et al. [44]	ML	SVM, random forest, knn, JRip, naïve Bayes, Iterative feature elimination (IFE)	Osteoarthritis transcriptome data from NCBI GEO and synovial inflammation proteome data	×	×	×	✓	×	True positive rate, true negative rate	The proposed rule guided IFE model identified biomarkers with TPR = 96% and TNR = 99%	Feature reduction method can be used in future for biomarker discovery with more availability of joint inflammation dataset

Table 2 (continued)

Author [Ref.]	ML/DL	Method	Dataset used	Type of analysis					Performance parameters	Evaluation/contribution	Future directions
				A	B	C	D	E			
Zuo et al. [45]	ML	Cox regression model, logistic regression, LASSO method	HCC cases from TU cohort (Egypt), GU cohort(Washington), breast cancer data from PRECOG website	✓	×	✓	✓	×	Accuracy, AUC	INDEED approach helps in discovery of biomarker candidates with improved accuracy	INDEED can be shared with the scientific community using developed R package. It can extend for biomarker discovery using integration of multi-omics data
Ramroach et al. [46]	ML	RF, GBM, NN,kNN, SVM	Gene expression dataset of TCGA	✓	×	×	✓	×	Accuracy, F1 score, AUROC	Random Forest outperforms for diagnosis of cancer	In future, testing of the model and for building ML models CTC gene expression dataset can be used
Ching et al. [47]	ML	Artificial neural network, Random Forest Survival, CoxBoost, Cox-PH	TCGA dataset (Breast, Bladder, Brain, Head and neck, liver, lung, stomach, ovarian, kidney cancer)	×	×	✓	×	×	C- index, IPCW score, log ranked p-value, Brier Score	Proposed Cox-nnet performed better than Cox-PH	Can analyse performance of Cox-nnet using different neural networks
Roadknight et al. [48]	ML	Logistic regression, SVM, J48 tree, classification, and regression tree	TNM stage data from City Hospital, Nottingham	×	×	✓	×	✓	accuracy	Proposed ensemble gives high accuracy and inverse ranking is a better method for attribute selection in anti-learning	Correlated Activity Pruning method and multiple kernel learning can be used to improve learning
Spirko-Burns and Devarajan [49]	ML	Continuum power regression	Glioblastoma and Head & Neck squamous cell carcinoma from TCGA, Oral cancer, Ovarian cancer	×	×	✓	×	×	AUC, Youden index	Proposed ACPR-AFT model outperforms CPR-AFT model	In analysis, variable importance projection threshold value can be investigated in future
Huang et al. [50]	DL	Neural network	583 breast cancer patients from GDAC Firehose	×	×	✓	×	×	Concordance index, log-rank test	Proposed SALMON using Cox-PH model	–
Lee et al. [51]	DL	Deep neural network, Cox-PH, RSF, MP-RForest, MP-AdaBoost, MP-LogitR, DeepSurv	METABRIC, UNOS, SEER database	×	×	✓	×	×	Time dependent c-index	Proposed DeepHit using loss function outperforms other models	–
Yousefi et al. [52]	DL	Bayesian optimization, Deep survival model	TCGA genome and clinical data using TCGAIntegrator	✓	×	✓	×	×	Accuracy	Used bayesian optimized with technique of risk propagation and deep learning	–

Table 2 (continued)

Author [Ref.]	ML/DL	Method	Dataset used	Type of analysis					Performance parameters	Evaluation/contribution	Future directions
				A	B	C	D	E			
Chaudhary et al. [6]	DL	Auto-encoder, SVM	Liver cancer of TCGA dataset	×	×	✓	×	×	Brier score, log-rank *p*-value, c-index	Proposed framework using deep learning	The model can be improved by collaboration with clinicians

Bayesian network model of machine learning for data integration of datasets containing clinical and genomic data. The authors performed classification using neural network approach and Bayesian approach of machine learning. FusionGP method is proposed by Savage and Yuan [60] to predict chemoinsensitivity. In this method, data fusion is done for breast cancer having copy number alteration, gene expression, and digital pathology image data. From the experimental results, it is found that proposed method performs better in classification than SVM, random forest, and logistic regression model. So, it is concluded that FusionGP uses heterogeneous data to select informative features for predicting outcome of treatment and disease. Kim et al. [61] proposed a method based on machine learning to predict cancer prognosis. The method firstly identifies biomarker genes that are needed for cancer prediction. In proposed method, GANs model (generative adversarial networks) of graph learning is used to specify the candidate genes, and PageRank algorithm is used to calculate gene scores. Five cancer types having omics data, including DNA methylation, gene expression, copy number, somatic mutation is used for prediction. The experimental results show that proposed model performs better than WGCNA (Weighted gene co-expression network analysis), CPR(Clustering and PageRank) methods, and also multi-omics data leads to improved prediction accuracy. Bica et al. [62] introduced a cross modal neural network architecture used for integration of multi-omics datasets when training examples of a limited number are available. The proposed model finds out the most relevant genes and influence of different types of omics on each other. The proposed model gives an accuracy of 99% on omics datasets. Priority-lasso is proposed by Klau et al. [63] for prediction analysis using multi-omics data. The proposed model is simple and fast, and it inherits all the properties of Lasso. In priority-lasso, irrespective of less priority, the variables of blocks with high priority are exploited. The leukaemia dataset is used for analysis, and in results, priority-lasso model gives better accuracy than lasso model in prediction. For clustering of multi-omics data, Rappoport and Shamir [64] and Wu et al. [65] developed methods. Rappoport and Shamir [64] proposed NEMO (Neighborhood based multi omics clustering) algorithm for the clustering of multi-omics data to find cancer subtypes. NEMO is applied to TCGA dataset with 3168 patients having omics data of methylation, gene expression and miRNA expression type for ten cancer types. From the experimental results, it is found that NEMO performs better for both full and partial datasets compared to an existing clustering algorithm, i.e., PVC. Similarly, Wu et al. [65] proposed an integrative model using low-rank approximation for clustering of multi-omics data. The algorithm is applied on TCGA dataset having cancer patients of gene expression, copy number variation, somatic mutations, and DNA methylations data types. From

the results, it is concluded that LRAcluster performs better than other existing methods. Ensemble strategy is applied by Lopes et al. [66] for classification based on distinct feature selection and logistic regression for TCGA dataset having RNA-seq and clinical data. The classification of Triple-Negative Breast Cancer is done using proposed strategy. Hybrid of feature selection method and classification model is used by Chang et al. [67] to develop prognostic model. The model is developed for oral cancer, having genomic and clinicopathologic features. The authors used SVM, ANN, ANFIS, and logistic regression as classifiers and GA, CC (Pearson's Correlation Coefficient), Relief-F, CC-GA, and ReliefF-GA as feature selection methods. The results show outperformance of ReliefF-GA-ANFIS model based on accuracy for oral cancer prognosis. Deep learning algorithms for multi-omics data analysis are used by Zhang et al. [2], Liang et al. [68], Islam et al. [69]. Zhang et al. [2] adopted Autoencoder, a deep learning algorithm for integration of multi-omics data and identification of two subtypes having significant survival differences, K-means clustering is used. The Autoencoder based classification gives better results for classification and prognostic subtypes identification in neuroblastoma compared with PCA, iCluster, and DGscore. Liang et al. [68] proposed a multimodal deep belief network to analyse integrated data from multiple platforms. In multimodal DBN, contrastive divergence (CD) algorithm is used for inferring parameters. The proposed model is used on two cancer datasets, i.e., breast cancer and ovarian cancer, to identify different subtypes and to predict survival time of patients. The results show that multimodal DBN is an effective approach to analyse big data problems. Islam et al. [69] demonstrated deep learning models for prediction of disease and its subtypes for omics data. As datasets are increased in size and samples are reduced in number, there is a limitation in using machine learning algorithms such as SVM and random forest. The authors applied deep learning model, i.e., DNN, for classification of breast cancer subtype. Also, multi-omics integrated with deep learning based DNN model results in an improvement to predict breast cancer subtypes.

The recurrence prediction of oral cancer is done using development of decision support system by Exarchos et al. [70]. The proposed system helps in integration of data of different types such as clinical, genomic, and image data. The authors firstly identified factors of progress of cancer using individual data source for 41 patients. The individual results are combined to discriminate between patients of cancer having and not having recurrence. Similarly, Park et al. [71] used a graph regularization approach and developed semi-supervised learning algorithm to predict recurrence of cancer. The graph structure is built from gene expression data, and informative gene set is done by integration of protein data with gene expression data. The authors used three datasets of Breast cancer, Colorectal cancer, and Colon cancer from GEO (Gene Expression Omnibus) database. The results show outperformance of proposed method compared to SVM, Random Forest, Naïve Bayes as supervised algorithms and TSVM as semi-supervised algorithm.

In biomarker discovery of multi-omics data, various methods are proposed by researchers. Kim et al. [72] proposed meta-analytic support vector machine (Meta-SVM) for detecting genes associated with disease by accommodating multi-omics data. The proposed method is applied on breast cancer data of TCGA and lung cancer having mRNA expression. The results show that Meta-SVM identifies potential biomarkers for multi-omics data effectively. Similarly, Mallik et al. [73] used minimal redundancy and maximum relevance to develop a framework for identification of epigenetic biomarkers for multi-omics data. The proposed framework decreases false positive rate and increase positive predictive rate for identification of biomarkers using Prostate Carcinoma dataset and Leukemia dataset. Similarly, Long et al. [74] identified and validated candidate biomarkers of pancreatic cancer using translational model and statistical learning on integrated omics data. For diagnosis and prognosis at an early stage, the candidate biomarkers are identified by using random forest model on gene expression data.

For survival analysis, El-Manzalawy [75] proposed a multi view feature selection algorithm to select discriminative features jointly from multi-omics data. The proposed method is based on statistical method, i.e., canonical correlation analysis (CCA). The method is applied to predict the survival of kidney renal clear cell carcinoma (KIRC) using gene expression, copy number alteration, RNA-seq data. In another work, Ma et al. [76] developed predictive models for predicting ten-year survival using omics data of breast cancer patients. Seventy-seven patients having data types, i.e., proteome, gene expression, phosphoproteome, and copy number variation, are used in the study. SVM (with linear, RBF kernel, polynomial kernel), random forest, and bayesian logistic regression are the predictive models used for conducting experiments. From the experimental results, it is found that using proteome data to build predictive model gives high accuracy of 0.725 compared to other omics data types. ANN models are developed by Chen et al. [77] for risk prediction of lung cancer survival in various laboratories. The authors used NSCLC cases having clinical data combined with gene expression data taken from NCI database for analysis. The results show 83% accuracy with prediction model applied on gene expression data. A multi-omic kernel machine learning method is proposed by Zhu et al. [78] to get the prognostic values of multi-omics profiles individually and with combination of clinic data for cancer. The results of integration of omics data with clinical give an improvement in prognostic performance. Kim et al. [79] proposed a framework to do tasks such as prediction of

survival data, integration of omics data, and identification of features interactions linked with survival. The authors developed a tool ATHENA which uses GENN, GESR two modeling methods, and it is used to perform feature selection, modeling, and interpretation in translational bioinformatics. The validation of proposed framework is done by taking TCGA dataset of breast cancer having 476 patients. Poirion et al. [80] used an autoencoder algorithm to develop an unsupervised pipeline for multi-omics integration. The authors used TCGA-Assembler to download, assemble, and process data of bladder cancer from TCGA portal. For analyzing and computing survivability, survival package and survdiff function are used. The proposed deep learning based pipeline predicts the survival subtype of new sample very effectively. A pipeline is developed by Dubourg-Felonneau et al. [5] using deep learning to integrate data from different sources for prediction of omics data. The success of proposed approach is demonstrated by predicting survival of breast cancer patients with integration of TCGA microarray, METABRIC microarray, and TCGA RNA-seq datasets. The comparative analysis of the existing work done in multi-omics data analysis is given below in Table 3.

Despite only genomic information, the availability of valuable information from medical images encouraged development of methods for data analysis as discussed in Sect. 4.3.

4.3 Radiomic Data Analysis/Omics Data Analysis Using Imaging Features

Radiomics is referred as extraction of features from medical images collected using MRI (magnetic resonance imaging), CT (computed tomography), US (ultrasonography), and PET (positron emission tomography) to find correlation within features and diagnosis of disease. The statistical and mathematical methods are used comprehensively on data extracted from medical images. In radiomics initially acquisition and pre-processing of data is done. In next step, segmentation is done to extract features which lead to discovery of knowledge and modeling. The advances in machine learning and data mining technology help in getting minable data from quantitative features extracted from medical images. The radiomics has gained attention in research for treatment and diagnosis. The prediction accuracy can be improved by using radiomics data alone or combined with other data. In literature, Incoronato et al. [81] reviewed various analysis methods used for radiomics and genomics. In radiomics, image biomarkers are created that helps in identification of genomics of a disease such as cancer. The authors defined that in various types of cancer, with an increase in imaging and genomic data amount, an integrated approach is encouraged to characterize tumor and for precision medicine. Although radiogenomics enhanced the diagnostic accuracy, still it requires imaging and genomic protocols standardization, image acquisition, and post-processing. Also, Acharya et al. [82] discussed the workflow of radiomics in which four stages are used. In first stage images are acquired, second stage is segmentation where distinct features are collected, third stage is feature extraction in which useful information is extracted, fourth stage is analysis in which algorithms are used for prediction. The authors also discussed the use of images like CT, MRI, and PET in radiomics to get results in various studies. The authors concluded that the use of imaging biomarkers derived from radiomics helps in getting more accurate and predictive analysis. Li et al. [83] predicted molecular classification of breast cancer with the help of tumour phenotypes extracted from magnetic resonance images. The authors used TCGA Assembler software to download clinical, molecular classification, pathology data from TCGA data portal. The receiver operating characteristic is used to evaluate the performance of the classifier model for molecular subtyping. From the analysis, it is concluded that to discriminate breast cancer subtypes, computer-extracted image phenotypes show promising results.

For radiomics data analysis, the machine learning method SVM is used by Zhou et al. [84], Takahashi et al. [85], Li et al. [86]. Zhou et al. [84] diagnosed distant metastasis in lung cancer with CT radiomic features. The clinical and radiomic features of 348 patients from a hospital in Sichuan University are extracted in the study. The effective features are selected using FSV (Feature selection via concave minimization). The developed SVM with Stochastic gradient descent model is applied for classification. In results, the use of only radiomic features gives AUC of 72.84%, and with clinical features, it is increased to 89.09%. It is concluded that diagnosis of distant metastasis is done effectively using radiomic features. Similarly, Takahashi et al. [85] predicted glioma grading using radiomic analysis of imaging type diffusion kurtosis and tensor. The 55 datasets are used with 504 significant features selected using recursive feature elimination for analysis. The results show AUC of 90% using logistic regression and 93% using support vector machine. It is concluded that with diffusion kurtosis imaging and machine learning, the glioma grade is predicted with high accuracy. Similarly, Li et al. [86] predicted survival and classification of gene expression status of pancreatic ductal adenocarcinoma using computed tomography (CT) radiomic features. The data of 111 patients is used for the study. Deep learning methods and log-rank test help in extraction of radiomic features. Support vector machine classified gene expression levels using accuracy, AUC, sensitivity, and specificity. The results show an AUC score of 0.91 and 0.90 for C-MYC and HMGA2 gene expression status prediction. The decision tree is used for data training by Clifton et al. [87] to assess tumour segmentation by utilizing radiomic features. The study used 25-pixel based texture features of 8 patients

Table 3 Comparison of work done in multi-omics data analysis (A: disease prediction, B: disease recurrence, C: survival analysis, D: biomarker discovery, E: ensemble technique, ML: machine learning, DL: deep learning)

Author [Ref.]	ML/DL	Method	Dataset used	Type of analysis					Performance parameters	Evaluation/ contribution	Future directions
				A	B	C	D	E			
Argelaguet et al. [53]	ML	Probabilistic Bayesian framework, Cox regression model	Chronic Lympho-cytic Leukaemia	✓	×	✓	×	×	Accuracy, Harrell's C-index	Proposed MOFA method performs unsupervised integration with various modalities	In omics data type, pathway database can be used In diagnosis, the uncertainties can be estimated by making use of comprehensive Bayesian treatment
Yan et al. [3]	ML	Composite Association Network, Graph Sharpening Integration, Bayesian network, RVM, SVM, Ada boost RVM	GAW 19, Ovarian cancer and breast cancer (TCGA project) having methylation, gene expression, miRNA, protein expression data	✓	×	×	×	×	Accuracy, AUC, F1 score	RVM, Ada-Boost RVM, Composite Association Network are better integration algorithms	The algorithms can be combined to develop an ensemble classifier
Dimitrakopoulos et al. [54]	ML	Fishers method	Five cancer types from TCGA dataset	✓	×	×	×	×	AUC, *p*-value	Proposed NetICS method performed better prioritization of cancer genes	In integration, mutational patterns of complex type can be used Data can be combined at differentially expressed gene level
Sun et al. [55]	ML	Multiple Kernel Learning	Breast cancer (TCGA dataset)	×	×	✓	×	×	Mathews correlation coefficient, accuracy, sensitivity, specificity, precision, c-index	GPMKL method gives better AUC of 82.1% for integrated data compared to AUC of 79.4% for genomic and AUC of 68.1% for pathological data	Subtype prediction can be done using GPMKL Deep learning methods can be applied for multimodal feature extraction and data fusion in GPMKL
Torshizi and Petzold [56]	ML	Graph based Semi-supervised algorithm, ANN, SVM	Ovarian cancer from the Human Genome Atlas (TCGA) having DNA methylation and gene expression	✓	×	×	×	×	AUC, Error rate	Proposed graph based SSL method of data integration outperforms in classification	The proposed method's performance can be boosted by use of transcriptome, epigenetic, and biological knowledge The performance of proposed method can be enhanced by taking different measures of statistical correlation

Table 3 (continued)

Author [Ref.]	ML/DL	Method	Dataset used	Type of analysis					Performance parameters	Evaluation/ contribution	Future directions
				A	B	C	D	E			
Fang et al. [57]	ML	Bayesian model	DNA methylation data of Child Asthma dataset from Children's Hospital of Pittsburgh	✓	×	×	×	×	AUC, sensitivity, RMSE	Full Bayesian model with missingness is proposed to incorporate missing data along with feature selection and model prediction	In omics applications, to increase computing speed, parallel computing in FBM can be employed
Gevaert et al. [58]	ML	Bayesian network	Breast cancer data from ITTACA database	✓	×	×	×	×	AUC	Partial integration of microarray and clinical data gives better AUC of 0.845	The approach can be validated upon availability of more public data
Subhani et al. [59]	ML	Bayesian network, neural network	Colorectal cancer dataset	✓	×	×	×	×	–	Developed data integration model for clinical and genomic data	The schema of model can be implemented in SAP HANA environment using public datasets. Also, meta-dimensioanl data model can be proposed
Savage and Yuan [60]	ML	FusionGP, Random Forest, a stepwise logistic regression and SVM	Gene expression, pathology image, copy number alteration data from META-BRIC breast cancer data	✓	×	×	✓	×	AUC	Proposed FusionGp for clinical outcome prediction which performs better than SVM and stepwise GLM	In different data types, complex structures need to be encoded The model which can learn new latent features is needed to be developed In FusionGP, fast approximations can be done to make it more valuable
Kim et al. [61]	ML/ DL	GANs model (Generative Adversarial Networks)	Cancer data from TCGA dataset having CNV, DNA methylation, mRNA, SNP data types	✓	×	×	✓	×	AUC, *p*-value	The proposed method gives more accurate prediction of cancer	–
Bica et al. [62]	ML	Superlayered Neural Network Architec-ture (SNN), RNN, MLP	Cancer dataset from TCGA	✓	×	×	×	×	Accuracy, MCC, precision, sensitivity, F1 score	Proposed cross modal neural network to integrate multi-omics data	In regression problem, to predict genes expression level and for survival time analysis of cancer patients SNN can be used
Klau et al. [63]	ML	priority-Lasso, Lasso	acute myeloid leukemia from Gene Expression Omnibus repository	✓	×	×	×	×	AUC, *p*-value, sensitivity, specificity	Priority-lasso is a flexible model compared to lasso model	–

Table 3 (continued)

Author [Ref.]	ML/DL	Method	Dataset used	Type of analysis					Performance parameters	Evaluation/ contribution	Future directions
				A	B	C	D	E			
Rappoport and Shamir [64]	ML	NEMO (Neighbor-hood based multi omics clustering)	TCGA dataset with 3168 cancer patients	✓	×	×	×	×	logrank *p*-value	For multi-omics clustering, NEMO is proposed which outperforms existing methods	In sparse graphs, spectral clustering methods can be used for improvement of NEMO's network
Wu et al. [65]	ML	LRAcluster (Low Rank Approximation)	11 cancer types from TCGA dataset	✓	×	×	×	×	Classification accuracy	The dimension reduction is done fast by LRA-cluster	The driving factors of cancer can be found by modelling the information of inter-omics using different pre-processing stage
Lopes et al. [66]	ML	Logistic Regression with Elastic Net Regularization, SPLS-DA,SGPLS	Triple-negative breast cancer data from TCGA dataset	✓	×	×	×	✓	MSE	The proposed ensemble method leads to significant classification of cancer patients	The proposed method can be used for other datasets and other regression models. It can help in precision medicine by improving biomarker selection
Chang et al. [67]	ML	ANN, SVM, ANFIS (Adaptive neuro-fuzzy inference system), logistic regression	31 patients of oral cancer from MOCDTBS	✓	×	×	×	×	AUC, accuracy	Accuracy of 93.81% and AUC of 0.90 is achieved using ReliefF-GA-ANFIS hybrid model	The dataset size can be improved with extra samples
Zhang et al. [2]	DL	Auto-encoder, iCluster, PCA, xgboost, naïve bayes, SVM, logistic regression	Gene expression and CNA data of neuroblastoma from TARGET project, SEQC project	✓	×	✓	×	×	Accuracy, sensitivity, c-indices, *p*-value	Autoencoder is used for integration of multi-omics data and outperforms for classification performance	Patients' prognosis prediction and personalised treatment can be done by clinicians using integrative classification
Liang et al. [68]	DL	Multi-modal Deep Belief Network, Restricted Boltzmann Machine	Breast cancer and Ovarian cancer data from The Cancer Genome Atlas dataset	✓	×	✓	×	×	Mean squared error	A multimodal deep belief network is proposed for identification of subtypes using integrative clustering	Large-scale problems can be addressed using a proposed framework
Islam et al. [69]	DL	DNN, SVM, RF	Gene expression and CNA data from Breast cancer of METABRIC dataset	✓	×	×	×	×	AUC, accuracy	DNN outperforms SVM in cancer subtypes classification	–
Exarchos et al. [70]	ML	SVM, Decision tree, Bayesian network, Random forest, ANN	Oral squamous cell carcinoma	×	✓	×	×	×	Accuracy, sensitivity, specificity	A multiparametric decision support system is proposed	The capability of proposed method can be enhanced by using richer set of patients

Table 3 (continued)

Author [Ref.]	ML/DL	Method	Dataset used	Type of analysis					Performance parameters	Evaluation/ contribution	Future directions
				A	B	C	D	E			
Park et al. [71]	ML	Naïve Bayesian, SVM, random forest, proposed semi-supervised method	Gene expression data of Breast cancer, colon cancer, colorectal cancer from GEO database	×	✓	×	×	×	Accuracy, AUC	Proposed semi-supervised method gives 24.9% more accuracy	A balanced number of samples are difficult to obtain in medical dataset. In future, problems of semi-supervised method can be solved
Kim et al. [72]	ML	Support Vector Machine	Breast Cancer and Lung Cancer data of the TCGA project	×	×	×	✓	×	Youden index, sensitivity, specificity	Proposed Meta-SVM performs better than meta logistic regression	Programming of low-level can be used to increase speed of computation For improving prediction accuracy and feature discovery, quadratic and radial basis kernels can be used In model interaction terms can be used to apply genomic features with complex association
Mallik et al. [73]	ML	KNN, Naive Bayes, SVM, Adaboost	Gene expression and methylation data of Prostate Carcinoma dataset and Leukemia dataset	×	×	×	✓	×	MCC, accuracy, precision, sensitivity, specificity	The proposed framework helps in identification of biomarkers with increased positive predictive value	The correctness and efficiency of model can be improved by incorporating statistical measures in the model
Long et al. [74]	ML	Random forest	Pancreatic cancer data from GEO (Gene Expression Omnibus)	×	×	×	✓	×	ROC, accuracy, sensitivity, specificity	Random forest performs better for diagnosis	Biomarkers can be validated using clinical and epidemiological integration of multi-omics data
Manzalawy [75]	ML	random forest, logistic regression, extreme gradient boosting, Canonical Correlation Analysis (CCA)	Kidney Renal Clear Cell Carcinoma (KIRC) data from TCGA dataset	×	×	✓	×	×	MCC, AUC, Sensitivity, Accuracy, specificity	Proposed CCA based feature selection method outperforms	CCA method having supervised variant can be incorporated To find nonlinear relationships in views, kernalized CCA method can be utilized
Ma et al. [76]	ML	SVM (with linear, rbf kernel, polynomial kernel), random forest and Bayesian logistic regression	breast cancer patients of TCGA project	×	×	✓	×	×	AUC	Proposed predictive model and identified protein data as an effective one	Ensemble methods can be used for prediction on different omics data types Data driven feature reduction methods can be employed

Table 3 (continued)

Author [Ref.]	ML/DL	Method	Dataset used	Type of analysis					Performance parameters	Evaluation/ contribution	Future directions
				A	B	C	D	E			
Chen et al. [77]	ML	ANN	gene expression and clinical data of Lung cancer from NCI caArray database	×	×	✓	×	×	Kaplan–Meier estimate, accuracy	83% accuracy is achieved for classification using gene expression data	In future, Cox model can be used for variables selection and ANN for survival modeling
Zhu et al. [78]	ML	Kernel Machine Learning	14 Cancer types from TCGA dataset	×	×	✓	×	×	C-index	Proposed multi-omics kernel learning method and mRNA expression gives highest c-index for analysis	In recently targeted therapies, studies can be done to find omics profiling prognostic value To improve precision, additional data types can be used in framework for prediction
Kim et al. [79]	ML	GESR (Grammatical evolution symbolic regression), GENN (Grammatical evolution neural network)	Gene expression, methylation, CNA, Protein expression breast cancer data from TCGA	×	×	✓	×	×	Fitness value, accuracy	The proposed model gives 73% prediction accuracy	Chemotherapy response or recurrence can be predicted using proposed framework In future, among various cancer types, interaction associated survival can be identified
Poirion et al. [80]	DL	Autoencoder, SVM	Methylation, mRNA,miRNA data of bladder cancer from TCGA portal	×	×	✓	×	×	Log-rank *p*-value	DeepProg with autoencoder is used for identification of survival subtypes	The performance of pipeline can be improved by using samples of good quality in future Analytical pipelines having diverse clinical interests can be created using proposed strategy
Dubourg-Felonneau et al. [5]	ML	SVM, Lasso regression, Random Forest, Naïve Bayes	Microarray data from TCGA, METABRIC and TCGA RNA-seq	✓	×	✓	×	×	AUC	The proposed pipeline predicted clinical outcomes using high dimensional omics data	–

with CT data. For test data, the model gives 83.9% AUROC for accurate prediction of tumour location. The random forest model is used by Bae et al. [88] and Chaddad et al. [89]. Bae et al. [88] predicted survival improvement for glioblastoma multiforme (GBM) using radiomic features. The data of 217 patients having clinical and genetic profiles are used. To predict survival, radiomic features are used for training of random survival forest model. The results evaluated using integrated area under ROC curve show that radiomic MRI phenotyping improved survival prediction. Similarly, Chaddad et al. [89] introduced multiscale radiomic features for prediction of overall survival (OS) and to predict progression free (PFS). Glioblastoma patients are used for the analysis. The classifier random forest is applied, which results in 85.54% and 85.37% AUC for prediction of OS and PFS in patients. The proposed texture features to characterize glioblastoma regions using Laplacian-of Gaussian filter provides efficient prediction. The gradient boosting algorithm is used by Kaissis et al. [90] and Sun et al. [91]. Kaissis et al. [90] predicted molecular subtypes in pancreatic ductal adenocarcinoma with supervised machine learning method. PyRadiomics is used for extraction of radiomic features. The tumor subtypes are predicted using gradient boosting tree algorithm. The fitting of overall and disease-free survival data is done with gradient boosted survival model of regression. The results ranked entropy as the most significant radiomic feature. The AUC, sensitivity, specificity of 0.93, 0.90, 0.92 respectively are achieved with machine learning algorithm. Similarly, Sun et al. [91] predicted overall survival of non-small cell lung cancer patients using radiomic features extracted from CT images. The authors used eight machine learning models and five feature selection methods for survival prediction. In results gradient boosting based on Cox partial likelihood provides best prediction compared to other models when CI feature selection method is used. The use of k nearest neighbour method for learning is done by D'Amico et al. [92] and Ferreira Junior et al. [93]. D'Amico et al. [92] differentiated malignant from benign using radiomic features of breast cancer. The study used 45 patients with extraction of 200 radiomic features. The features are selected using TWIST (Training with input selection and testing) algorithm, and the classification is done using kNN method of machine learning. The kNN classifier gives 90% accuracy with 35 selected features. Similarly, Ferreira Junior et al. [93] predicted histopathology and metastases of lung cancer using radiomics based CT features. The analysis is done by k nearest neighbor, naïve Bayes, and RBF neural network machine learning algorithms. The classification performance is predicted using sensitivity, specificity, and AUC parameters. The results give AUC of 0.92 for histopathology pattern recognition, AUC of 0.89 for nodal metastasis, and AUC of 0.97 for distant metastasis. HI-MKL, a histopathological integrating multiple kernel learning, is developed by Zhang et al. [94]. HI-MKL method uses integrated multi-omics and histopathological images data. The method is applied on glioblastoma multiforme dataset of TCGA for analysis. The results give high accuracy of HI-MKL and make it robust for prognosis prediction compared to SimpleMKL and MeanMKL method. The feature extraction using LASSO method is done by Lao et al. [95] and Fu et al. [96]. Lao et al. [95] proposed a model of deep learning in radiomics using transfer learning. LASSO model is used to construct six-deep-feature signature. The dataset used consists of patients having Glioblastoma Multiforme (GBM) from TCGA database. The proposed signature of radiomics model predicts overall survival with better performance in GBM patients. Similarly, Fu et al. [96] predicted lymphovascular invasion in rectal cancer. From single region and multi region, the radiomics features along with clinical features are collected. The feature extraction of radiomics is done using LASSO algorithm and Spearman correlation. On the extracted 21 features a ridge classification model is applied. The results concluded that multi regional model outperforms with accuracy 79% compared to single regional model with accuracy 67%. ANN model is applied by Chufal et al. [97] on lung cancer data for prognostic modeling. The study used 422 patients of non-small cell lung carcinoma having radiomics and clinical data. Both direct and hybrid modelling is applied for outcome prediction. In direct modeling, Multilayer perceptron gives better accuracy of 79.2% compared to radial basis function with accuracy 61.4%. In hybrid modeling, after selecting features using clustering MLP gives 80% accuracy. BiCNN method is developed by Weil et al. [98] using deep CNN for classification of images of breast cancer. The authors used dataset BreakHis having histopathological images of tumor patients. The proposed model gives an accuracy of 97% using classification for diagnosis of breast cancer. The comparative analysis of the existing work done in radiomics data analysis using machine learning is given below in Table 4.

Effective analysis of omics data demands metaheuristic techniques for developing predictive models, as discussed in Sect. 4.4.

4.4 Metaheuristic Techniques for Omics Data Analysis

A tremendous amount of data is available due to rapid advances in omics technologies such as genomics, metabolomics, proteomics. Use of conventional intelligence techniques makes the interpretation and analysis of data a challenging task for bioinformaticians. This leads to development of hybrid intelligent approaches in which multiple standard intelligent methods are integrated. In hybrid methods, there is extensive use of metaheuristic algorithms such as genetic algorithm, ant colony optimization, and

Table 4 Comparison of work done in radiomics data analysis using machine learning

Author [Ref.]	Method	Dataset used	Contribution	Limitation	Future directions
Incoronato et al. [81]	–	Cancer dataset	Integrated approaches are reviewed	Standardization is required for post-processing, image acquisition It is a challenge to get signatures from radiogenomics	For precision medicine integrated approach using genetics and phenotype is encouraged
Acharya et al. [82]	–	–	Discussed workflow of radiomics	In image acquisition, there is no standard procedure In clinical research, sharing of radiological images is always a problem	Features from radiological images can be correlated with omics data to develop an effective model Deep learning can be incorporated in radiomics For precision medicine, development in radio-genomics is needed
Li et al. [83]	–	Breast cancer data of TCGA dataset	Predicted breast cancer subtypes using image phenotypes	The sample of patients is small due to unavailability of MR images The current MRI technology is not used as images collected ten years ago	Protocols of image acquisition with high standards can be developed Genomic data merged with image phenotypes can improve predictions
Zhou et al. [84]	SVM	Lung cancer	Predicted distant metastasis with CT radiomic features selected using FSV	The results are biased as patients are taken from Asia The CT images are nonenhanced and are non-comparable as attained at various time points	Performance can be improved by an extra study of phenotypic features
Takahashi et al. [85]	Logistic regression, SVM	Glioma grade	Predicted glioma grading using radiomic analysis with diffusion kurtosis imaging	In institution, a small number of cases are available for study Very important MRI sequence GdT1 is excluded in study	Conventional and advanced MRI sequences can be combined to create an effective machine learning model
Li et al. [86]	SVM	Pancreatic ductal adenocarcinoma	Predicted survival and classification of gene expression status using CT radiomic features	External validation lacks as large cohort of patients is required for model verification Overall characteristics of tumor are not possible as region used is of maximum tumor area	–
Clifton et al. [87]	Decision tree	Lung cancer	Segmented lung tumour location with radiomic features	–	PET scans needed to get additional functional data to verify independent tumour site

Table 4 (continued)

Author [Ref.]	Method	Dataset used	Contribution	Limitation	Future directions
Bae et al. [88]	Random survival forest	glioblastoma multiforme	Predicted survival using radiomic features	It is time consuming to outline the tumors semiautomatically There is lack of external validation	–
Chaddad et al. [89]	Random Forest	40 glioblastoma patients from TCGA dataset, MR images from TCIA dataset	Proposed multiscale texture features for prediction of survival	The images FLAIR MRI and T1-WI are only used for analysis	The increase in image modalities can help in improvement of developed model performance The correlation between survival and texture can be used to extend the model
Kaissis et al. [90]	Gradient boosting tree	pancreatic ductal adenocarcinoma	Predicted molecular subtypes using supervised learning on radiomic features extracted using PyRadiomics	A small cohort size is used in study External testing lacks along with investigation used is of single center nature	–
Sun et al. [91]	GB-Cox, CoxBoost, Cox, GB-Cindex, RFS, BST, SVCR, SR	Lung cancer patients from 'NSCLC-Radiomics' collection	Investigated machine learning models using radiomic features for survival prediction	–	The prognostic models to be used can be referenced with machine learning models compared in this study
D'Amico et al. [92]	kNN	Breast cancer dataset	Classified breast cancer using features selected by TWIST algorithm	The study used small sample size	To improve the robustness and performance of method clinical data can be added in dataset
Ferreira Junior et al. [93]	KNN, RBF, NB	Local image dataset of Lung cancer	Recognized pattern of lung cancer using radiomics based features	The size of dataset is a limitation which produces overfitted results	Different approaches can be used to extract features, and prediction can be improved by using clinical outcomes In future large datasets can be used to train, test, and validate
Zhang et al. [94]	Multiple kernel learning	Glioblastoma multiforme dataset from TCGA	MKL method is used for survival and prognosis prediction using radiomic features and multi-omics data	The SimpleMKL method gives no improved results by using histopathological features as they are heterogeneous and less in number	–
Lao et al. [95]	LASSO model	GBM patients from TCGA	Predicted survival accurately using radiomic signature with deep learning	The sample size is small for retrospective study The association within features of deep learning and characteristics of genetic was not studied	The correlation of radiomics and genomics should be explored in dataset of a patient The feature extraction model can be trained with fine tuning The generalization of model can be measured using large scale multicentre data

Table 4 (continued)

Author [Ref.]	Method	Dataset used	Contribution	Limitation	Future directions
Fu et al. [96]	LASSO method	Rectal cancer dataset	Predicted rectal cancer using LASSO for radiomic feature selection	There is a lack of external testing validation, which leads to overfitting The same MR scanner is used to perform MRI, which decreases the model robustness	Data from different MRI scanners can be used for robustness improvement External validation with large sample size can be studied
Chufal et al. [97]	ANN	Non-small cell lung carcinoma	Applied ANN model on lung cancer data for prognostic modeling	–	The algorithm to be externally validated using an institutional or public dataset The variation in tumor segmentation can be studied
Weil et al. [98]	CNN	Break his breast cancer dataset	Classified breast cancer using developed BiCNN model	–	–

particle swarm optimization because of their robustness and efficiency.

Genetic algorithm is widely used optimization by many researchers for feature selection. Lu et al. [99] proposed GAOGB model for survival prediction of breast cancer. In proposed model, genetic algorithm for optimal feature selection is used along with gradient boosting model. Also, adaptive linear regressor is selected as the base learner to enhance the adaptiveness. The authors evaluated GAOGB model using SEER BC dataset, WBCD dataset, Ionosphere dataset, and Spambase dataset. GAOGB model is compared with OLR, OSELM, OGB, OAB learning models using accuracy, sensitivity, specificity, AUC, variation, retraining time parameters. From the results, GAOGB outperforms compared to existing learning models and base learners. Yang et al. [100] proposed fused KPLS (fKPLS) using multilevel omics data for disease classification and prediction. In proposed model, genetic algorithm(GA) is used for optimization of kernel parameters and kernel weights. From the results, it is shown that GA-fKPLS model improved the prediction of subtype (triple-negative and non-triple negative) in breast cancer. Li et al. [101] developed an approach for gene selection using a hybrid of support vector machine and genetic algorithm. For analysis, the dataset of diffuse large B cell lymphoma having microarray data is selected. The prediction accuracy of 99% is achieved using microarray data with resultant classifier. Also, for complex biological phenotype key feature genes are identified. Ram and Kuila [102] used genetic algorithm with SVM classifier to develop model for selection of minimal feature set. The microarray dataset of Prostate, DLBCL cancer is used for simulation to get accuracy, sensitivity, and specificity. The results show an increase in accuracy using proposed GA algorithm compared to Differential Evolution. Fortino et al. [103] developed GARBO algorithm for biomarker discovery. GARBO is a developed genetic algorithm that uses random forest classifier and fuzzy logic. For prediction of drug sensitivity and cancer patient stratification, TCGA, GDSC, and CCLE omics datasets are used. GARBO selects minimum feature set with Pareto-based selection compared to weighted sum method. The results show that biomarker model selection using proposed algorithm gives better classification accuracy compared to existing GA methods. Kečo et al. [104] proposed two-step prediction framework with machine learning and genetic algorithm implemented on the cloud platform. Hadoop MapReduce is used by extended parallel algorithm to reduce computation time. The authors used SVM and ANN on GEMS datasets having tumor, and results achieved effective accuracy. Yu et al. [105] proposed ACOSampling method for classification of imbalanced dataset of DNA microarray. The evaluation of method is done using SVM on dataset of lung cancer, colon, glioma, and central neural system. The authors considered accuracy, F-measure, AUC,

G-mean metrics for performance evaluation. From results, ACOSampling is detected as an effective technique to get informative samples from imbalanced dataset compared to other sampling algorithms. Xu et al. [106] developed HI-DFNForest framework for classification of cancer subtype by integrating multi-omics data. In HI-DFNForest model, particle swarm optimization is used to optimize the parameters for base classifier. The data representations are learned using stacked autoencoder, and then deep flexible neural network is used for cancer subtype classification. The authors used gene expression, DNA methylation, miRNA data from TCGA for BRCA, OV, GBM datasets. From the results, it is concluded that multi-omics data integration increases the classification accuracy compared to gene expression data. The comparative analysis of existing work done in omics data analysis using metaheuristic techniques is given below in Table 5.

To perform analysis on large omics datasets, there are open-source tools available for researchers as discussed in Sect. 5.

5 Tools for Omics Data Analysis

It is a challenge for researchers without bioinformatics skills to integrate and analyse high volume of omics data. So, several works developed tools for integration and analysis of complex data created from omics technologies. Sangaralingam et al. [107] provided O-miner, an efficient web tool to automatically combine and analyse data generated by omics technologies. The tool helps for identification of significant pathways and for prioritizing biomarkers from datasets having transcriptome, genome, and methylation data along with clinical/biological information. The pipelines developed for tool uses statistical methods of Bioconductor packages and run in the R environment. Colaprico et al. [108] developed TCGAbiolinks, a package used for analysing TCGA data. The tool also provides functions to download, query, and prepare TCGA data for analysis. It is the first tool that helps in integration of gene expression data with copy number or DNA methylation, providing an opportunity for users to perform integrative analysis. Yu et al. [109] developed first cloud-based Omics AnalySIs System for PRecision Oncology (OASISPRO) for visualizing and analysing omics information from TCGA dataset having transcriptome, proteome, DNA methylation, microRNA and clinical data. The proposed system is web based that provides predictions by visualizing clinical profiles of patients and executing machine learning method of choice. The tool also helps in prediction of survival outcome of patient and identification of genes linked strongly with cancer. The tool will contribute towards precision medicine by finding an association between clinic phenotypes and omics. The authors proposed the system to bridge the gap of developing data mining tools that need no training of programming and bioinformatics for physicians, scientists, and medical researchers. Martinez-Mira et al. [110] developed MOSim to simulate multi-omics datasets having various omics data, including gene expression data. MOSim is a tool built in R and is used for benchmarking the analysis pipeline and for testing the performance of integrative methods. The functions offered by MOSim package also include creating synthetic multi-omics dataset, modification of omics data, and recovery of simulation results. Cumbo et al. [111] proposed TCGA2BED for searching and retrieving TCGA data. The tool supports integration and conversion of TCGA data in BED format and also in other standard formats like JSON, CSV, XML, and GTF. The tool allows extension of TCGA data with other genomic databases. Ulfenborg [112] developed miodin package to do integration and analysis of multi-omics data. The package works for both horizontal and vertical integration of omics data. In the package, there are workflows to import and process low-level data without use of technical experts and for transparent data analysis. Deng et al. [113] developed Web-TCGA for integrative analysis of molecular datasets of TCGA. The tool is used to create global molecular profiles, and it provides interactive tables and views for detailed analysis. Compared to other tools, there is no need of installation and configuration for analysis of datasets in Web-TCGA. Hernandez-Ferrer et al. [114] proposed MultiDataSet, to manage multiple datasets using simple methods of sample selection and feature subsetting. MultiDataset is a R class developed under Bioconductor standards. In proposed class, integrative analysis using packages of third party can be done. For integration, new methods and functions can be created. Also, data from biological experiment in unimplemented form can be encapsulated. Singh et al. [115] developed a method DIABLO, for identification of biomarkers relevant to disease by integrating multiple omics datasets. The method uses latent component conversion of omic dataset for analysis. This mixOmics framework for multi-group classification is capable and versatile for high dimensional data of both real-world and simulated environment. Wang et al. [116] used cloud-based technology and developed WebMeV tool to visualize and analyse Cancer Genome data. This web-based tool is used to upload own datasets and to access large datasets which are publicly available. To help users deploy their hardware architecture, WebMeV docker container image is provided in this tool. Zhu et al. [117] developed a software package, TCGA Assembler that retrieves, process, and assembles publicly available TCGA data automatically. The package contains two modules, one to streamline downloading of data and another for analysis

Table 5 Comparison of work done in omics data analysis using metaheuristic techniques

Author [Ref.]	Algorithm used	Optimization technique	Dataset used	Contribution	Future directions
Lu et al. [99]	Gradient boosting machine	Genetic algorithm	SEER BC dataset, WBCD dataset, Ionosphere dataset, and Spambase dataset	Proposed GAOGB model for survival prediction of breast cancer using optimized features	–
Yang et al. [100]	Fused KPLS (fKPLS)	Genetic algorithm	Breast cancer data from TCGA dataset	Proposed GA-fKPLS for prediction of disease with kernel weights and parameters optimized with genetic algorithm	To improve computational efficiency, different kernel methods of approximation can be explored
Li et al. [101]	SVM	Genetic algorithm	large B cell lymphoma	Developed approach for gene selection using a hybrid of SVM and genetic algorithm	In various biological investigations having microarray data, the proposed method can be considered
Ram and Kuila [102]	SVM	Genetic algorithm	Prostate, DLBCL cancer dataset	Used genetic algorithm with SVM classifier to develop model for selection of minimal feature set	–
Fortino et al. [103]	random forest, fuzzy logic	Genetic algorithm	TCGA, GDSC, and CCLE dataset	Developed GARBO algorithm for biomarker discovery	The proposed method can be applied to new biomarkers generated from sources of omics data
Kečo et al. [104]	ANN and SVM	Genetic algorithm	11 cancer datasets from GEMS database	Two-step cancer classification is done using framework having GA merged with SVM and ANN	–
Yu et al. [105]	SVM	ACO	Lung cancer, colon, glioma, and central neural system	Proposed ACOSampling method for classification of imbalanced dataset	The efficiency of ACOSampling can be improved with formation rule modification by considering its computation and storage cost In class imbalance problems of real-world ACOSampling can be useful In multiclass tasks, ACOSampling can be applied
Xu et al. [106]	Stacked autoencoder, deep flexible neural network	Particle swarm optimization	BRCA, OV, GBM datasets of TCGA	Developed HI-DFNForest framework for classification of cancer subtype	–

by processing the data. Wei et al. [118] developed TCGA-assembler 2 package, as TCGA-Assembler was not able to download data as data was shifted from TCGA portal to GDC. Also, some data of proteomics provided by CPTAC program was not supported to download by TCGA Assembler. The proposed version of TCGA assembler is used to download data from both CPTAC and GDC portal in an integrated form. In TA2, the functions give more reliability and performance as they are optimized and modified. Xie et al. [119] developed a repository, MOBCdb, for integration of clinical, genomic, epigenomic, transcriptomic data. The database is developed for user to extract gene expression, SNV, DNA methylation, microRNA data of breast cancer patients. In MOBCdb to simultaneously visualize multi-omics data of various samples, an interface is available. Also, the survival analysis of this data is done using survival module of MOBCdb. As a complete web interface MOBCdb helps in precision medicine by identifying novel biomarkers of various subtypes of breast cancer. Chen et al. [120] developed OmicsARules, an R package, to integrate datasets of multi-omics. OmicsARules is an open-source and free package used to find intensive changes with association rule mining. The authors used Lamda3 in OmicsARules for pattern evaluation and gene prioritization. The datasets of breast cancer and lung cancer having DNA methylation and RNA seq data are used for analysis. The evaluation shows better results using Lamda3 compared to other rule ranking methods. Koh et al. [121] developed iOmicsPASS, a supervised analysis technique for omics data to discover predictive subnetworks. In omics data, to find dense subnetworks connection, the tool firstly calculates a score for interaction in molecules and then used nearest shrunken centroid algorithm and predicted accuracy of sample group. For TCGA/CPTAC dataset of breast cancer, iOmicsPASS helps in finding its subtype as positive marker using network regulation, which was not possible if omics data analysed individually. Fisch et al. [122] developed Omics Pipe, a framework for the analysis of both available and newly produced data. Omics Pipe is used in the cloud to automate analysis of multi-omics data pipeline. For testing, Omics Pipe is evaluated using datasets of TCGA breast patients. For each sample, automatic download and processing on compute cluster of high performance are done by Omics Pipe, and results are generated. The better analysis shows the use of efficient methods in the proposed framework. Jang et al. [123] developed MONGKIE, a platform to integrate omics data analysis tools with network visualization. Network clustering algorithms are used for network analysis. MONGKIE is used to describe complex biomolecular reactions and to display multi-omics data using options in visualization unit. The glioblastoma dataset from TCGA is used for analysis with MONGKIE. Polpitiya et al. [124] developed DAnTE, a tool for complete analysis of omics data. The methods like imputation of missing value, data normalization, and investigative plots are available in DAnTE for data analysis. DAnTE includes ANOVA method for hypothesis testing and Partial least square method as analytical algorithm. The tool is applicable for analysis of proteomics as well as microarray data. Eren et al. [125] developed Anvi'o, a platform for visualization, and advanced analysis of omics data. The tool offers a single display of omics data by linking multiple sources. The publicly available datasets are analysed using the tool to identify genomic changes. It is an open-source tool and even used for analysis of large omics datasets by researchers without bioinformatics knowledge. Guhlin et al. [126] developed ODG, a tool to generate, analyse, and query multi-omics data. ODG is Omics Database Generator which uses available genomic data to generate customized database. A multi-dimensional database is created by ODG using experimental and omics data. ODG is preferred to conduct multi-omics queries for understudied species. Surujon and van Opijnen [127] developed a web-based application, ShinyOmics, for management, exploration, and online sharing of omics data. Omics data of two human pathogens having proteomic, microarray, RNA-seq, Tn-Seq is loaded in ShinyOmics. The proposed application is easy to customize and set up. Blatti et al. [128] proposed a cloud platform KnowEnG (Knowledge Engine for Genomics) for analysis of genomic data. The platform contains data mining and machine learning algorithms for analysis. KnowEnG is used for analysis of gene set, prioritization ofgene, analysis of expression signature, and for sample clustering. Using KnowEnG, the execution of software packages is done on user computer with installation. Zhao et al. [129] proposed a cloud-based tool, Rainbow, for automatic analysis of whole genome sequencing data. The tool is built using Crossbow, and it uses Amazon web services to process genomic data. The main advantage of Rainbow is that for load balancing, large scale files can be splitted. Also, it can handle both FASTQ and BAM input files. It is an open-source, cost-effective, and scalable tool. Chiesa et al. [130] developed GARS (Genetic Algorithm for identification of Robust Subset), a tool based on genetic algorithm to identify features subset accurately. In high dimensional data, it is an appropriate tool for feature selection. GARS show high classification accuracy compared to wrapper, filter based, embedded methods. Also, analysis of high dimensional data without any step of pre-filtering can be done only by GARS. To solve complex problems in real-world applications, GARS is feasible to be applied. Mohammed et al. [131] developed CancerDiscover for prediction of cancer class and identification of cancer biomarkers. The tool helps in normalization and provides multiple feature selection methods to select the features of

best performance. Using CancerDiscover, high throughput raw dataset can be processed automatically and efficiently. In proposed integrative pipeline, various models can be developed for identification of cancer types and subtypes. CancerDiscover is a tool available free for use and is open-source. The comparative analysis of existing tools for omics data analysis is given below in Table 6.

6 Discussion

In this review article, the techniques and tools used for analysis of omics data are presented. In this survey, latest and relevant research publications are reviewed. The objective of this article is to mainly focus on machine learning methods, tools, and metaheuristic techniques applied for analysis of omics data as it is expected to be the area of focus because of its need for precision medicine in future.

In response to first research question, the omics types genome, proteome, transcriptome, and metabolome are discussed. Extensive research is done by taking either one omics data type or multiple data types for analysis.

In response to second research question, integrative analysis is explained by differentiating it from single level analysis. Also, the categories of integration: concatenation based, transformation based, model based are explained. In literature, for integration various techniques are developed as a single data type is not sufficient to understand the biological system.

In response to third research question, the detailed information of techniques developed for analysis of omics data, multi-omics data, and radiomic data are presented. Currently, various machine learning, deep learning, statistical methods are proposed for effective analysis of omics, multi-omics, radiomic data for learning to predict the disease, discover the biomarkers, predict the recurrence of disease, and for survival analysis.

In response to fourth research question, metaheuristic techniques used for omics data analysis are discussed. In literature, various optimization techniques are used for the improvement of results in analysis. A genetic algorithm is a popular technique used in the existing studies for omics data analysis.

In response to fifth research question, various existing freely available tools along with limitations are presented. Mostly tools are open source and are available for users to perform the required tasks. Some tools are developed on the cloud platform, and the packages are available on user demand.

In the end, sixth research question is answered by discussing various challenges for omics data analysis. Future research directions are given based on the complete study of related articles for omics data analysis in the literature.

6.1 Challenges for Omics Data Analysis

Based on the existing studies, some challenges are there while performing omics data analysis, as shown in Fig. 10 and discussed below.

Heterogenous datasets: In healthcare, different datasets contain heterogeneous data in terms of variables count, scaling, distribution, modality. This heterogeneity in data makes the analysis purpose difficult in healthcare. To solve the problem of heterogeneity, there are various machine learning approaches that can be used like clustering, graphs and networks, kernel learning, and deep learning.

Complex datasets: In healthcare, compared to other domains the datasets are complicated. There is difficulty to know about the growth and cause of the disease as it is heterogeneous. In healthcare dataset, there is a limited number of patients, which makes the analysis to be performed difficult.

Temporal data: With time, the disease in healthcare progress and change in a non-deterministic manner. The machine learning algorithms face difficulty to tackle the factor of time as they assume inputs to be static.

Data missingness: Missingness in data means some observations are missing because of different reasons like protein having less sensitivity, less handle of next-generation sequence data. The missingness creates a problem when most of the samples in dataset are having missing data. Listwise deletion is done as a solution to this problem in which whole sample of missing variables is deleted from dataset. But this approach can make loss of information if the percentage of missing data is more. Imputation technique is also used in which some random values, mostly mean or median is assigned to the variable with missing data. Primarily, k-nearest neighbors technique is used to impute missing values.

Dimensionality curse: In healthcare, mostly the datasets contain fewer samples with a high number of features or variables. This increase in the number of variables, develop a problem of overfitting for various machine learning algorithms. In healthcare, datasets contain different types like imaging, clinical, and omics data. There are multiple techniques of data reduction to be applied to make the high dimensional datasets capable of learning. The techniques are generally of feature extraction and feature selection. In feature extraction, methods like Principal Component Analysis (PCA), Non-negative matrix factorization (NMF) are used. Feature extraction is primarily used for unsupervised learning where labels of response are not known. In multi-omics data, machine learning using feature extraction is used mostly to discover subgroups of a disease. In feature selection, a subset of features is selected. Filter, Wrapper, and Embedded methods are types of feature selection. In Filter method, features subset is selected without the use of any model. Various filter methods used are such as Pearson

Table 6 Comparison of existing tools for omics data analysis

Author [Ref]	Tool	Year	Technology/platform	Availability link	Limitations
Sangaralingam et al. [107]	O-miner	2019	R statistical environment	http://www.o-miner.org	Needs pre-processing of Next Generation Sequencing inputs For sequencing data, sequence alignment and QC needs to be conducted by user The integration of various analytical layers results is not possible
Colaprico et al. [108]	TCGAbiolinks	2015	R/Bioconductor	http://bioconductor.org/packages/TCGAbiolinks/	–
Yu et al. [109]	OASISPRO	2017	cloud	http://tinyurl.com/oasispro	–
Martinez-Mira et al. [110]	MOSim	2018	R package	https://bitbucket.org/ConesaLab/mosim/	–
Cumbo et al. [111]	TCGA2BED	2017	Java programming language	http://bioinf.iasi.cnr.it/tcga2bed/	–
Ulfenborg [112]	miodin	2018	R programming language	https://gitlab.com/algoromics/miodin	–
Deng et al. [113]	Web-TCGA	2016	R language	https://github.com/mariodeng/web-TCGA	–
Hernandez-Ferrer et al. [114]	MultiDataSet	2017	R programming language	https://bioconductor.org/packages/release/bioc/html/MultiDataSet.html	One of the limitations of MultiDataSet is that memory management is not optimized as it uses straightforward extensions of Bioconductor's classes MultiDataSet can only be used with three integration pipelines: by mcia (omicade4) and iClusterPlus through a wrapper and as input object for MEAL package
Singh et al. [115]	DIABLO	2016	mixOmics R package	http://mixomics.org/mixdiablo/	–
Wang et al. [116]	WebMeV	2017	R version repository	http://mev.tm4.org	The performance can be affected by limited number of nodes assigned to application
Zhu et al. [117]	TCGA-Assembler	2014	R programming language	https://bio.tools/TCGA-Assembler	–
Wei et al. [118]	TCGA-assembler 2	2018	R programming language	http://www.compgenome.org/TCGA-Assembler	–
Xie et al. [119]	MOBCdb	2018	Perl, R, MySQL database	http://bigd.big.ac.cn/MOBCd b/	Exome sequencing is used to obtain SNV data which makes only 1.5% part of human genome Normal tissue samples are having no sample of healthy human, and it is 10% of sample
Chen et al. [120]	OmicsARules	2019	R programming using ARM framework	https://github.com/Bioinformatics-STU/OmicsARules	–

Table 6 (continued)

Author [Ref]	Tool	Year	Technology/platform	Availability link	Limitations
Koh et al. [121]	iOmicsPASS	2019	R programming	https://github.com/cssblab/iOmicsPASS	iOmicsPASS has no functionality to predict phenotype group Poorly represented markers in network data are discarded in the proposed tool
Fisch et al. [122]	Omics Pipe	2015	Python, Cloud	http://sulab.scripps.edu/omicspipe	–
Jang et al. [123]	MONGKIE	2016	Java (NetBeans architecture)	http://yjjang.github.io/mongkie	–
Polpitiya et al. [124]	DAnTE	2008	C#, R language	http://omics.pnl.gov/software/	–
Eren et al. [125]	Anvi'o	2015	R programming	http://merenlab.org/projects/anvio	In anvi'o the clustering for human-guided binning cannot be done for splits count more than 25,000. It is because, for time complexity O(n^2) or high, the hierarchical algorithms can not be used
Guhlin et al. [126]	Omics Database Generator	2017	Java programming language	https://github.com/jguhlin/odg	In ODG, the data is available only in standard file formats such as TSV, GFF3, OBO, FASTA, etc
Surujon and van Opijnen [127]	ShinyOmics	2020	R programming language	https://github.com/dsurujon/ShinyOmics	–
Blatti et al. [128]	KnowEnG	2020	Cloud platform	https://knoweng.org/analyze/	For bioinformatics analysis, the platform is not one size that can fit all solution
Zhao et al. [129]	Rainbow	2013	Cloud computing,	http://s3.amazonaws.com/jnj_rainbow/index.html	In cloud there is network congestion or failure as to move large datasets is not trivial In cloud to debug workflows, it is a difficult task
Chiesa et al. [130]	GARS	2020	R programming language	https://www.bioconductor.org/packages/GARS/	–
Mohammed et al. [131]	CancerDiscover	2017	WEKA, Affy R package	https://github.com/HelikarLab/CancerDiscover	–

correlation, ANOVA, Information Gain. In wrapper method, prediction model is used repetitively to get best subset of features by keeping the worst subsets on side. Various wrapper methods used are such as Boruta, Jackstraw, and Recursive Feature Elimination (RFE). On large datasets, wrapper methods do computation expensively. In terms of complexity of computation, embedded method lies between wrapper and filter method. LASSO is an embedded method used. In classification and regression, feature selection is used mainly for supervised learning. Before integration of datasets, feature selection method is used as all features are not informative in high dimensional datasets.

Imbalance data: In healthcare, an imbalance of data occurs when target class in dataset has positive values very less compared to samples having negative values. In machine learning, there are various ways to handle the imbalance data like sampling of data, modification of algorithms, and ensemble learning. In sampling, the dataset is balanced before performing classification using machine learning. Mostly, over-sampling and under-sampling are done in combination to solve the issue of an imbalanced dataset. In modification of algorithms, the algorithms are modified by using the same dataset with imbalance. The algorithm is modified by giving high weights to samples

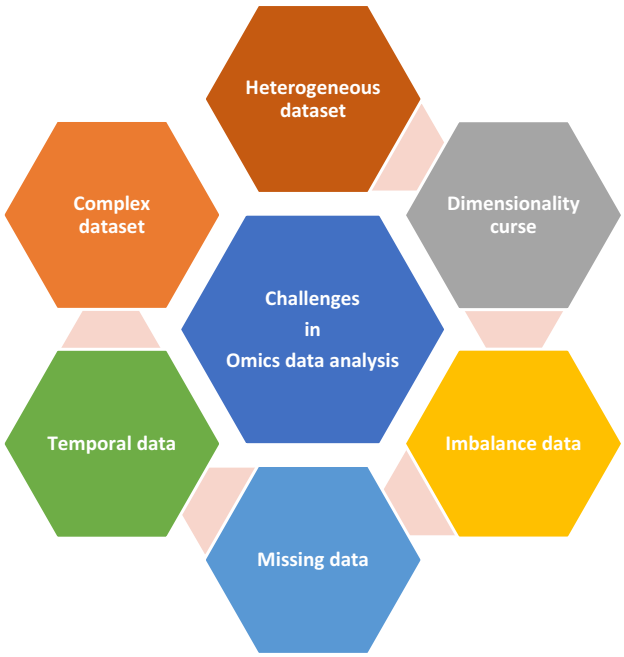

Fig. 10 Challenges for omics data analysis

with minority types. In ensemble learning, aggregation of decisions of all classifiers is calculated by applying every single model subset of majority type samples and including all of minority samples.

6.2 Future Research Directions

In this area of research, based on the existing literature following are the possible future directions.

- For effective analysis, more genomic data needed to be collected, and new machine learning, deep learning methods can be developed for feature selection and disease prediction [36, 37, 39, 43, 48].
- The methods developed for omics data analysis can be extended for analysis of multi-omics data [42, 45].
- For multi-omics data analysis, additional complex data types at different levels can be integrated [54, 56, 60, 78].
- Machine learning, deep learning and ensemble models can be improved for analysis [3, 55, 64, 72, 73, 75, 76].
- The dataset having more samples to be used for enhancement of model capability [58, 67, 70, 71, 80].
- Effective learning models need to be developed using genomic data integrated with radiomic images features extracted using advanced MRI scanners for precision medicine [80–83, 85].
- The validation of developed algorithms for radiomic data analysis to be done using large data samples taken from an institute or a public dataset [96, 97].
- In radiomics, deep learning methods can be incorporated [82].

7 Conclusions

Computationally intelligent techniques aids to gain insights into Omics data analysis for disease prediction, biomarker discovery, recurrence prediction, and survival analysis. Supervised learning has surfaced as the most significant application to date in the field of biomedical analysis. There is no universal model or method that is applicable in all scenarios. The model varies problem to problem. The blending of omics technology with artificial intelligent technologies has built fertile grounds for efficient data analysis along with novel challenges for bioinformatics analysts. Deep learning will remain the fuelling technology in the present era of increasing volumes of medical, omics, clinical health records, medical images, and publically available data.

In this paper, various approaches, existing techniques, and possible caveats in the integration of omics data, omics data analysis, multi-omics data analysis, radiomic data analysis, and various metaheuristic techniques have been investigated. A comparative study has been carried out by analysing tools and techniques developed by various researchers. It is concluded that single omics data is not sufficient for analysis purposes. The integrative methods helps in the growth of multi-omics data, which leads to more effective analysis. Also, the use of imaging features from radiomics data integrated with genomic data makes the data analysis research more informative and promising. With the increase in different types of omics and biomedical data, futuristic approaches should be targeted towards the development of standardised predictive analytic pipelines of multi-omics data analysis.

Funding The authors have no funding to report.

Compliance with Ethical Standards

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

Ethical Standards The author declares that this article complies the ethical standard.

References

1. Miotto R, Wang F, Wang S et al (2017) Deep learning for healthcare: review, opportunities and challenges. Brief Bioinform 19(6):1236–1246. https://doi.org/10.1093/bib/bbx044

2. Zhang L, Lv C, Jin Y et al (2018) Deep learning-based multi-omics data integration reveals two prognostic subtypes in high-risk neuroblastoma. Front Genet 9:477. https://doi.org/10.3389/fgene.2018.00477
3. Yan KK, Zhao H, Pang H (2017) A comparison of graph- and kernel-based -omics data integration algorithms for classifying complex traits. BMC Bioinform 18(1):539. https://doi.org/10.1186/s12859-017-1982-4
4. Manogaran G, Vijayakumar V, Varatharajan R et al (2018) Machine learning based big data processing framework for cancer diagnosis using hidden Markov model and GM clustering. Wirel Pers Commun 102(3):2099–2116. https://doi.org/10.1007/s11277-017-5044-z
5. Dubourg-Felonneau G, Cannings T, Cotter F et al (2018) A framework for implementing machine learning on omics data. arXiv preprint arXiv:1811.10455
6. Chaudhary K, Poirion OB, Lu L, Garmire LX (2018) Deep learning–based multi-omics integration robustly predicts survival in liver cancer. Clin Cancer Res 24(6):1248–1259. https://doi.org/10.1158/1078-0432.CCR-17-0853
7. Kim M, Tagkopoulos I (2018) Data integration and predictive modeling methods for multi-omics datasets. Mol Omics 14(1):8–25
8. Antonelli L, Guarracino MR, Maddalena L, Sangiovanni M (2019) Integrating imaging and omics data: a review. Biomed Signal Process Control 52:264–280
9. Zhang Z, Zhao Y, Liao X et al (2019) Deep learning in omics: a survey and guideline. Brief Funct Genom 18(1):41–57. https://doi.org/10.1093/bfgp/ely030
10. Li Y, Wu FX, Ngom A (2018) A review on machine learning principles for multi-view biological data integration. Brief Bioinform 19(2):325–340. https://doi.org/10.1093/bib/bbw113
11. Rappoport N, Shamir R (2018) Multi-omic and multi-view clustering algorithms: review and cancer benchmark. Nucleic Acids Res 46(20):10546–10562. https://doi.org/10.1093/nar/gky889
12. Wei Y (2015) Integrative analyses of cancer data: a review from a statistical perspective. Cancer Inform 14:173–181. https://doi.org/10.4137/CIN.S17303
13. Wu C, Zhou F, Ren J et al (2019) A selective review of multi-level omics data integration using variable selection. High Throughput 8(1):4. https://doi.org/10.3390/ht8010004
14. Hasin Y, Seldin M, Lusis A (2017) Multi-omics approaches to disease. Genom Biol 18(1):83
15. Kodama Y, Shumway M, Leinonen R (2012) The sequence read archive: explosive growth of sequencing data. Nucl Acids Res 40(D1):D54–D56. https://doi.org/10.1093/nar/gkr854
16. Clough E, Barrett T (2016) The gene expression omnibus database. In: Mathé E, Davis S (eds) Statistical genomics. Methods in molecular biology, vol 1418. Humana Press, New York, pp 93–110
17. Vizcaino JA, Deutsch EW, Wang R, Csordas A, Reisinger F, Rios D, Dianes JA, Sun Z, Farrah T, Bandeira N, Binz PA (2014) ProteomeXchange provides globally coordinated proteomics data submission and dissemination. Nat Biotechnol 32(3):223–226. https://doi.org/10.1038/nbt.2839
18. Jones P, Côté RG, Martens L, Quinn AF, Taylor CF, Derache W, Hermjakob H, Apweiler R (2006) PRIDE: a public repository of protein and peptide identifications for the proteomics community. Nucl Acids Res 34(1):D659–D663. https://doi.org/10.1093/nar/gkj138
19. Wilhelm M, Schlegl J, Hahne H et al (2014) Mass-spectrometry-based draft of the human proteome. Nature 509(7502):582–587. https://doi.org/10.1038/nature13319
20. Kale NS, Haug K, Conesa P et al (2016) MetaboLights: an open-access database repository for metabolomics data. Curr Protoc Bioinform 53(1):14.13.1-14.13.18. https://doi.org/10.1002/0471250953.bi1413s53
21. Conesa A, Mortazavi A (2014) The common ground of genomics and systems biology. BMC Syst Biol 8(S2):S1
22. Evangelou E, Ioannidis JPA (2013) Meta-analysis methods for genome-wide association studies and beyond. Nat Rev Genet 14(6):379–389. https://doi.org/10.1038/nrg3472
23. Kristensen VN, Lingjærde OC, Russnes HG et al (2014) Principles and methods of integrative genomic analyses in cancer. Nat Rev Cancer 14(5):299–313. https://doi.org/10.1038/nrc3721
24. Dhillon A, Singh A (2020) eBreCaP: extreme learning-based model for breast cancer survival prediction. IET Syst Biol 14(3):160–169
25. Ding H (2016) Visualization and integrative analysis of cancer multi-omics data. Dissertation, The Ohio State University
26. Ritchie MD, Holzinger ER, Li R et al (2015) Methods of integrating data to uncover genotype-phenotype interactions. Nat Rev Genet 16(2):85–97. https://doi.org/10.1038/nrg3868
27. Dhillon A, Singh A, Vohra H, Ellis C, Varghese B, Gill SS (2020) IoTPulse: machine learning-based enterprise health information system to predict alcohol addiction in Punjab (India) using IoT and fog computing. Enterp Inf Syst 1–33. https://doi.org/10.1080/17517575.2020.1820583
28. Dhillon A, Singh A (2019) Machine learning in healthcare data analysis: a survey. J Biol Today's World 8(2):1–10
29. Omics (2020) https://en.wikipedia.org/wiki/Omics. Accessed 20 March 2020
30. Weissenbach J (2016) The rise of genomics. CR Biol 339(7–8):231–239. https://doi.org/10.1016/j.crvi.2016.05.002
31. Jou WM, Haegeman G, Ysebaert M, Fiers W (1972) Nucleotide sequence of the gene coding for the bacteriophage MS2 coat protein. Nature 237(5350):82–88
32. Kitchenham B, Brereton OP, Budgen D, Turner M, Bailey J, Linkman S (2009) Systematic literature reviews in software engineering–a systematic literature review. Inf Softw Technol 51(1):7–15
33. Tao M, Song T, Du W et al (2019) Classifying breast cancer subtypes using multiple kernel learning based on omics data. Genes 10(3):200. https://doi.org/10.3390/genes10030200
34. Liu Y (2004) Active learning with support vector machine applied to gene expression data for cancer classification. J Chem Inf Comput Sci 44(6):1936–1941. https://doi.org/10.1021/ci049810a
35. Xu X, Zhang Y, Zou L, Wang M, Li A (2012) A gene signature for breast cancer prognosis using support vector machine. In: 2012 5th international conference on biomedical engineering and informatics, IEEE, 16–18 October 2012, Chongqing, China, pp 928–931
36. Chen Y, Sun J, Huang LC et al (2015) Classification of cancer primary sites using machine learning and somatic mutations. Biomed Res Int 2015:1–9. https://doi.org/10.1155/2015/491502
37. Anaissi A, Goyal M, Catchpoole DR et al (2016) Ensemble feature learning of genomic data using support vector machine. PLoS ONE 11(6):e0157330. https://doi.org/10.1371/journal.pone.0157330
38. Cai Z, Xu D, Zhang Q et al (2015) Classification of lung cancer using ensemble-based feature selection and machine learning methods. Mol BioSyst 11(3):791–800. https://doi.org/10.1039/c4mb00659c
39. Ruan J, Jahid MJ, Gu F et al (2019) A novel algorithm for network-based prediction of cancer recurrence. Genomics 111(1):17–23. https://doi.org/10.1016/j.ygeno.2016.07.005
40. Long NP, Park S, Anh NH et al (2019) High-throughput omics and statistical learning integration for the discovery and validation of novel diagnostic signatures in colorectal cancer. Int J Mol Sci 20(2):296. https://doi.org/10.3390/ijms20020296

41. Bravo-Merodio L, Williams JA, Gkoutos GV, Acharjee A (2019) Omics biomarker identification pipeline for translational medicine. J Trans Med 17(1):155. https://doi.org/10.1186/s12967-019-1912-5
42. Moon M, Nakai K (2018) Integrative analysis of gene expression and DNA methylation using unsupervised feature extraction for detecting candidate cancer biomarkers. J Bioinform Comput Biol 16(02):1850006. https://doi.org/10.1142/S0219720018500063
43. Hamzeh O, Rueda L (2019) A gene-disease-based machine learning approach to identify prostate cancer biomarkers. In: Proceedings of the 10th ACM international conference on bioinformatics, computational biology and health informatics. Association for Computing Machinery, Niagara Falls, NY, USA, pp 633–638
44. Swan AL, Stekel DJ, Hodgman C et al (2015) A machine learning heuristic to identify biologically relevant and minimal biomarker panels from omics data. BMC Genom 16(S1):S2. https://doi.org/10.1186/1471-2164-16-S1-S2
45. Zuo Y, Cui Y, di Poto C et al (2016) INDEED: Integrated differential expression and differential network analysis of omic data for biomarker discovery. Methods 111:12–20. https://doi.org/10.1016/j.ymeth.2016.08.015
46. Ramroach S, Joshi A, John M (2020) Optimisation of cancer classification by machine learning generates enriched list of candidate drug targets and biomarkers. Mol Omics 16(2):113–125. https://doi.org/10.1039/c9mo00198k
47. Ching T, Zhu X, Garmire LX (2018) Cox-nnet: an artificial neural network method for prognosis prediction of high-throughput omics data. PLoS Comput Biol 14(4):e1006076. https://doi.org/10.1371/journal.pcbi.1006076
48. Roadknight C, Suryanarayanan D, Aickelin U et al (2015) An ensemble of machine learning and anti-learning methods for predicting tumour patient survival rates. In: 2015 IEEE international conference on data science and advanced analytics. IEEE, 19–21 Oct. 2015, Paris, France, pp 1–8
49. Spirko-Burns L, Devarajan K (2020) Supervised dimension reduction for large-scale "omics" data with censored survival outcomes under possible non-proportional hazards. IEEE/ACM Trans Comput Biol Bioinf. https://doi.org/10.1109/TCBB.2020.2965934
50. Huang Z, Zhan X, Xiang S et al (2019) Salmon: survival analysis learning with multi-omics neural networks on breast cancer. Front Genet 10:166. https://doi.org/10.3389/fgene.2019.00166
51. Lee C, Zame WR, Yoon J, van der Schaar M (2018) DeepHit: a deep learning approach to survival analysis with competing risks. In: Thirty-second AAAI conference on artificial intelligence, pp 2314–2321
52. Yousefi S, Amrollahi F, Amgad M et al (2017) Predicting clinical outcomes from large scale cancer genomic profiles with deep survival models. Sci Rep 7(1):1–11. https://doi.org/10.1038/s41598-017-11817-6
53. Argelaguet R, Velten B, Arnol D et al (2018) Multi-omics factor analysis—a framework for unsupervised integration of multi-omics data sets. Mol Syst Biol 14(6):e8124. https://doi.org/10.15252/msb.20178124
54. Dimitrakopoulos C, Hindupur SK, Hafliger L et al (2018) Network-based integration of multi-omics data for prioritizing cancer genes. Bioinformatics 34(14):2441–2448. https://doi.org/10.1093/bioinformatics/bty148
55. Sun D, Li A, Tang B, Wang M (2018) Integrating genomic data and pathological images to effectively predict breast cancer clinical outcome. Comput Methods Prog Biomed 161:45–53. https://doi.org/10.1016/j.cmpb.2018.04.008
56. Torshizi AD, Petzold LR (2018) Graph-based semi-supervised learning with genomic data integration using condition-responsive genes applied to phenotype classification. J Am Med Inform Assoc 25(1):99–108. https://doi.org/10.1093/jamia/ocx032
57. Fang Z, Ma T, Tang G et al (2018) Bayesian integrative model for multi-omics data with missingness. Bioinformatics 34(22):3801–3808. https://doi.org/10.1093/bioinformatics/bty775
58. Gevaert O, Smet FD, Timmerman D et al (2006) Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks. Bioinformatics 22(14):e184–e190
59. Subhani MM, Anjum A, Koop A, Antonopoulos N (2016) Clinical and genomics data integration using meta-dimensional approach. In: 2016 IEEE/ACM 9th international conference on utility and cloud computing (UCC). IEEE, Shanghai, China, pp 416–421
60. Savage RS, Yuan Y (2016) Predicting chemoinsensitivity in breast cancer with 'omics/digital pathology data fusion. R Soc Open Sci 3(2):140501. https://doi.org/10.1098/rsos.140501
61. Kim M, Oh I, Ahn J (2018) An improved method for prediction of cancer prognosis by network learning. Genes 9(10):478. https://doi.org/10.3390/genes9100478
62. Bica I, Velickovic P, Xiao H, Li P (2018) Multi-omics data integration using cross-modal neural networks. In: ESANN 2018 proceedings, European symposium on artificial neural networks, computational intelligence and machine learning. Bruges, Belgium, pp 385–390
63. Klau S, Jurinovic V, Hornung R et al (2018) Priority-Lasso: a simple hierarchical approach to the prediction of clinical outcome using multi-omics data. BMC Bioinform 19(1):322. https://doi.org/10.1186/s12859-018-2344-6
64. Rappoport N, Shamir R (2018) NEMO: cancer subtyping by integration of partial multi-omic data. Bioinformatics 35(18):3348–3356. https://doi.org/10.1093/bioinformatics/btz058
65. Wu D, Wang D, Zhang MQ, Gu J (2015) Fast dimension reduction and integrative clustering of multi-omics data using low-rank approximation: application to cancer molecular classification. BMC Genom 16(1):1022. https://doi.org/10.1186/s12864-015-2223-8
66. Lopes MB, Veríssimo A, Carrasquinha E et al (2018) Ensemble outlier detection and gene selection in triple-negative breast cancer data. BMC Bioinform 19(1):168. https://doi.org/10.1186/s12859-018-2149-7
67. Chang SW, Abdul-Kareem S, Merican AF, Zain RB (2013) Oral cancer prognosis based on clinicopathologic and genomic markers using a hybrid of feature selection and machine learning methods. BMC Bioinform 14(1):170. https://doi.org/10.1186/1471-2105-14-170
68. Liang M, Li Z, Chen T, Zeng J (2015) Integrative data analysis of multi-platform cancer data with a multimodal deep learning approach. IEEE/ACM Trans Comput Biol Bioinform 12(4):928–937. https://doi.org/10.1109/TCBB.2014.2377729
69. Islam MM, Wang Y, Hu P (2018) deep learning models for predicting phenotypic traits and diseases from omics data. In: Artificial intelligence—emerging trends and applications. IntechOpen, pp 333–351
70. Exarchos KP, Goletsis Y, Fotiadis DI (2011) Multiparametric decision support system for the prediction of oral cancer reoccurrence. IEEE Trans Inf Technol Biomed 16(6):1127–1134. https://doi.org/10.1109/TITB.2011.2165076
71. Park C, Ahn J, Kim H, Park S (2014) Integrative gene network construction to analyze cancer recurrence using semi-supervised learning. PLoS ONE 9(1):e86309. https://doi.org/10.1371/journal.pone.0086309
72. Kim S, Jhong JH, Lee J, Koo JY (2017) Meta-analytic support vector machine for integrating multiple omics data. BioData Min 10(1):2. https://doi.org/10.1186/s13040-017-0126-8
73. Mallik S, Bhadra T, Maulik U (2017) Identifying epigenetic biomarkers using maximal relevance and minimal redundancy based feature selection for multi-omics data. IEEE Trans Nanobiosci 16(1):3–10. https://doi.org/10.1109/TNB.2017.2650217

74. Long NP, Jung KH, Anh NH et al (2019) An integrative data mining and omics-based translational model for the identification and validation of oncogenic biomarkers of pancreatic cancer. Cancers 11(2):155. https://doi.org/10.3390/cancers11020155
75. El-Manzalawy Y (2018) CCA based multi-view feature selection for multi-omics data integration. In: 2018 IEEE conference on computational intelligence in bioinformatics and computational biology (CIBCB). IEEE, St. Louis, MO, USA, pp 1–8
76. Ma S, Ren J, Fenyö D (2016) Breast cancer prognostics using multi-omics data. In: AMIA summits on translational science proceedings, pp 52–59
77. Chen YC, Ke WC, Chiu HW (2014) Risk classification of cancer survival using ANN with gene expression data from multiple laboratories. Comput Biol Med 48:1–7. https://doi.org/10.1016/j.compbiomed.2014.02.006
78. Zhu B, Song N, Shen R et al (2017) Integrating clinical and multiple omics data for prognostic assessment across human cancers. Sci Rep 7(1):1–13. https://doi.org/10.1038/s41598-017-17031-8
79. Kim D, Li R, Dudek SM, Ritchie MD (2015) Predicting censored survival data based on the interactions between meta-dimensional omics data in breast cancer. J Biomed Inform 56:220–228. https://doi.org/10.1016/j.jbi.2015.05.019
80. Poirion OB, Chaudhary K, Garmire LX (2018) Deep Learning data integration for better risk stratification models of bladder cancer. AMIA Summits on Translational Science Proceedings. pp 197–206
81. Incoronato M, Aiello M, Infante T et al (2017) Radiogenomic analysis of oncological data: a technical survey. Int J Mol Sci 18(4):805. https://doi.org/10.3390/ijms18040805
82. Acharya UR, Hagiwara Y, Sudarshan VK et al (2018) Towards precision medicine: from quantitative imaging to radiomics. J Zhejiang Univ Sci B 19(1):6–24. https://doi.org/10.1631/jzus.B1700260
83. Li H, Zhu Y, Burnside ES et al (2016) Quantitative MRI radiomics in the prediction of molecular classifications of breast cancer subtypes in the TCGA/TCIA data set. NPJ Breast Cancer 2(1):1–10. https://doi.org/10.1038/npjbcancer.2016.12
84. Zhou H, Dong D, Chen B et al (2018) Diagnosis of distant metastasis of lung cancer: based on clinical and radiomic features. Trans Oncol 11(1):31–36. https://doi.org/10.1016/j.tranon.2017.10.010
85. Takahashi S, Takahashi W, Tanaka S et al (2019) Radiomics analysis for glioma malignancy evaluation using diffusion kurtosis and tensor imaging. Int J Radiat Oncol Biol Phys 105(4):784–791. https://doi.org/10.1016/j.ijrobp.2019.07.011
86. Li K, Xiao J, Yang J et al (2019) Association of radiomic imaging features and gene expression profile as prognostic factors in pancreatic ductal adenocarcinoma. Am J Trans Res 11(7):4491–4499
87. Clifton H, Vial A, Miller A et al (2019) Using machine learning applied to radiomic image features for segmenting tumour structures. In: 2019 Asia-Pacific signal and information processing association annual summit and conference (APSIPA ASC). IEEE, Lanzhou, China, pp 1981–1988
88. Bae S, Choi YS, Ahn SS et al (2018) Radiomic MRI phenotyping of glioblastoma: improving survival prediction. Radiology 289(3):797–806. https://doi.org/10.1148/radiol.2018180200
89. Chaddad A, Sabri S, Niazi T, Abdulkarim B (2018) Prediction of survival with multi-scale radiomic analysis in glioblastoma patients. Med Biol Eng Comput 56(12):2287–2300. https://doi.org/10.1007/s11517-018-1858-4
90. Kaissis G, Ziegelmayer S, Lohöfer F et al (2019) A machine learning algorithm predicts molecular subtypes in pancreatic ductal adenocarcinoma with differential response to gemcitabine-based versus FOLFIRINOX chemotherapy. PLoS ONE 14(10):1–16. https://doi.org/10.1371/journal.pone.0218642
91. Sun W, Jiang M, Dang J et al (2018) Effect of machine learning methods on predicting NSCLC overall survival time based on radiomics analysis. Radiat Oncol 13(1):1–8. https://doi.org/10.1186/s13014-018-1140-9
92. D'Amico NC, Grossi E, Valbusa G et al (2020) A machine learning approach for differentiating malignant from benign enhancing foci on breast MRI. Eur Radiol Exp 4(1):5. https://doi.org/10.1186/s41747-019-0131-4
93. Ferreira Junior JR, Koenigkam-Santos M, Cipriano FEG et al (2018) Radiomics-based features for pattern recognition of lung cancer histopathology and metastases. Comput Methods Prog Biomed 159:23–30. https://doi.org/10.1016/j.cmpb.2018.02.015
94. Zhang Y, Li A, He J, Wang M (2020) A novel MKL method for GBM prognosis prediction by integrating histopathological image and multi-omics data. IEEE J Biomed Health Inform 24(1):171–179. https://doi.org/10.1109/JBHI.2019.2898471
95. Lao J, Chen Y, Li ZC et al (2017) A deep learning-based radiomics model for prediction of survival in glioblastoma multiforme. Sci Rep 7(1):1–8. https://doi.org/10.1038/s41598-017-10649-8
96. Fu Y, Liu X, Yang Q et al (2019) Radiomic features based on MRI for prediction of lymphovascular invasion in rectal cancer. Chin J Acad Radiol 2(1–2):13–22. https://doi.org/10.1007/s42058-019-00016-z
97. Chufal KS, Ahmad I, Pahuja AK et al (2019) Application of artificial neural networks for prognostic modeling in lung cancer after combining radiomic and clinical features. Asian J Oncol 5(02):050–055. https://doi.org/10.1055/s-0039-3401438
98. Wei B, Han Z, He X, Yin Y (2017) Deep learning model based breast cancer histopathological image classification. In: 2017 IEEE 2nd international conference on cloud computing and big data analysis (ICCCBDA), IEEE, April 28–30, Chengdu, China, pp 348–353
99. Lu H, Wang H, Yoon SW (2019) A dynamic gradient boosting machine using genetic optimizer for practical breast cancer prognosis. Expert Syst Appl 116:340–350. https://doi.org/10.1016/j.eswa.2018.08.040
100. Yang H, Cao H, He T et al (2018) Multilevel heterogeneous omics data integration with kernel fusion. Brief Bioinform 21(1):156–170. https://doi.org/10.1093/bib/bby115
101. Li L, Jiang W, Li X et al (2005) A robust hybrid between genetic algorithm and support vector machine for extracting an optimal feature gene subset. Genomics 85(1):16–23. https://doi.org/10.1016/j.ygeno.2004.09.007
102. Ram PK, Kuila P (2019) Feature selection from microarray data: genetic algorithm based approach. J Inform Optim Sci 40(8):1599–1610. https://doi.org/10.1080/02522667.2019.1703260
103. Fortino V, Scala G, Greco D (2020) Feature set optimization in biomarker discovery from genome scale data. Bioinformatics, btaa144. https://doi.org/10.1093/bioinformatics/btaa144
104. Kečo D, Subasi A, Kevric J (2018) Cloud computing-based parallel genetic algorithm for gene selection in cancer classification. Neural Comput Appl 30(5):1601–1610. https://doi.org/10.1007/s00521-016-2780-z
105. Yu H, Ni J, Zhao J (2013) ACOSampling: an ant colony optimization-based undersampling method for classifying imbalanced DNA microarray data. Neurocomputing 101:309–318. https://doi.org/10.1016/j.neucom.2012.08.018
106. Xu J, Wu P, Chen Y et al (2019) A hierarchical integration deep flexible neural forest framework for cancer subtype classification by integrating multi-omics data. BMC Bioinform 20(1):1–11. https://doi.org/10.1186/s12859-019-3116-7
107. Sangaralingam A, Dayem Ullah AZ, Marzec J et al (2019) "Multi-omic" data analysis using O-miner. Brief Bioinform 20(1):130–143. https://doi.org/10.1093/bib/bbx080

108. Colaprico A, Silva TC, Olsen C et al (2016) TCGAbiolinks: an R/bioconductor package for integrative analysis of TCGA data. Nucl Acids Res 44(8):e71. https://doi.org/10.1093/nar/gkv1507
109. Yu KH, Fitzpatrick MR, Pappas L et al (2018) Omics analysis system for precision oncology (OASISPRO): a web-based omics analysis tool for clinical phenotype prediction. Bioinformatics 34(2):319–320. https://doi.org/10.1093/bioinformatics/btx572
110. Martínez-Mira C, Conesa A, Tarazona S (2018) MOSim: multi-omics simulation in R. *bioRxiv*. https://doi.org/10.1101/421834
111. Cumbo F, Fiscon G, Ceri S et al (2017) TCGA2BED: extracting, extending, integrating, and querying the cancer genome atlas. BMC Bioinform 18(1):6. https://doi.org/10.1186/s12859-016-1419-5
112. Ulfenborg B (2019) Vertical and horizontal integration of multi-omics data with miodin. BMC Bioinform 20(1):649. https://doi.org/10.1101/431429
113. Deng M, Brägelmann J, Schultze JL, Perner S (2016) Web-TCGA: an online platform for integrated analysis of molecular cancer data sets. BMC Bioinform 17(1):72. https://doi.org/10.1186/s12859-016-0917-9
114. Hernandez-Ferrer C, Ruiz-Arenas C, Beltran-Gomila A, González JR (2017) MultiDataSet: an R package for encapsulating multiple data sets with application to omic data integration. BMC Bioinform 18(1):36. https://doi.org/10.1186/s12859-016-1455-1
115. Singh A, Shannon CP, Gautier B et al (2018) DIABLO: from multi-omics assays to biomarker discovery, an integrative approach. *bioRxiv*, 067611. https://doi.org/10.1101/067611
116. Wang YE, Kutnetsov L, Partensky A et al (2017) WebMeV: a cloud platform for analyzing and visualizing cancer genomic data. Can Res 77(21):e11–e14. https://doi.org/10.1158/0008-5472.CAN-17-0802
117. Zhu Y, Qiu P, Ji Y (2014) TCGA-assembler: open-source software for retrieving and processing TCGA data. Nat Methods 11(6):599–600
118. Wei L, Jin Z, Yang S et al (2018) TCGA-assembler 2: software pipeline for retrieval and processing of TCGA/CPTAC data. Bioinformatics 34(9):1615–1617. https://doi.org/10.1093/bioinformatics/btx812
119. Xie B, Yuan Z, Yang Y et al (2018) MOBCdb: a comprehensive database integrating multi-omics data on breast cancer for precision medicine. Breast Cancer Res Treat 169(3):625–632. https://doi.org/10.1007/s10549-018-4708-z
120. Chen D, Zhang F, Zhao Q, Xu J (2019) OmicsARules: a R package for integration of multi-omics datasets via association rules mining. BMC Bioinform 20(1):1–8. https://doi.org/10.1186/s12859-019-3171-0
121. Koh HWL, Fermin D, Vogel C et al (2019) iOmicsPASS: network-based integration of multiomics data for predictive subnetwork discovery. NPJ Syst Biol Appl 5(1):1–10. https://doi.org/10.1038/s41540-019-0099-y
122. Fisch KM, Meißner T, Gioia L et al (2015) Omics pipe: a community-based framework for reproducible multi-omics data analysis. Bioinformatics 31(11):1724–1728. https://doi.org/10.1093/bioinformatics/btv061
123. Jang Y, Yu N, Seo J et al (2016) MONGKIE: an integrated tool for network analysis and visualization for multi-omics data. Biol Direct 11(1):10. https://doi.org/10.1186/s13062-016-0112-y
124. Polpitiya AD, Qian WJ, Jaitly N et al (2008) DAnTE: a statistical tool for quantitative analysis of -omics data. Bioinformatics 24(13):1556–1558. https://doi.org/10.1093/bioinformatics/btn217
125. Eren AM, Esen OC, Quince C et al (2015) Anvi'o: an advanced analysis and visualization platformfor 'omics data. PeerJ 3:e1319. https://doi.org/10.7717/peerj.1319
126. Guhlin J, Silverstein KAT, Zhou P et al (2017) ODG: omics database generator—a tool for generating, querying, and analyzing multi-omics comparative databases to facilitate biological understanding. BMC Bioinform 18(1):367. https://doi.org/10.1186/s12859-017-1777-7
127. Surujon D, van Opijnen T (2020) ShinyOmics: collaborative exploration of omics-data. BMC Bioinform 21(1):1–8. https://doi.org/10.1186/s12859-020-3360-x
128. Blatti C, Emad A, Berry MJ et al (2020) Knowledge-guided analysis of "omics" data using the KnowEnG cloud platform. PLoS Biol 18(1):e3000583. https://doi.org/10.1371/journal.pbio.3000583
129. Zhao S, Prenger K, Smith L et al (2013) Rainbow: a tool for large-scale whole-genome sequencing data analysis using cloud computing. BMC Genom 14(1):425. https://doi.org/10.1186/1471-2164-14-425
130. Chiesa M, Maioli G, Colombo GI, Piacentini L (2020) GARS: genetic algorithm for the identification of a robust subset of features in high-dimensional datasets. BMC Bioinform 21(1):54. https://doi.org/10.1186/s12859-020-3400-6
131. Mohammed A, Biegert G, Adamec J, Helikar T (2017) CancerDiscover: an integrative pipeline for cancer biomarker and cancer class prediction from high-throughput sequencing data. Oncotarget 9(2):2565–2573