Archives of Computational Methods in Engineering
https://doi.org/10.1007/s11831-022-09821-9

SURVEY ARTICLE

A Systematic Review on Biomarker Identification for Cancer Diagnosis and Prognosis in Multi-omics: From Computational Needs to Machine Learning and Deep Learning

Arwinder Dhillon[1] · Ashima Singh[1] · Vinod Kumar Bhalla[1]

Received: 2 May 2022 / Accepted: 5 September 2022

Abstract

Biomarkers, also known as biological markers, are substances like transcripts, deoxyribonucleic acid (DNA), genes, proteins, and metabolites that indicate whether a biological activity is normal or abnormal. Markers play an essential role in diagnosing and prognosis of diseases like cancer, diabetes, and Alzheimer's. In past years, in healthcare, an enormous amount of omics data, including genomics, proteomics, transcriptomic, metabolomics, and interatomic data, is becoming available, which helps researchers to find markers or signatures needed for disease diagnosis and prognosis and to provide the best potential course of therapy. Furthermore, integrative omics, often known as multi-omics data, are also proliferating in biomarker analysis. Therefore, various computational methods in healthcare engineering, including machine learning (ML) and deep learning (DL), have emerged to identify the markers from the complex multi-omics data. This study examines the current state of the art and computational methods, including feature selection strategies, ML and DL approaches, and accessible tools to uncover markers in single and multi-omics data. The underlying challenges, recurring problems, limitations of computational techniques, and future approaches in biomarker research have been discussed.

1 Introduction

With the initiation of precision medication and treatment therapy, genes have become increasingly popular for accurate diagnosis and prognosis of diseases in healthcare [1]. Biomarkers are commonly known as biological markers and biomarkers, that is, the identifiers that can be used to classify a biological event or condition and track certain biological events or processes. Due to their properties, genes, transcripts, proteins, and metabolites are categorized as biomarkers. Biomarkers are of seven types, including risk (markers showing a risk of getting a disease), diagnostic (markers confirming the existence of disease), prognostic (markers predicting the recurrence of disease), predictive (marker used to detect the reaction of the patient to specific therapy), monitoring (markers that are monitored periodically), safety (markers used to measure the toxicity before and after treatment) and response biomarkers (markers use to measure the response) [2]. This study is focused only on diagnostic, prognostic, and predictive biomarkers. Circular RNAs are recently identified diagnostic markers in Hepatocellular carcinoma patients [3]. Further, long non-coding RNA (lncRNA), including H19 and UCA1, are recognized as diagnostic and prognostic markers in gastric cancer [4]. Figure 1 shows the example of some common genes of different types of cancers. These genes are needed in the diagnosis and prognosis of cancer. The main focus of this survey article is on diagnostic, prognostic, and predictive biomarkers identification. Numerous cutting-edge innovations, like next-generation sequencing and microarray technologies, have appeared in the last couple of decades, entering a new age of omics in identifying biomarkers [5]. A large volume of omics data, including genome, transcriptome, proteome, and metabolome, have been created and used in various projects like The cancer genome portal (TCGA) [6], Therapeutically Applicable Research to Generate Effective Treatments (TARGET) [7] and International Cancer Genome

✉ Arwinder Dhillon
adhillon_phd19@thapar.edu

Ashima Singh
ashima@thapar.edu

Vinod Kumar Bhalla
vkbhalla@thapar.edu

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

Published online: 23 September 2022

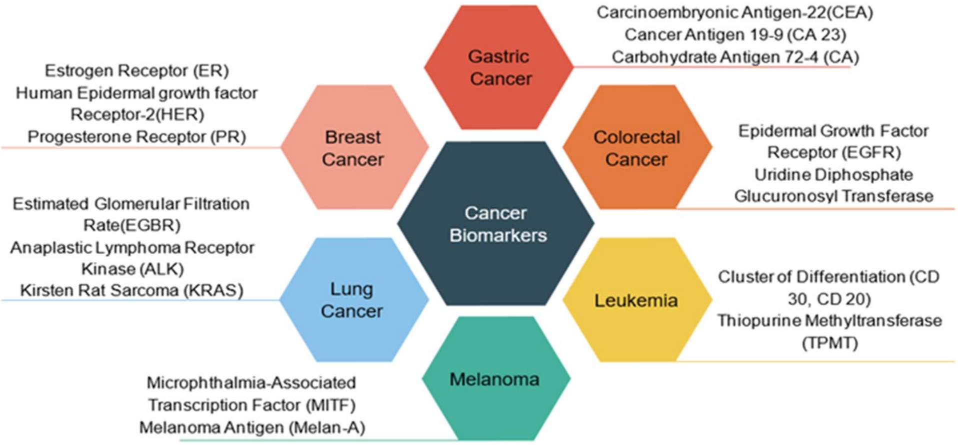

Gastric Cancer
Carcinoembryonic Antigen-22(CEA)
Cancer Antigen 19-9 (CA 23)
Carbohydrate Antigen 72-4 (CA)
Breast Cancer
Estrogen Receptor (ER)
Human Epidermal growth factor Receptor-2(HER)
Progesterone Receptor (PR)
Colorectal Cancer
Epidermal Growth Factor Receptor (EGFR)
Uridine Diphosphate Glucuronosyl Transferase
Cancer Biomarkers
Lung Cancer
Estimated Glomerular Filtration Rate(EGBR)
Anaplastic Lymphoma Receptor Kinase (ALK)
Kirsten Rat Sarcoma (KRAS)
Leukemia
Cluster of Differentiation (CD 30, CD 20)
Thiopurine Methyltransferase (TPMT)
Melanoma
Microphthalmia-Associated Transcription Factor (MITF)
Melanoma Antigen (Melan-A)

Fig. 1 Examples of common cancer biomarkers

Consortium (ICGC) [8]. These omics resources are used to identify signatures for disease diagnosis and prediction. Sometimes, it is impossible to identify biomarkers using a single type of omics data [9]. Consequently, integrated omics or multi-omics is required for the discovery of biomarkers. While the accessibility of multi-omics makes it a lot easier to assess markers/signatures for disease diagnosis and prognosis, identifying biomarkers that can accurately recognize or detect diseases in the presence of tens of millions of genes and billions of variants is still a challenging task [10]. The complexity of multi-omics data can be handled by computational methods in healthcare engineering [11], including machine learning (ML) and deep learning (DL) approaches can be employed for the identification of biomarkers. ML and DL technologies have been reviewed for biomarker identification using single omics and multi-omics datasets in this survey article. Features extracted and gene-prioritization are treated as biomarkers that are further passed to ML and DL for disease prognosis and diagnosis.

1.1 Motivation and Contribution

Identifying biomarkers is extremely difficult because of the enormous size of multi-omics datasets in healthcare science. Therefore, there is a need to study the literature on biomarker identification using multi-omics data vastly. The contributions of this research are:

- A thorough analysis was undertaken to examine the importance of current approaches in improving biomarker identification.
- The study of existing feature selection techniques for biomarker identification, ML and DL techniques for diagnostic, prognostic, and predictive biomarker identification using omics and multi-omics data is done.
- Tools required for biomarker identification using multi-omics data, which can be easily accessible by the users, are reviewed.
- Based on available features, emerging methods and tools are compared. For research groups and data scientists, the comprehensive analysis aids in the selection of future research directions.

1.2 Existing Studies and Our Research

In recent years, several authors have conducted surveys on biomarker identification. For example, Swan et al. [12] offers a study on identifying biomarker using proteomics data with the help of ML. Qin et al. [13] proposed ML algorithms to identify predictive biomarkers as molecular networks using interatomic data. Further, popular interatomic resources required for performing the experiments are also discussed. Jagga et al. [14] presented various ML and feature extraction techniques to discover diagnostic and prognostic markers. Popular omics resources and projects are also deliberated. Dragani et al. [15] reviewed various ML algorithms to discover diagnostic biomarkers required for early cancer prediction. Shi et al. [16] presented various machine learning algorithms, including supervised, unsupervised, and clustering algorithms, to identify diagnostic, prognostic, and predictive biomarkers using integrated omics data. Kaur et al. [17] identified the biomarkers using various machine learning diagnostic, predictive, and prognostic biomarkers identification tools. The comparison of the proposed work with the existing studies is shown in Table 1.

Following a review of current surveys, it was discovered that omics data tools are often used to address diagnostic biomarkers. The current ML, DL, and feature selection approach for discovering prognostic, diagnostic, and predictive biomarkers using omics and multi-omics data analysis must be summarized. This survey incorporates current methods and tools studies and is an improvement on previous studies. The comparison of the current state-of-the-art work with our presented research is shown in Table 2 below.

Table 1 Comparison of the proposed research with existing biomarker studies

Author [Ref]	Omics	Multi-omics	Omics resources	Biomarker identification			ML	DL	Tools
				Diagnostic	Prognostic	Predictive			
Swan et al. [12]	✓	✗	✗	✓	✗	✗	✓	✗	✗
Qin et al. [13]	✓	✗	✓	✗	✗	✓	✓	✗	✗
Jagga et al. [14]	✓	✗	✓	✓	✓	✗	✓	✗	✗
Dragani et al. [15]	✓	✗	✗	✓	✗	✗	✓	✗	✗
Shi et al. [16]	✗	✓	✓	✓	✓	✓	✓	✗	✗
Kaur et al. [17]	✗	✓	✗	✓	✓	✓	✓	✗	✓
Our survey	✓	✓	✓	✓	✓	✓	✓	✓	✓

1.3 Structure of Survey Paper

The study is divided into seven parts. Section 2 presents background information, including a description of biomarkers and their various types, multi-omics data, and learning used for biomarker identification. Section 3 discusses various research questions and review techniques. A systematic analysis of existing approaches for biomarker identification using omics and multi-omics data is presented in Sect. 4. Based on general features, the techniques are evaluated and defined. Section 5 illustrates different biomarker identification tools that are currently available. The analysis is outlined in Sect. 6, which includes open problems and future study directions. Section 7 brings the study to a conclusion and suggests future studies.

2 Background

2.1 Biomarkers

Biomarkers are the molecules like genes, DNA, proteins, and metabolite that signify whether a process going on in the body is regular or irregular, and it can be used as a symptom of any disease or disorder. Biomarkers are generated from the cancer tissue, and they can be present in any part of the body, including stool, blood, tumor tissue, urine, body fluids, and any other tissue or cell. Biomarkers are found in every disease, including cancer, multiple sclerosis, diabetes, and heart diseases [18]. There are seven types of biomarkers, including risk, diagnostic, prognostic, predictive, monitoring, safety, and response biomarkers. Figure 2 shows the types of biomarkers, and their explanation is described below:

2.1.1 Risk Biomarkers

A biomarker shows the risk of getting a disease or health issue in someone who might not currently have the disease or health problem. A genetic biomarker that determines whether a person has an elevated chance of contracting cancer later in life is an example of a risk biomarker. Risk biomarkers are most useful in clinical practice for guiding prevention measures. One of the commonly identified risk biomarkers is BRCA1/2 mutation, which assesses the probability of producing breast carcinoma [2].

2.1.2 Diagnostic Biomarkers

A marker that predicts or confirms the existence of a disorder of interest or classifies people according to the disease subtype is a diagnostic marker. For example, diffuse large B-cell lymphoma patients can be divided into subgroups of distinct tumor cell signatures using gene expression profiling as a diagnostic biomarker [19].

2.1.3 Prognostic Biomarkers

A prognostic biomarker predicts the occurrence of a potential clinical condition, disease recurrence, or relapse in an identified sample [20]. Biomarkers, including tumor size, the percentage of lymph nodes active for tumor cells, and the existence of malignancy, have been used to predict prognosis in the future. High low-density lipoproteins (LDL) cholesterol is an example of a prognostic marker for a person who recently suffered from a heart attack [2].

2.1.4 Predictive Biomarkers

A predictive biomarker is a marker used as a test to classify people who are more likely to react to a particular medicinal substance or chemical product. A symptomatic gain may increase longevity, or an adverse effect may be the result [21]. A predictive biomarker is considered a gene prioritization problem where the gene can signify the occurrence of some particle disease with some known disease genes [16].

Table 2 Popular omics databases and repositories

	Omics type	Links	Description
Genomics	NCBI Genome	https://www.ncbi.nlm.nih.gov/genome	This database includes complete information of the genomic data, including maps, assemblies, annotations, and chromosomes
	GOLD	https://gold.jgi.doe.gov/	It provides complete information about genome sequencing programs, as well as the metadata associated with them
	JGI	https://genome.jgi.doe.gov/portal/	It provides access to all the genomic databases along with their annotations
	EBI's Ensembl	http://asia.ensembl.org/index.html	A browser for accessing and downloading genomic data of humans, mouse and other species
	GDV	https://www.ncbi.nlm.nih.gov/genome/gdv/	A browser used for visualizing, exploring and retrieving genomic data of humans and integrate genomic data from various sources
	dbGAP	https://www.ncbi.nlm.nih.gov/gap/	A database of complete interaction between genomic and phenotype in humans
	ENCODE	https://www.encodeproject.org/	It provides complete information about genomic data of humans including mapping of DNA elements and regulatory elements
	dbVar	https://www.ncbi.nlm.nih.gov/dbvar	A database of structural variation of human genomics and involve insertions, deletions, inversions and complex variants
	UCSC Genome Browser	https://genome.ucsc.edu/	An online tool to analyze, download, and visualize genomic data
Transcriptome	Array Express	https://www.ebi.ac.uk/arrayexpress/	It is a repository of gene expression dataset holding databases from all microarray platforms
	GEO	https://www.ncbi.nlm.nih.gov/gds	It is a data repository of genomics allowing downloading for various gene expression datasets
	BioXpress	https://hive.biochemistry.gwu.edu/bioxpress	A database of gene expression and miRNAs in which the expression levels are mapped to their genes,
	Gene Expression Atlas	https://www.ebi.ac.uk/gxa/home	It is a database of gene expression profiles collected under different biological conditions
	GEA	https://www.ddbj.nig.ac.jp/gea/	A database of genetic, genomic and sequencing data including microarray profiles
Proteomics	PRIDE	https://www.ebi.ac.uk/pride/archive/	An online public available large data repository of mass spectrometry data based on proteomic data
	YRC PDR	http://www.yeastrc.org/pdr/	A protein data repository of images database including localization of proteins in the image
	Peptide Atlas	http://www.peptideatlas.org/	It is a multi-organism, freely open database of peptides discovered through tandem mass spectrometry proteomics
	GPMD	https://gpmdb.thegpm.org/	A repository of evidence for detectingproteins, and peptides using advanced tandem mass spectrometry-based proteomics
	ProteomicsDB	https://www.proteomicsdb.org/	An online public database of mass spectrometry protein data
	Human Proteome Map	http://www.humanproteomemap.org/	A database developed by integrating sequencing results of peptides
Metabolome	HMDB	https://hmdb.ca/	A freely accessible database including metabolites information
	Human MetaboLights	https://www.ebi.ac.uk/metabolights/	A database of derived information and metabolomics experiments
	BiGG	http://bigg.ucsd.edu/	A database of metabolites and pathways developed for humans and other different species
	MetabolomeExpress	https://www.metabolome-express.org/	A public repository for GC/MS metabolomics datasets to be processed, interpreted, and shared

Table 2 (continued)

	Omics type	Links	Description
Interatomic	METLIN	https://metlin.scripps.edu/	A repository for mass spectrometry metabolite data
	STRING	https://string-db.org/	A freely available database of protein–protein interaction
	IntAct	https://www.ebi.ac.uk/intact/	It's a freely accessible open-access data system and analysis forum for molecular interactions
	KEGG	http://www.kegg.jp/	It is a repository of high-level functional data for various species

NCBI National Cancer for Biotechnology information, *GOLD* Genomics Online Database, *JGI* Joint Genome Institute, *GDV* Genome Data Viewer, *ENCODE* Encyclopedia of DNA elements, *GEA* Genomic Expression Archive, *GEO* Gene Expression Omnibus, *PRIDE* Proteomic Identification Database, *YRC PDR* Yeast Resource center protein data repository, *HMDB* Human Metabolome Database, *GPMD* Global Proteomic Machine Database

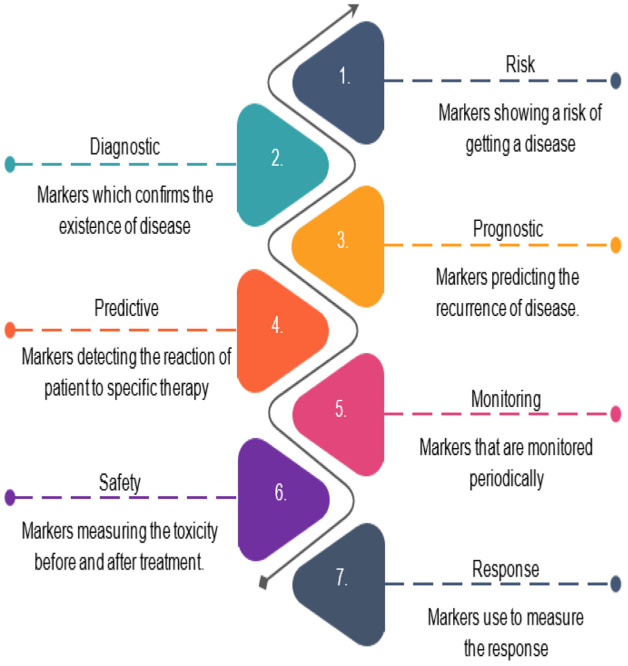

1.
Risk
Markers showing a risk of getting a disease
2.
Diagnostic
Markers which confirms the existence of disease
3.
Prognostic
Markers predicting the recurrence of disease.
4.
Predictive
Markers detecting the reaction of patient to specific therapy
5.
Monitoring
Markers that are monitored periodically
6.
Safety
Markers measuring the toxicity before and after treatment.
7.
Response
Markers use to measure the response

Fig. 2 Types of biomarkers

2.1.5 Monitoring Biomarkers

A marker that is measured periodically over time to determine disease incidences, such as the emergence of new disease symptoms, the deterioration of preexisting anomalies, or changes in clinical outcomes or particular anomalies, is a monitoring biomarker. CA 125 is an example of monitoring biomarker in ovarian cancer patients to measure disease activity or burden before and after surgery [2].

2.1.6 Safety Biomarkers

A biomarker is assessed before or after access to a therapeutic drug or an environmental agent to determine the probability, occurrence, and severity of toxicity as an adverse impact. Serum creatinine is an example of a safety biomarker in patients on medications that impair kidney function [22].

2.1.7 Response Biomarkers

A biomarker indicating patient's biological reaction to a medical substance or an environmental agent is a response biomarker. For example, plasma microRNAs act as a Hodgkin lymphoma response biomarker [23].

As biomarkers are molecules, omics and multi-omics data are required for their identification.

2.2 Multi-omics Data

In recent years, multi-omics data have been used as molecular biomarkers using the integration of omics data types, including genomic, transcriptomic, proteomic, metabolites, and interatomic, for the prognosis and diagnosis of some specific diseases [24]. The discovery of disease biomarkers with multi-omics data would aid in the stratification of various patient cohorts, but it would also include early diagnosis knowledge that may enhance patient care and possibly mitigate adverse outcomes [25]. There are different tools and techniques available for multi-omics data integration, which can be used for biomarker identification, disease diagnosis, and progression [26]. The types of omics data are discussed below.

2.2.1 Genome

The whole sequence of DNA in an organism, including all of its chromosomes, is referred to as a genome. Genomics seeks to characterize and quantify all of the genes of an organism and their interrelationships and effects on the organism. The primary goal of genomics research in medicine is to find genetic variants linked to disease, therapeutic response, and patient prognosis [25].

2.2.2 Proteome

The entire universe of proteins in the cell is called proteome. Proteomics is a technique for detecting protein expression variations in response to a particular stimulus at a specific time and determining protein structure networks at the tissue, organism, or cell level [27]. Proteomics is based on three vital technical elements: a tool for fractionating complex protein or peptide combinations, mass spectrometry (MS) for acquiring the data needed to classify specific proteins, and computational biology for analyzing and assembling the MS data [28].

2.2.3 Transcriptome

A transcriptome is a collection of mRNA, miRNA, and lncRNA molecules in which their sequence produced in a particular cell is called "transcriptome." RNA lies between proteins and DNA and acts as the primary function of DNA readouts [29]. RNA-Seq technique is used to profile the transcripts or raw data.

2.2.4 Metabolome

The metabolome contains a complete collection of small-molecule groups called metabolites, including carbohydrates, amino acids, sugars, and fatty acids. Similarly, quantitative measurements of metabolites are performed using the MS technique like proteins. Metabolomics tasks are executed at different metabolite levels, and any relative distributions and disturbances signify the disease when they occur outside of the normal range [30].

2.2.5 Interatomic

An interatomic is a multi-dimensional description of functional associations between molecules inside a cell or throughout the whole organism. A protein–protein interaction comes under this category of omics data [31]. The popular omics resources, databases, and repositories, along with their description and links, is shown in Table 2

2.3 Learning for Biomarker Identification

Machine learning analysis in biomarker identification deals with the different types of omics data and their integration for disease prediction and prognosis and guides treatment therapies based on the identified biomarkers [14]. Figure 3 shows ML and DL's workflow for the biomarker identification using multi-omics data. The steps involved in learning are data preprocessing, feature extraction, biomarker identification, and modeling, and are discussed.

2.3.1 Data Preprocessing

It is the method of transforming or encoding multi-data so that the computer can quickly process it. Data preprocessing, including data cleaning in which the missing values and noisy values are removed; data transformation in which the data is converted into some specific range using normalization and selection techniques; and data reduction in which the high-dimensional multi-omics dataset is reduced to low dimensional dataset [32]. The attribute selection and dimensionality reduction techniques are described below under feature selection and extraction.

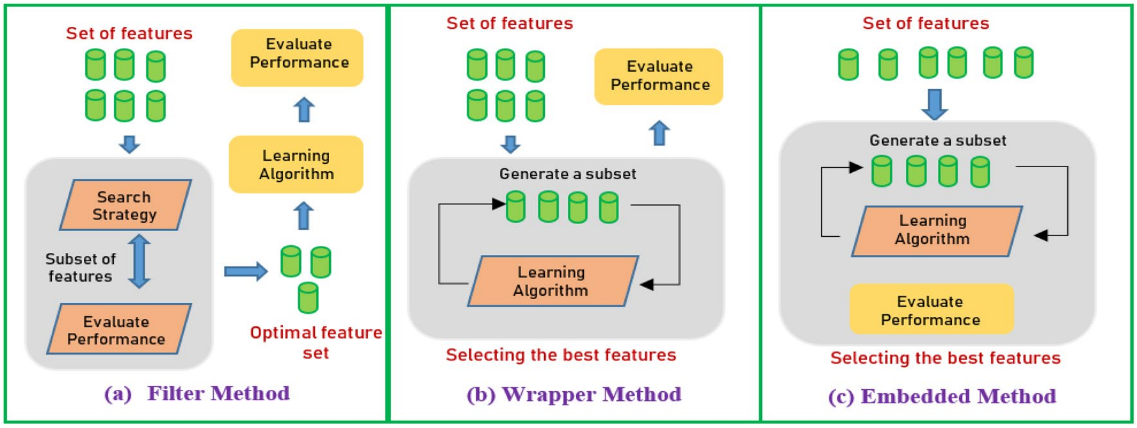


Fig. 3 Feature selection methods

2.3.2 Feature Extraction and Selection

Feature Extraction reduces the feature space of high dimensional multi-omics data to low dimensional feature space [33]. This low dimensional feature space consists of important information only required for biomarker identification, disease detection, and prognosis. There are different techniques available for the extraction of features for integrated omics, including Principal Component Analysis (PCA) [34], Canonical Correlation Analysis (CCA) [35], and Linear Discriminant analysis [35], and Non-Negative Matrix Factorization (NNMF) [36]. However, these techniques integrate only linear multi-omics data. In case non-linear integration is required, these fail to work, for example, in integrating gene expression and interatomic data. The Non-linear feature extraction techniques are required to integrate non-linear data. These are Kernel Principal Component Analysis (KPCA) [37], Locally Linear Embedding (LLE) [38], t-distributed Stochastic Neighbor Embedding (t-SNE) [39], and auto-encoders [40]. However, feature extraction only returns a subset of features. Some relevant features need to be selected for biomarker identification, which is done using feature selection.

Feature selection is a method of electing valuable and informative features by removing duplicate and noisy features [41]. Feature selection techniques are of three types, including filter, wrapper, and embedded methods and are shown in Fig. 3. The filter method works by assigning a rank to the features and selecting only higher-rank features. Different filter method techniques include Pearson Correlation Coefficient (PCC), chi-square, t-test, and Analysis of Variance [42], which works by finding the correlation between the features and target variable. In biomarker identification, chi-square and t-test are used by various researchers to rank the differentially expressed genes and select the top-ranked genes [43–45]. However, there is a disadvantage to filter selection. In the filter method, each feature works independently, i.e., they do not interact with each other. Nevertheless, there is a complex relationship between the features in omics data, so the filter method does not work correctly in this scenario. Also, the filter method works independently of the classifier, resulting in poor performance of the selected features [46]. Wrapper methods are used to overcome these disadvantages. Wrapper methods work by selecting the features iteratively and evaluating their performance using a classifier. Initially, there was no feature set. Each time a feature is added and performance is checked. This is done until the most relevant features are not selected [33]. The wrapper method selects the features in two ways, including forward feature selection and backward feature selection. Some of the standard techniques of the wrapper method are Recursive Feature Elimination (RFE) [47], Sequential Feature Elimination (SFE) [48], and Genetic Algorithms [49]. Various authors have worked on wrapper methods for biomarker identification, for example, RFE is used to identify miRNA biomarkers [50], and hybrid wrapper methods are used in multi-omics data to identify diagnostic markers [a4]. There is complete interaction between features and classifier. Therefore, it solves the problem of the filter method, but the wrapper method leads to overfitting. The embedded methods have been introduced to solve this problem. The embedded method combines the function of both the filter and wrapper method. It works by integrating the feature selection algorithm with the training algorithm and selecting the feature subset [51]. Least Absolute Shrinkage and Square Estimator (LASSO) is one of the most common techniques of feature selection which is implemented by several researchers in diagnostic and prognostic biomarker identification [52–54].

2.3.3 Modelling

Both ML and DL can be used in the modeling of a dataset. ML is a data processing technique that automates the growth of analytical models. It is a branch of artificial intelligence that allows computers to learn from their mistakes, interpret data, identify patterns, and make educated decisions with little or no human interference. [55]. ML is of four types supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning. In supervised learning, a computer is trained with well-labeled data. Some of the datasets have already been labeled with the correct answer. Afterward, the machine provides the test data, which is analyzed by the supervised learning algorithm that generates an accurate result from classified data. Supervised Learning can be a classification problem or regression problem [56]. In classification, the outcome variable is a class or categorical variable, and in regression, the outcome variable is a real value. There are different supervised learning algorithms, including SVM, Linear regression, Random Forest (RF), Adaboost, K-Nearest Neighbor (KNN), Naïve Bayes (NB), and Decision Tree [57]. On the other side, Unsupervised learning is training a computer to work on knowledge that is neither categorized nor labeled. The machine works by organizing the unlabeled data into groups or clusters based on similarities, variations, and discrepancies without any previous data knowledge. Hierarchical clustering and K-mean clustering are unsupervised learning algorithms [58]. Semi-supervised is a mixture of both supervised and semi-supervised learning, but in this, the labeled data is of minimal size compared to unlabeled data [59]. Reinforcement learning is about taking the proper steps to optimize the incentive in a given situation. Various algorithms and computers use it to determine the best possible action or direction in a given scenario [55]. *Deep Learning* is a form of ML inspired by the human brain's structure. Deep learning analyzes data using a predetermined conceptual form to draw similar results as humans. Deep learning uses a multilayered system of algorithms known as neural networks [60]. The neural network's architecture is focused on the configuration of the human brain. Neural networks can be trained to recognize trends and interpret various kinds of data in the same way our brains do. The brain attempts to compare new knowledge to existing items once we encounter it. Deep neural networks operate on the same principle. Classification, clustering, regression, and many other tasks can be performed in neural networks. We can aggregate or filter unlabeled data using neural networks based on the similarity between the samples. The are many algorithms available in DL comprising convolutional neural network (CNN), recurrent neural networks (RNN), U-net, deep belief networks (DBN), long short term memory (LSTM), and many more which help in the classification and prognosis of diseases [61]. The complete workflow of biomarker identification using ML and DL is shown in Fig. 4.

In this research, classification is used for disease diagnosis, regression for disease prognosis, and feature selection and extraction are used as biomarker identification. Once the biomarkers are identified, they are passed to machine and deep learning algorithms which further classify them into diagnostic, prognostic, and predictive markers. Their survival analysis is checked with various models for prognostic markers, including univariate cox, multi-variate cox, and LASSO model. The risk score is calculated from which higher risk markers are identified as prognostic markers. Based on this, drugs and treatment therapies can be recommended. On the other side, predictive markers are considered a gene-prioritization problem from which a biomarker can be discovered from some known disease biomarkers. Several gene-prioritized algorithms have been reviewed for the identification of predictive markers. The complete taxonomy of biomarker identification is shown in Fig. 5.

2.4 Biomarker Identification Research Evolution

The first biomarker identified was a protein biomarker discovered by Bence-Jones in 1847 in multiple myeloma patients [62]. It was approved by FDA in 1986 when they are reported again in serum markers of myeloma patients [63]. In 1867, Sir Michal Forster identified urinary amylase marker in pancreatic cancer patients. The biological marker term was introduced in 1950 and gained popularity in the 1980s [64]. Carcinoembryonic antigen (CEA) was identified by Dr. Joseph gold in 1965 and was discovered in the malignant tissues of cancer patients [65]. In the 1970s, three more markers, i.e., Cancer Antigen (CA) CA 199, CA 15-3, and CA 125, were discovered in colorectal, breast, and ovarian cancer patients. Furthermore, prostate-specific antigen (PSA) was discovered in the 1980s, and till now, a variety of biomarkers have been discovered [65]. The complete history of biomarkers is shown in Fig. 6 below. Figure 7 shows the trends of the biomarker identification using multi-omics data, which shows an increase in 2018–2021.

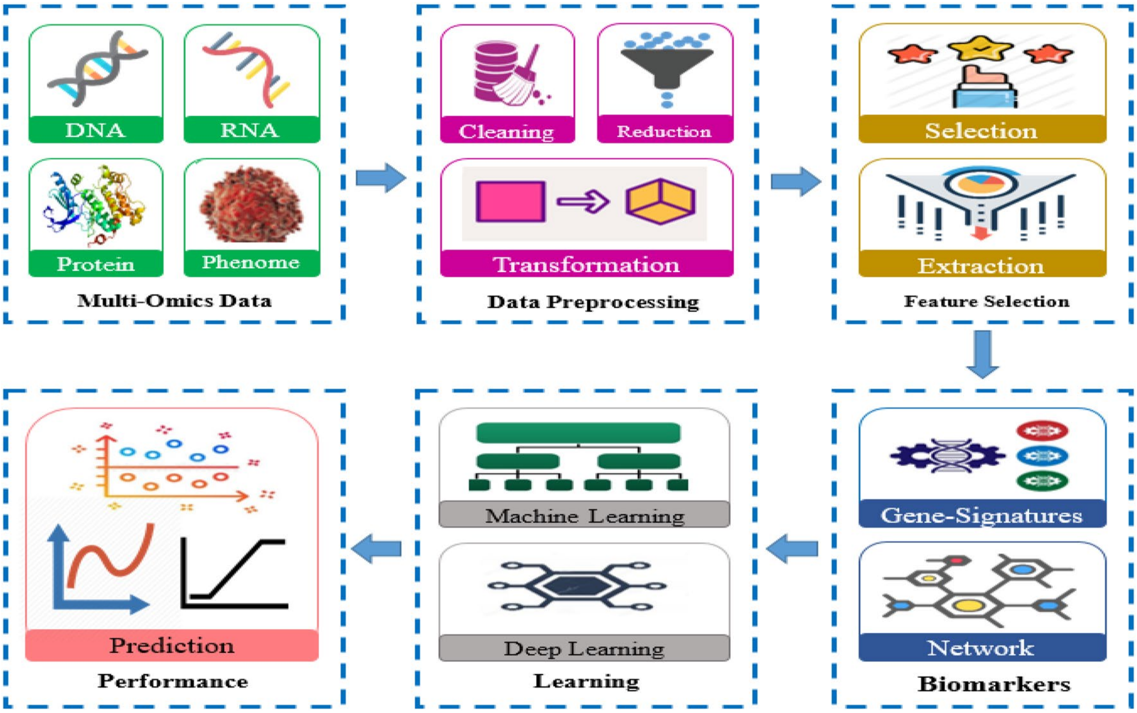


Fig. 4 Workflow of biomarker identification using ML and DL

Fig. 5 Biomarker identification taxonomy

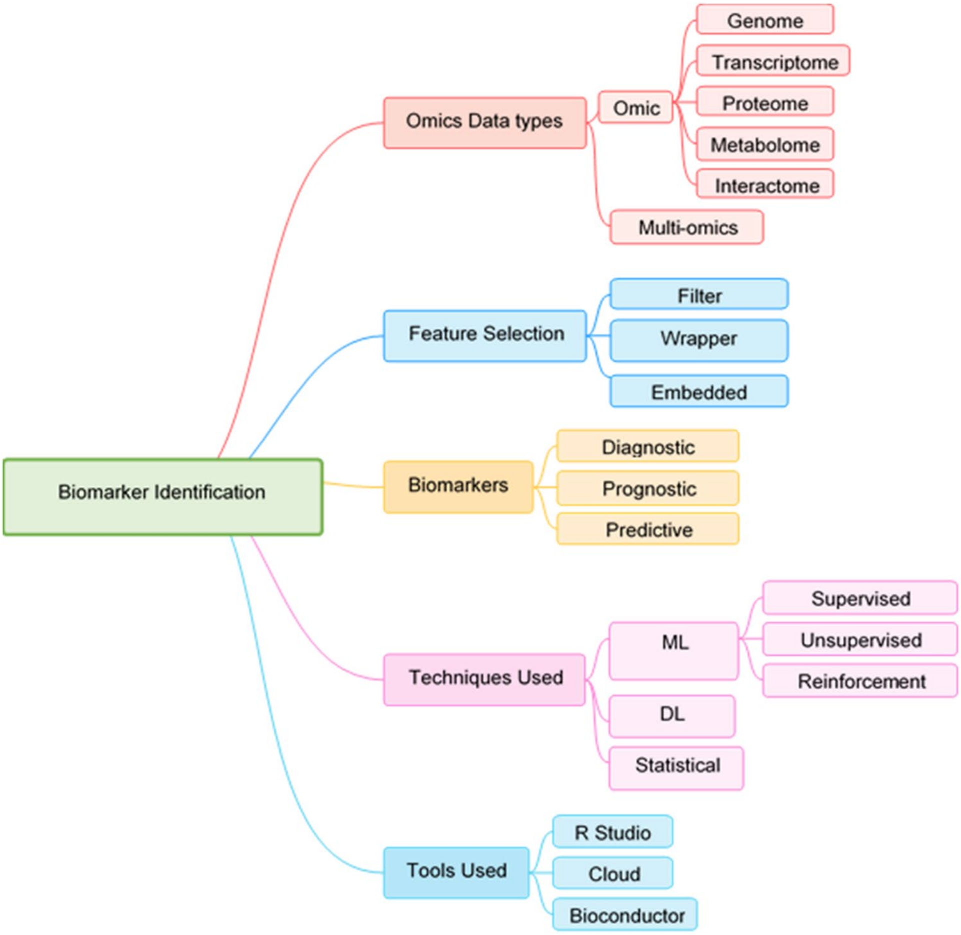
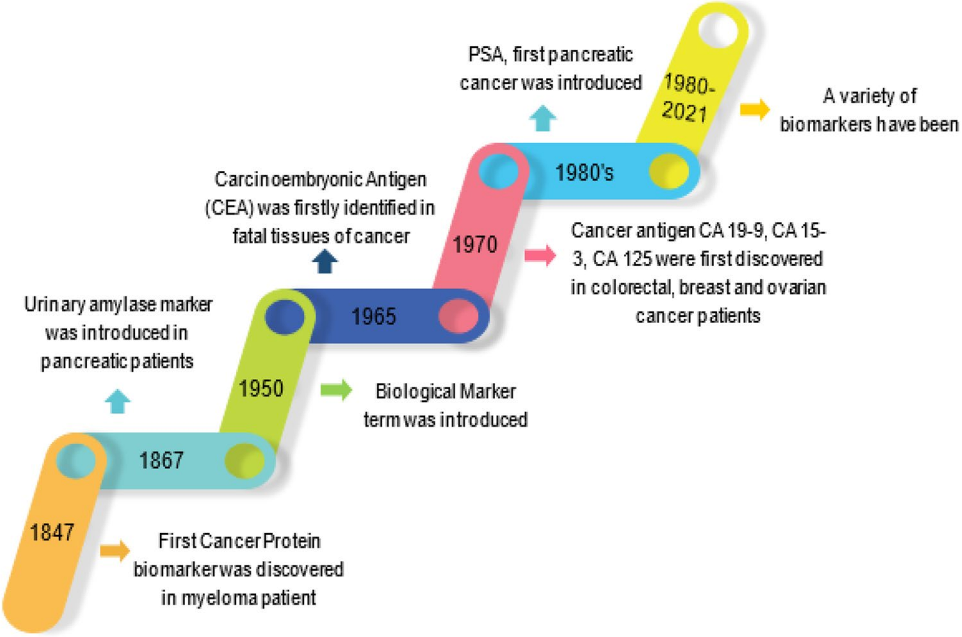


Fig. 6 History of biomarkers


1847
First Cancer Protein biomarker was discovered in myeloma patient
1867
Urinary amylase marker was introduced in pancreatic patients
1950
Biological Marker term was introduced
1965
Carcinoembryonic Antigen (CEA) was firstly identified in fatal tissues of cancer
1970
Cancer antigen CA 19-9, CA 15-3, CA 125 were first discovered in colorectal, breast and ovarian cancer patients
1980's
PSA, first pancreatic cancer was introduced
1980-2021
A variety of biomarkers have been

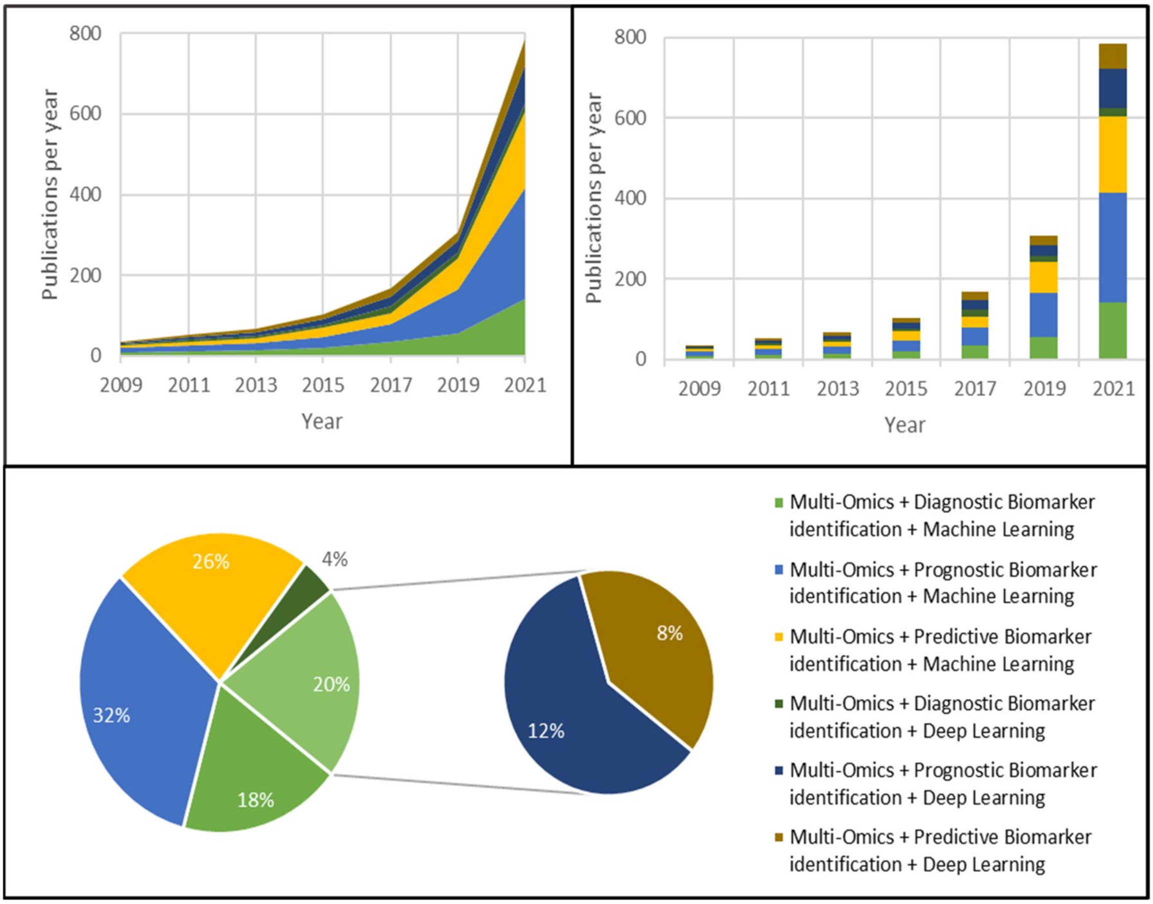


Fig. 7 Trends in biomarker identification using multi-omics data

It also shows the percentage of publications in biomarker identification using multi-omics data. In this research, classification is used for disease diagnosis, regression for disease prognosis, and feature selection and extraction are used as biomarker identification. Once the biomarkers are identified, they are passed to machine and deep learning algorithms which further classify them into diagnostic, prognostic, and predictive markers. Their survival analysis is checked with various models for prognostic markers, including univariate cox, multi-variate cox, and LASSO model. The risk score is calculated from which higher risk markers are identified as prognostic markers. Based on this, drugs and treatment therapies can be recommended. On the other side, predictive markers are considered a gene-prioritization problem from which a biomarker can be discovered from some known disease biomarkers. Several gene-prioritized algorithms have been reviewed for the identification of predictive markers. The complete taxonomy of biomarker identification is shown in Fig. 5.

3 Review Method

Following the methods of Kitchenham et al. [66], a thorough study of tools and techniques required for biomarker identification using multi-omics data analysis is conducted to summarize current work and highlight scientific limitations. The analysis process begins with several research problems to be answered, as outlined in Sect. 3.1. The fundamental goal of this study is to address the most recent approaches and tools used for identifying biomarkers by responding to the research problems. Various keywords have been used to search the articles in different libraries required for literature review. Lastly, the data collection process is simplified by using an inclusion–exclusion process.

3.1 Research Problems

This survey gives detailed information about the most recent tools and technologies required for biomarker identification by responding to the following research problems.

P1: What do you mean by biomarkers, and what are its various types?
P2: What do you mean by multi-omics data?
P3: What are feature extraction and selection techniques? What are its types, and why are they needed?
P4: What strategies have been developed for identifying biomarkers using omics and multi-omics data?
P5: What are the tools developed for biomarker identification using multi-omics data?
P6: In the field of biomarker identification, what are the current problems and opportunities?

3.2 Article Resources

The different online sites have been used to search the articles from various publications, including Springer, IEEE explore, Google Scholar, Elsevier, Web of Science, Science Direct, and Wiley Online Library. Various documents like research articles, conference papers, survey articles, editorial materials, and book chapters can be retrieved from the aforementioned resources.

3.3 Criterion Used for Searching

We start with the title "biomarker identification", "biomarker identification in multi-omics", "techniques used for biomarker identification in multi-omics data", "tools for biomarker identification in multi-omics data". Using these keywords, different string has been formed and are shown as below:

- "Biomarker identification" + "omics"
- "Biomarker identification" + "multi-omics"
- "Diagnostic biomarker identification" + "omics" + "machine learning"
- "Prognostic biomarker identification" + "omics" + "machine learning"
- "Predictive biomarker identification" + "omics" + "machine learning"
- "Diagnostic biomarker identification" + "multi-omics" + "machine learning"
- "Prognostic biomarker identification" + "multi-omics" + "machine learning"
- "Predictive biomarker identification" + "multi-omics" + "machine learning"
- "Biomarker identification" + "multi-omics" + "deep learning"
- "Biomarker identification" + "multi-omics" + " tools"

Scientific papers from numerous publications, journals, chapters, and conferences have been included in the search.

3.4 Inclusion–Exclusion Process

In the data inclusion and exclusion method, 98 studies were chosen for this study. The discovery process begins with a search string that returns unrelated papers to the report. Figure 8 depicts the beginning of the process, which starts with 1510 research papers returned. With the title-based exclusion, the count is decreased to 704. The number is reduced to 420 based on the related abstract. Since considering the entire text, only 180 articles remained. Ultimately, 98 research articles were chosen for the literature review.

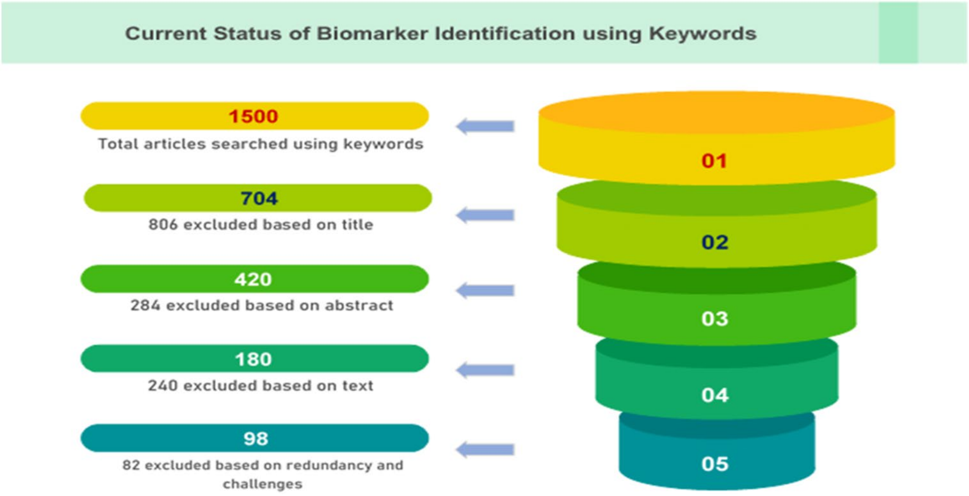

Current Status of Biomarker Identification using Keywords
1500
Total articles searched using keywords
704
806 excluded based on title
420
284 excluded based on abstract
180
240 excluded based on text
98
82 excluded based on redundancy and challenges
01
02
03
04
05

Fig. 8 Data inclusion and exclusion process

4 Biomarker Identification Literature Review

In this section, a thorough review of feature selection techniques required for identifying biomarkers using multi-omics data set is performed, followed by the identifying biomarkers using omics and multi-omics data with the help of ML and DL.

4.1 Feature Selection and Extraction Techniques for Biomarker Identification

Feature Selection and extraction work by reducing feature space's dimensionality and selecting the most relevant features. The complete description of feature extraction and selection is discussed in Sect. 2. Here we discussed some of the work done by the authors on feature selection techniques. Malik et al. [67] used maximal-relevance and minimal redundancy (mRMR) feature selection technique on gene expression and DNA methylation data to reduce their dimensionality. A sample of prostate cancer patients has been taken, then passed to preprocessing and mRMR method, which selects the most relevant genes as the top-ranked markers. Fujita et al. [68] used a feature selection method to discover biomarkers using a multi-omics dataset. The authors took the dataset in three matrix forms and applied the JNMF method, which generates four clusters by reducing the dimensionality of the matrices. This method also reduces noisy values and selects the relevant features, further passed for training purposes. This method successfully identifies the candidate genes and finds the association between the genes and the drugs given for treatment. Jia et al. [69] used PCA to reduce the dimension of enormous feature space from miRNA, mRNA, DNA, and single nucleotide polymorphisms (SNP) data of pancreatic cancer and identified 12 risk biomarkers. First, the dimensionality is reduced using PCA, and then the most relevant features are selected using Filter methods, including chi-square and t-test. The proposed method identified 12 markers successfully. These methods work well for linear multi-omics data integration only. For non-linear integration, non-linear methods are required, as explained in Sect. 2. Southekal et al. [70] used the t-SNE feature extraction method to identify markers from Gene expression and DNA methylation dataset of 35 cancer types. Dataset was taken and passed to preprocessing and t-SNE algorithm to reduce dimensionality. Further, a filter method identified a correlation between the features and selected top-ranked genes. The survival analysis was performed to check the risk scores for the selected genes based on selected genes. Similarly, Moon et al. [71] use deep learning auto-encoders with Kernel PCA to identify candidate biomarkers using integrative omics datasets. The authors used the DNA methylation and Gene Expression dataset and passed it to preprocessing stage to remove the noisy values. Further, the preprocessed data is passed to PCA and stacked auto-encoder to convert the data from multi-dimensional space to a single dimension. Then the differentially expressed genes are identified, performance is checked, and candidate markers are selected related to the disease. Hamzeh et al. [72] used to filter and wrapper methods to identify prostate cancer patients' biomarkers. At first, preprocessing of the dataset is performed, passed to filter methods including Information Gain and chi-square test to rank the genes. Further, the wrapper method mRMR method to select the most relevant biomarkers. Multiple algorithms of ML, including NB, RF, and KNN, have been used to check the performance. Further, feature extraction/selection and ML and DL algorithms are used in the next section to identify different types of biomarkers (Prognostic, Diagnostic, and Predictive Biomarkers) using omics and multi-omics datasets.

4.2 Biomarker Identification Using ML and DL for Omics Data

The biomarker identification is divided into three sections: identification of diagnostic, prognostic, and predictive biomarkers using omics data with the ML and DL approaches.

4.2.1 Diagnostic Biomarkers

Diagnostic markers are the markers used to confirm the presence of disease and identify the markers in different sub-types of cancer. For example, Zhao et al. [73] used the machine learning algorithm RF to identify novel diagnostic biomarkers in hepatocellular carcinoma. miRNA genes of 373 patients were downloaded from TCGA data and passed to the Random Forest model for biomarker identification. The experiment was validated on the GSE63046 dataset. The results found that the proposed method identified five diagnostic biomarkers. Kloten et al. [74] presented a technique to discover the new markers in breast cancer patients. The authors examined the promoter methylation of seven putative tumor suppressor genes (ITIH5, SFRP1, WIF1, SFRP2, RASSFIA, SFRP5, and DKK3). Our findings showed that ITIH5 and DKK3 promoters methylation as possible biomarkers achieve a precision of 93%. Rehman et al. [75] proposed machine learning algorithms to validate the importance of miRNA as breast cancer biomarkers. miRNA sample of breast cancer patients has been taken and passed to the preprocessing stage to eliminate all the missing values from the dataset. It is further given to feature selection techniques comprising LASSO, Chi-Squared (CHI2), and Information Gain (IG) to rank the features according to their importance. The training of these samples is performed using machine learning algorithms comprising RF and SVM, and it was investigated that 11 top-ranked miRNAs as biomarkers can be beneficial in predicting breast cancer and are treated as diagnostic biomarkers. Alkhateeb et al. [76] introduced machine learning algorithms to identify the transcripts for prediction and guide the treatment related to prostate cancer progression. Transcripts dataset have been taken and passed to preprocessing stage for cleaning the dataset, which is then passed on featuring extraction technique to obtain the differentially expressed genes. Machine learning algorithms, including SVM, RF, NB, and Decision Tree (DT), were used for modeling, and it has been evident that SVM outpaced with 90% accuracy. It was also found that HEATR5B, DDC, GABPB1-AS1, NREP, PTGFR, SCARNA22, FLVCR2, DOCK9, IK2F3, CLASP1, and USP13 are the potential biomarkers used for the progression of prostate cancer. Jin et al. [77] developed a model using a semi-restricted Boltzmann machine named ECMarker to predict biomarkers for the different stages of diseases like the early prediction of cancer. Gene-expression of non-small lung cancer patients was taken, and the ECMarker model was applied, which achieved an accuracy of 85%. The nine genes are identified as the diagnostic biomarkers, including KRAS, ALK, BRAF, PIK3CA, NRAS, AKTI, RET, EGFR, and ROS1. It was also used to prioritize biomarkers genes responsible for the early prediction of lung cancer. Tyanova et al. [78] discriminate between three breast cancer subtypes using a protein expression dataset to detect protein biomarkers using a machine learning model. Embedded feature selection techniques were followed by training and cross-validation using a SVM model. The experiment was performed, and it was identified that the detected biomarkers correctly classified the breast cancer subtypes with an AUC value of 91%. A total of eight markers (Her2, Grb7, MCM5, STMN1, GLS, RCL1, C9ORF114, and ENO1) are detected as diagnostic biomarkers. Xie et al. [43] used ML methods to identify diagnostic biomarkers using a metabolomics dataset. A sample of 110 patients was collected from the Hubei Taihe Hospital and passed to PCA to select the metabolites. Then the Statistical analysis is performed considering only those metabolites having a value less than 0.05. Further, the selected metabolites were passed to ML comprising RF, KNN, SVM, NN, NB, and Adaboost, and it was found that NN performed best with accuracy, specificity, sensitivity, and AUC value of 99%. It was also found that ten metabolomics biomarkers including L-Kynurenine, Proline, Spermidine, Palmitoyl-l-carnitine, Amino-hippuric acid, Phenylalanine, Taurine, L-Valine, *o*-Tyr, Carnitine plays a critical function in tumor diagnosis. Muazzam [79] used deep neural networks (DNN) to identify diagnostic biomarkers using the RNA-Seq dataset of breast cancer patients. First, the size of the dataset is reduced and then passed to the Stacked Denoising encoder for biomarker identification. Further, the identified biomarkers are passed to DNN to classify cancer patients. Pathway analysis identifies three genes (PIK3C2G, PCDHB8, WNT10A) in multiple cancers. Khattri et al. [44] proposed an ML algorithm for identifying diagnostic markers using transcriptomic data from Pancreatic Adenocarcinoma (PDAC) patients. Dataset was collected from Array Express and GEO and passed to preprocessing and t-test to discover genes. The identified genes were then passed to SVM to classify cancer patients. The experiment was performed, and the tested result proved that the proposed framework successfully identified nine genes comprising IFI27, CTSD, ITGB5, EFNA4, PLBD1, GGH, HTATIP2, CTSA, and IL1R2 with 97% accuracy. Liu et al. [45] presented two multilayers feed-forward NN using DL to identify markers using a DNA methylation dataset. The t-statistics test was passed to LASSO and the RF algorithm to identify genes. This test identifies 12 CpG markers and 13 promotor markers. Further, these markers are passed to the Deep learning model, which achieved a sensitivity of 92% for CpG markers and 89% for promotor markers. Toth et al. [80] presented a random forest-based classification model to detect biomarkers for prostate cancer. DNA methylation dataset was downloaded from TCGA and passed to the preprocessing and feature extraction stage to extract the relevant features. It was then given to the random forest model to identify the biomarker for prostate cancer. The results are evaluated, and it was proved that the random forest-based modeling identified the top 30 methylation genes and performed best with an AUC value of 77%.

4.2.2 Prognostic Biomarkers

Prognostic markers are used to predict the occurrence of a potential clinical condition, disease recurrence, or relapse in an identified sample. The work done by the various authors in prognostic markers is described as follows. Ma et al. [81] proposed machine learning algorithms to identify 16 gene prognosis markers to predict lung adenocarcinoma (LUAD). Clinical and RNA-seq dataset from the TCGA portal was used for the experiment. At first, survival-related genes were identified using Cox and random survival forest (RSF) method, and then prognostic-related genes were identified from integrated clinical and RNA-seq data. Furthermore, to validate the results, GEO was used. The experiment was performed and compared with existing prediction models. The result was calculated using three metrics comprising hazard ratio (HR), concordance index (CI), and p-value, and it is evident from the results that the proposed method outperforms with the c-index value of 67%. It was also found that 13 new biomarkers, including PITX3, LINC00908, GJB3, MELTF, CRCT1, LOC105370802, BAIAP2L2, GABRA2, RHOV, ARF3, KRT18, TRIM7, ZNF710.AS1 and LOC100996732 were identified as compared to existing studies. Hossain et al. [82] used machine learning algorithms to discover biomarkers related to ovarian cancer. Clinical and Gene-expression information of OC patients has been integrated and passed to ML models, which discover the biomarkers affecting Ovarian Cancer patients. Cai et al. [83] used a machine learning algorithm random survival forest to identify prognostic biomarkers of HCC patients. Gene signatures from two HCC datasets were used and passed to a random survival forest to predict the gene signatures. Then a Protein–Protein Interaction (PPI) network is generated, and it identifies RAD21, CDK1, and HDAC2 markers. Further to check the validity of identified markers, a multi-variate Cox survival analysis was performed, showing that CDK1 is the only prognostic marker for HCC patients. Ghosal et al. [84] use ML algorithms to identify the prognostic markers in noncoding RNAs dataset. First, data is passed to a statistical test to identify the differentially expressed genes, then passed to the multi-variate cox regression model. Further, four ML algorithms comprising LASSO, elastic net, cart, and ridge were used to classify the sample into five cancer subtypes. Then Kaplan–Meier analysis identified five lincRNAs (LINC00472, RP4-806 M20.3, RP1-40E16.9, RP11-254F7.2 RP11-455B3.1) as prognostic markers. Li et al. [85] used weighted gene-coexpression network analysis (WGCNA), LASSO, and multi-variated cox analysis for the identification of prognostic markers using the RNA-Seq dataset of cervical cancer (CC) patients. The proposed algorithms identified two biomarkers, including ACAP1 and RASGRP1, by calculating their risk score using the LASSO cox algorithm of machine learning. Liu et al. [86] aimed to identify prognostic genes of Osteosarcoma using machine learning. RNA-Seq samples of 94 Osteosarcoma were collected and passed to ML to identify prognostic markers. The experiment was performed, and the results evidenced that the proposed framework identifies four markers (RPL7AP28, RPL11-551L14.1, RP11-326A19.5, and RP4-706A16.3). Yu et al. [50] identify miRNA prognostic markers using ML algorithms. A sample of 119 patients was collected from the TCGA database, where data from TCGA is used as a validation set. At first, differentially expressed miRNAs were calculated using a p-value. The optimal feature subset is selected using Recursive Feature Elimination (RFE). The selected optimal features are passed to the SVM model, classifying the patients into early-stage and last-stage samples. Then the risk is calculated using the survival analysis method univariate cox regression model, which identifies five prognostic markers. Xing et al. [87] used survival-related cox regression analysis models to identify prognostic small nucleolar RNAs (snoRNAs). The data is passed to LASSO regression, and snoRNAs having a p-value less than 0.5 are selected as candidate biomarkers. Then these markers are passed to Cox proportional hazard model (multi-variate) to identify the prognostic markers. Further, correlation analysis was performed to validate the identified markers. The presented approach identified five snoRNAs as prognostic markers. Long et al. [88] proposed a deep learning method to identify prognostic markers in cervical cancer patients. Dataset was downloaded from GEO and Array and passed to statistical analysis test to identify differentially expressed genes. These genes are passed to a deep learning model for cancer classification. Further, survival analysis is also calculated using cox models, and it identified two genes (ZNF281, EPHB6) as the prognostic markers for cervical cancer. Using a deep multilayer perceptron network algorithm, Wong et al. [89] identify prognostic genes for glioblastoma (GBM). Gene expression data was collected and passed to partial likelihood to calculate the loss function required for survival analysis. The features extracted were then passed univariate and multi-variate regression models, identifying the ten prognostic makers comprising TNR, POSTN, BCAN, TMSB15B, GAD1, SCG3, NNMT, PLA2G2A, ELAVL4, and CHI3L1 by calculating the concordance index value.

4.2.3 Predictive Biomarkers

A predictive biomarker is a test that can classify people who are more likely to react to a specific medicinal substance or chemical product. A symptomatic gain may increase longevity, or an adverse effect may be the result. In this research, a predictive biomarker is considered as a gene prioritization problem where the gene can signify the occurrence of some particle disease with some known disease genes. Therefore, the work done on gene prioritization algorithms is discussed in this section. Nam et al. [90] designed a Gene Ranker method to identify genes using a gene expression dataset. First, a PPI network was created and used as a base network. Then, the network is generated using WGCNA. An integrated network is generated and passed to the gene ranker algorithm to generate a score. The higher rank genes, including OTC, B3GNT9, and Clorf167 are identified as the predictive markers. The ten known genes, including CSNK2A3, IFNL2, UCN3, POU3F4, TIW1, IL22, UCN2, PSG1, HTRA1, and CD68 are also identified. These genes have a strong relationship with the above-mentioned identified genes. Zhao et al. [91] presented a graph convolutional network (GCN) to prioritize protein coding genes using the lncRNAs dataset. The lncRNA dataset was used and passed to the feature selection technique in which the gene expression and position of the gene is identified, and the gene network is created. This network is then passed to a graph convolutional network, which prioritizes the target genes of lncRNAs. The method is also validated and compared with existing methods, and it is found that GCN works well with an AUC and AUPR value of 90% and 91%, respectively. Zhang et al. [92] proposed a network-based Deep-learning

Approach for identifying genes by prioritizing them. First, a human molecular interaction network (HMIN) was constructed in which nodes represent the proteins corresponding to their gene and edges represent the interaction between the genes. This network is then passed to a graph convolutional network (GNN), which trains the dataset and prioritizes the genes based on their influence on the patients. The experiment was performed, and it is evident from the results that the proposed framework works well by selecting ten genes, including RUNX1T1, MAG12, GRIA3, MVCRP2, AKAP6, PTPRD, AUTS2, MYO9A, AB12, and PLXNA2 respectively. Jiang et al. [93] presented a generative adversarial network (GAN) with de-noising auto-encoder (DAE) as the generator and multilayer perceptron Multilayer Perceptron (MLP) as a discriminator (GAN-DAEMLP) to prioritize genes by taking miRNA dataset. The dataset sample was taken and passed to GAN-DAEMLP, which calculates the disease and non-disease prediction score. Finally, a risk score is calculated, and a genes risk list is generated. The experiment was performed, and it is proved from the results that the GAN-DAEMLP performed best by selecting ten disease-related genes. Table 3 shows the work done by various authors using ML and DL for omics data.

4.3 Biomarker Identification Using Machine and Deep Learning for Multi-omics Data

As single omics is not enough for the correct discovery of biomarkers, multi-omics analysis is required. This section performs biomarker identification from multi-omics data using ML and DL for each diagnostic, prognostic and predictive marker.

4.3.1 Diagnostic Biomarkers

Sinkalaet al. [94] proposed machine learning algorithms that accurately identify the set of proteins, mRNAs, miRNAs, and DNA methylation biomarkers to classify the pan-cancer into its subtypes. Sample of pan-cancer patients has been obtained from TCGA and cBioPortal, which were then passed on featuring an extraction technique called neighborhood component analysis (NCA) which identifies marker sets involving 49 methylated genes, 50 mRNAs, 20 miRNAs, and 14 proteins. After that, KNN and SVM models are applied, which effectively classify the cancer subtypes with 99% and 97% accuracy, respectively. Hamzeh et al. [95] used ML to calculate the Gleason score for prostate cancer and to identify the potential biomarker for each Gleason group accurately. mRNA and miRNA genes were taken from the NCBI GEO repository and passed to hybrid feature selection techniques to extract features. The experiment was performed, and it has been found that the proposed framework works well with 93% accuracy. PIAS3 and UBE2V2 were also identified, which will strongly correlate with prostate cancer progression. Xu et al. [96] identified biomarkers related to cervical cancer using a hybrid feature selection ML method by integrating multi-omics data. DNA methylation profiles of 12 types of cancer have been taken, and adopted machine learning techniques were applied. The results were evaluated, and it has been found that four cancer-specific markers comprising cg12205729 (GABRA2), cg07211381 (RAB3C), cg26490054 (SLC5A8), and cg20708961 (ZNF257) could identify the tumor cells with 92% AUC value. Guo et al. [97] presented a deep learning framework using a denoising autoencoder to identify subtypes of ovarian cancer and to identify genes related to ovarian cancer. The multi-omics dataset was collected using TCGA Assembler and integrated using denoising autoencoder. Further, the dataset was passed to the k-mean clustering technique to select the relevant features. These features were then given to the L1-penalized logistic regression (LR) to recognize the subtypes. Long et al. [98] used statistical learning and ML algorithms to identify biomarkers in pancreatic cancer patients. Transcriptomic, Genomic, and protein datasets were taken and passed a statistical test. Further, the identified survival analysis is performed using the Cox survival model. The identified biomarkers were also passed to the random forest model to classify cancer into normal and tumor patients. The proposed framework also shows that the protein expression of identified genes is highly correlated in pancreatic cancer patients. As diagnostic and prognostic markers, the proposed framework identified four genes, including LAMC2, ANXA2, ADAM9, and APLP2. Long et al. [99] used Different random forest features selection approaches comprising Boruta, AUC-RF, and Vita to identify diagnostic biomarkers. Transcriptomic data, including mRNA and miRNA sequencing data, was passed on, featuring extraction techniques identifying eight diagnostic biomarkers. Further, to check the performance of identified signatures, ML algorithms including RF, LR, NB, and KNN were used, showing that the identified markers are highly correlated in colorectal cancer. Feng et al. [100] proposed joint kernel learning on the multi-omics dataset to identify diagnostic genes from lung and liver cancer. The isoform expression profile, DNA methylation, and gene expression sample were collected and passed to the KPCA method for feature extraction. Then the extracted features were converted to kernel metrics using the Gaussian Kernel function, which is then passed to the clustering algorithm. The clustering algorithms divide the features into clusters for different cancers. The proposed framework identifies GMPS, EPHA10, C10orf54, and MAGEA6 for lung cancer and FAU, DEPDC6, VPS24, LOC100133469, RCBTB2, and SLC35B4 for liver cancer. Kwon et al. [101] proposed an ML algorithm to discover diagnostic signatures in pancreatic cancer patients. miRNA and mRNA expression

Table 3 Biomarker identification using ML and DL for omics data

Author [Ref]	Type	ML/DL	Algorithm	Data source	Result	Future studies
Zhao et al. [73]	Diagnostic biomarkers	ML	RF	miRNA dataset from TCGA and GSE63046	RF identified 5 genes as diagnostic markers with AUC value of 89%	Large sample size can be used
Kloten et al. [74]		ML	Fisher Extract Tests	Breast cancer serum samples from TCGA	Proposed tests identified 7 driver genes with 93% precision	Comfortable and non-invasive blood-borne screening procedure can be used
Rehman et al. [75]		ML	RF, SVM, CHI2, IG, LASSO	Breast cancer patients miRNA data from TCGA	RF outperforms and identified 11driver genes as the diagnostic markers with 99%, 96% and 99% accuracy, specificity and sensitivity	The proposed machine learning algorithms can be used for the identification of miRNA's biomarkers for other diseases
Alkhateeb et al. [76]		ML	mRMR, SVM, RF, DT, and NB	RNA seq dataset from NCBI	SVM with RBF kernel outperforms and identified 11 diagnostic biomarkers with 80% accuracy	Wet-lab experiments and clinical assays would be necessary to confirm the existence of these biomarkers
Jin et al. [77]		ML	ECMarker	Lung Cancer gene expression dataset from TCGA	The proposed ECMarker performed well and identified 9 markers with 85% accuracy	Multi-omics analysis can be carried out to predict biomarkers for disease prediction and guiding the treatment
Tyanova et al. [78]		ML	SVM, ANNOVA rank method	Protein data of breast cancer from TCGA	SVM model identified 8 diagnostic markers with an AUC value of 91%	It can be used in translational cancer research in which treatment therapies can be provided using the detected genes
Xie et al. [43]		ML	PCA,SVM, NN, RF, Adaboost, KNN, NB	Metabolomics data of lung cancer patients collected from Hubei Taihe Hospital	NN outperforms by identifying 10 with sensitivity, specificity, accuracy and AUC value of 99% respectively	This research focused only on metabolites level. Further information like age, history of smoking, past medical history can be included for better performance
Muazzam [79]		DL	DNN, KPCA, PCA	RNA-seq dataset of breast cancer patients from TCGA portal	DNN with stacked auto-encoder performs well and identified 3 marker with 95% accuracy	This study can beapplied to dataset of larger size and on multi-omics dataset for better performance
Khatri et al. [44]		ML	SVM, Statistical analysis	Transcriptomic data of PDAC patients from ARRAY and GEO	The proposed ML methods along with statistical analysis identified 9 markers with 97% accuracy	The proposed framework performed well in blood biomarkers and will be an ideal for future studies
Liu et al. [45]		DL	NN, RF, LASSO, t-statistics test	DNA methylation datasets from TCGA and GEO	12 CpG markers and 13 promotor markers are identified with 92% and 89% sensitivity	Future studies involves more statistical, ML and DL algorithms for the identification of biomarkers

Table 3 (continued)

Author [Ref]	Type	ML/DL	Algorithm	Data source	Result	Future studies
Toth et [80]		ML	RF	DNA methylation dataset of prostate cancer patients download from TCGA	RF performed best and identified top 30 markers with an AUC value of 77%	Additional biomarkers will be discovered using whole-genome bisulfite sequencing (WGBS)
Ma et al. [81]	Prognostic Biomarkers	ML	Cox, RSF	RNA-seq data from TGCA, and GEO	13 prognostic markers were identified with c-index value of 67%	DL can be used for biomarker discovery and cancer prognosis
Hossain et al. [82]		ML	Univariate, Multivariate Cox model	Gene-expression dataset of Ovarian Cancer from OMIM	Proposed model identified 5 genes with an hazard ratio close to 1 for each gene	The search can be applied in the identification of biomarkers for different types of cancers
Cai et al. [83]		ML	RSF, Multi-variate cox regression	Gene-expression data from GEO and Array express databases	Cox survival analysis identified only one marker as prognostic marker in HCC patients	Identification of gene signatures can be done that discuss recurrence risk related to disease
Ghosal et al. [84]		ML	Ridge, Elastic net	Non-coding RNAs dataset from TCGA portal	5 lincRNAs were identified as prognostic markers	In future, these studies can be extended for further identification of lincRNAs
Li et al. [85]		ML	WCGNA, LASSO	RNA sequencing dataset of CC patients from TCGA	Two mRNAs were identified with hazard ratio of 0.13	–
Liu et al. [86]		ML	Univariate, Multivariate, Lasso cox	RNA sequencing data of osteosarcoma patients from TCGA	Four markers were identified with hazard ration of 0.65, 0.32, 1.89, 0.52 respectively	Future Studies involves the validation of experiment using different datasets
Yu et al. [50]		ML	RFE, SVM, univariate cox	Two datasets of ESCC patients from GEO (GSE43732) and TCGA	Five microRNA markers were identified as prognostic markers	DL can be employed in future for the identification of prognostic biomarkers using large size dataset
Xing et al. [87]		ML	LASSO, cox	Dataset from UCSC Xena database	Proposed approach identified 5 markers as prognostic markers	–
Long et al. [88]		DL	DL, COX models	Cervical cancer dataset from GEO and ARRAY	Two markers were identified with 97% accuracy	In future, cancer types can be identified using the proposed method
Wong et al. [89]		DL	Deep Multilayer perceptron	Gene expression data of lung cancer patients retrieved from TCGA	Deep learning outperforms several ML algorithms and identified 10 markers as the prognostic markers	–
Nam et al. [90]	Predictive biomarkers	ML	WGCNA network	Gene Expression data from GEO Database	Proposed method successfully identify 3 genes as predictive markers with an AUC value of 76%	Disease categories can be identified and different methods can be created to integrate the networks
Zhao et al. [91]		DL	Graph CNN	lncRNAs dataset downloaded from TCGA	The presented method prioritize the candidate genes with 91% AUC	–

Table 3 (continued)

Author [Ref]	Type	ML/DL	Algorithm	Data source	Result	Future studies
Zhang et al. [92]		DL	PANDA, Graph CNN	Protein Database of Autism patients from Simons Foundation OMIM	PANDA successfully identified 10 genes and outperforms several ML algorithms with an accuracy of 89%	This framework help researchers to learn and perform better by understanding the complex genetic architecture of diseases
Jiang et al. [93]		DL	GAN-DAEMLP	miRNA expression dataset from http://www.hdinhd/	GAN-DAEMLP identified 9 markers with AUC value 0f 90%	–

data samples are integrated and passed to the SVM model to identify biomarkers. Further, validation is performed to select more relevant features using two independent datasets, including GEO and TCGA data repositories. This method identifies six biomarkers and 705 multi-markers related to the identified markers. Joshi et al. [102] proposed a DNN named Sparse CRossmodel Superlayered Neural Network (SCR-SNN) to integrate mRNA and DNA Methylation data and the biomarker identification for lung cancer patients. The dataset was passed to PCA for data filtering. Further Biomarker selection is performed using SCR-SNN, including LR with L1 penalty, L1-regulalized NN, and L1-regularized cross modular NN. The proposed method identifies 15 markers, including WFDC5, TATDN1, LPP, CPLX2, CXCL13, COLI17A1, CEL, CDSN, TMPRSS2, FOXD1, DSC1, LPIN2, MMS4A8, B3GALT2, and AQP10 as the diagnostic markers for lung cancer patients. The proposed method is also compared with exiting ML algorithms employed on a single omics dataset. Cheng et al. [103] use LR to identify diagnostic biomarkers using an integrated omics dataset of HCC patients. Correlation between gene expression and DNA profiles is performed using PCC. The proposed method identified six profiles including DSE (cg11481534), FAM55C (cg03509671, cg21908638, and cg11223367), NEBL (cg23565942), and GALNT3 (cg05569109) as the diagnostic markers with a sensitivity value of 92%. Zhang et al. [104] used cox survival analysis and the BayesNet model to identify diagnostic biomarkers from breast cancer patients using a DNA methylation and gene expression dataset. Dataset was collected and passed to statistical tests to identify potential biomarkers. These markers are then passed to the BayesNet model to classify the patients from healthy candidates. Further, the candidate markers are passed to the Cox regression model to identify the survival value, which identified seven differentially methylated sites (DMSs) comprising TUFT1, TRERF1, CCND1, SRGAP1, PER1, ENPP2, and PER1 as diagnostic and prognostic markers. Zhang et al. [105] used the RF feature selection method to identify the diagnostic biomarkers using osteoporosis patients' lncRNAs, mRNAs, and miRNAs dataset. A network was created of 105 nodes passed to functional analysis, which shows the involvement of DysCeNet in osteoporosis. Further, RF was used, which identified 25 features as diagnostic biomarkers. The identified genes are also validated using the LOOCV method, which shows that the identified genes show good performance in cancer classification. Liu et al. [52] identifies diagnostic and prognostic markers using machine learning algorithms, including RF and LASSO-Cox, from epigenetic, transcriptomic, and metabolomics dataset. 9398 CPGs and 2478 genes were collected and passed to Random forest, which selected 134 CpGs and 54 genes from the integrated dataset. These are then passed to LASSO for the identification of diagnostic markers. Moreover, prognosis analysis is performed using univariate Cox and LASSO cox methods. The proposed framework identifies five diagnostic and eight prognostic markers, respectively.

4.3.2 Prognostic Biomarkers

Zhao et al. [106] suggested a technique to the rank gene by calculating a score to identify the biomarkers. An integrated dataset of 13 cancer types was taken, and gene ranking was performed using the Cox (multivariate) proportional method. The experiment was performed, and higher ranks genes were identified as the prognostic biomarkers. Further C-index was used to validate the results, and it was found that in comparison to single omics, multi-omics works well with a c-index value of 0.95. When contrasting the genes related to 13 types, seven genes (API5, SLK, BTBD2, VPS37A, PTAR1, ZRANB1, and EIF2B1) be linked with several cancer prognoses. Kumar et al. [107] presented survival models to identify biomarkers in pancreatic ductal adenocarcinoma (PDAC). A sample of 153 PDAC patients was taken from the TCGA database for multi-omics analysis, consisting of DNA methylation, lncRNA, gene expression, and miRNA data. Then preprocessing and feature extraction was performed to identify the genes positively correlated with survival. For survival analysis, Cox and Kaplan–Meier estimations were done. The experiment was performed, and the results proved that the presented work performed well, with an AUC value of 95%. It was also identified that five genes comprising (B3GNT3, DMBT1, PVT1, DEPDC1B, and Gata6-AS) are strongly connected with the survival of PDAC patients. Zhuang et al. [108] used ML to identify gene biomarkers or the prognosis of acute myelocytic leukemia (AML). Copy Number Variation (CNV), RNASeq, and SNP data were used, and ten feature genes comprising CAMK2A, FAT2, TCERG1, PTGIS, GDF9, DOC2B, PREX1, DNTTIP1, C22orf42, and CRISPLD1 were identified which were then passed to univariate cox regression analysis to develop a signature gene which is responsible for the prediction of AML. The results are validated using GEO datasets. Dong et al. [109] developed prognostic models for the early prediction of LUAD using trans-omics biomarkers. The authors integrated the clinical, DNA methylation, and gene-expression dataset of 825 patients and used the Ranger algorithm for screening the biomarkers associated with prognosis. The experiment was performed, and it is evident from the results that the developed method improved the performance by 18.3% with an 87.2% AUC value. The concordance-index value shows an improvement of 4% compared to various existing models. Ouyang et al. [110] proposed an integration method comprising fisher ratio, classified information index, Spearman correlation coefficient (SCC), and decision trees (DTs) ensemble for the discovery of biomarkers in hepatocellular liver carcinoma (LIHC) using unbalanced datasets. The multi-omics datasets consisting of miRNA expression data, somatic mutation data, and DNA-methylation data were utilized from TCGA, and 34 Differentially expressed genes (DEGs) were identified. These identified genes are used to discriminate tumor cells from normal cells in LHC patients with an AUC value of 89%. Peng et al. [111] proposed a DL framework called Capsule Network-based Modelling of Multi-omics data (CapsNetMMD) to detect signatures related to breast cancer. Sample of 770 breast cancer patients, including DNA Methylation, miRNA expression, and CNA, have been taken and converted into a matrix form. It was then passed to CapsNetMMD for the extraction of genes. The experiment was performed, and the results were evaluated. The results were also compared with different ML algorithms comprising XGBOOST, NN, SVM, Adaboost, and KNN, and it was marked that CapsNetMMD outperforms with 90% Accuracy. Lim et al. [53] used a deep learning framework called Artificial Neural Network (ANN) to analyze genetic data and discover disease-related genes. TCGA dataset of breast cancer patients was taken, and the experiment was performed. For parameter optimization, the lasso penalty activation function was used. The model was compared using the Youden J index with other ML algorithms, including meta LR and meta-SVM. It is estimated from the results that the suggested DL framework is more robust in the discovery of genes. Lai et al. [112] proposed DNN to identify novel biomarkers from non-small cell lung cancer patients (NSCLC). A sample of 614 patients with gene expression and clinical data was integrated with 15 biomarkers to develop an integrative DNN model. The biomarkers are discovered using the StepMiner algorithm. The experiment was performed, and it was found that the proposed framework works well by accurately identifying the markers with 70% accuracy. Cui et al. [113] proposed u-net to identify prognostic biomarkers. A sample of 191 patients has been taken from the TCGA portal, and a u-net is applied to segment the images. The Cox-proportional hazard model has been used to predict survival. Four biomarkers comprising × 70, × 93, × 107, and × 164 were discovered, guiding lung cancer patients' survival. Mo et al. [114] used RF to identify prognostic biomarkers from breast cancer patients' integrated omics (SNP, RNASeq, and CNV) dataset. The integrated genes are then passed to the Random forest for feature selection. This technique identified 120 candidate genes. These genes were then passed to the Cox regression model to identify prognostic genes. The experiment was performed, and it is evident from the experiment that the proposed algorithm successfully identifies six genes, including CD24, PRRG1, IQSEC3, MRGPRX, RCC2, and CASP8 prognostic markers. Mo et al. [115] presented a clustering approach to identify the prognostic value of bladder cancer patients from the multi-omics dataset. A sample of 388 patients, including Somatic mutation, DNA methylation, RNASeq, and CNA, was passed to the iClusterBayes method. This will divide the data into two clusters that are basal and luminal subtypes clusters. These clusters are validated using Markov Chain Monte Carlo (MCMC) method. Only those genes are considered whose posterior probability is greater than 0.5. A total of 42 genes are identified, which are further passed to statistical analysis tests, including Fisher's exact test, two-sample t-test, and Analysis of Variance (ANOVA) methods, which identifies seven genes, including KRT5, CD44 KRT6B/C, TGM1, KRT14, PI3, and DSC3 as prognostic markers. Zhang et al. [54] identified immune-related prognostic signatures using a multi-omics dataset of lung cancer patients. A sample of 553 RNA seq and 504 DNA methylation data was collected and passed to the ESTIMATE algorithm to create the Tumor Microenvironment (TME). A score is calculated on which the patients with higher tumor priority are selected. Then, multi-omics analysis extracted the relevant genes and passed them to the cox and lasso regression model for further analysis. The c-index value was calculated, and based on it, six expression genes (FOXN4, PROZ, LCN15, CD70, UNC5D, and BIRC3), five methylation genes (cg04240491, cg08780166, cg01090026, cg26904049, and cg25407540) and two mutation genes (PTPRT and COL22A1) are identified as prognostic markers. Xu et al. [116] used ML algorithms to identify prognostic markers of pancreatic adenocarcinoma patients. RNA sequencing, SNP, and CNV datasets were passed to GISTIC 2.0 and Mutsig 2.0 to preprocess the omics data. Fifty-four candidate genes are identified and then integrated and passed to the LASSO risk prediction model, selecting nine markers comprising TSPYL4, UNC13B, KLHDC7B, MICAL1, AIM1 KLHL32, DCBLD1, ARHGAP18, and CACNA2D4 as the prognostic markers. Chang et al. [117] proposed a pipeline to discover the markers in colorectal cancer. The somatic copy number, RNA expression, and gene expression data are used and passed to Wilcoxon rank-sum test to identify DEGs. the genes with a value greater than 0.3 are selected and passed to cox regression analysis. Finally, PCC was calculated, which identifies six-driver genes, including WDR5B, NDUFB4, IQCB1, GTF2E1, SEC22A., and KPNA1) which show poor prognosis related to cancer. Yuan et al. [118] developed clustering algorithms on multi-omics data to identify the prognostic biomarkers in brain tumor. A sample of 117 glioblastoma patients, including mRNA expression, DNA copy, SNP, DNA methylation, and clinical information, was used to experiment. MutSigCV was used to analyze SNP data, which decreased the number of false positives. For CNV data, GISTIC was used to extract the important CNV genes. Then, the genes are integrated and passed to a cluster of cluster analysis (CoCA) algorithm that divides the data into HX-1 and HX-2. The survival analysis of these clusters is performed, which identifies three methylations including DUSP1, PHOX2B, HOXA7 cg169573, and 15 gene mutations including CYP27B1, PCDH1, LPIN3, BCL6, GPR32, OR4Q3, SKIV2L, MAGI3, PCSK5, UBE3B, AKAP12, MAP4, F5, TP53BP1, and RHOBTB1 as the prognostic markers.

4.3.3 Predictive Biomarkers

Dimitrakopolos et al. [119] developed a Network-based Integration of Multi-omics data (NetICS) method to prioritize cancer-related genes by integrating genetic aberrations, mRNA and miRNA, and DNA expression datasets. A bidirectional network diffusion is created, which generates a rank list for each sample. This rank list is then passed to rank aggregation techniques, generating a global ranking. NetICS identified the top 5% genes from breast cancer (TP53, PTEN, ERBB2, and CDH1) and Lung Cancer (EGFR, AKT1, KRAS, PIK3CA, and NRAS), respectively. Shang et al. [120] developed an integrative rank method to identify predictive markers in integrative omics data of HCC patients. A multiplex network is generated using multi-omics data by calculating the differentially mutual information (DMI). This DMI is then passed to the PageRank algorithm, and the final rank is obtained by aggregating the rank of multiple networks with an accuracy of 81%. Guan et al. [121] designed feature selection methods and support machines to prioritize the predictive genes multi-omics data. The PCC of genes was calculated, and their correlation scores were combined to generate a rank. The ten most predictive features are used, including ASAP2, BCL9L, PTPRF, PTPN12, ANXA1, AJUBA, CYTIP, SH3D19, CMTM4 EIF2C2, were selected. Yao et al. [122] proposed a method, MetPriCNet, to prioritize and predict the metabolites using a multi-omics dataset. The authors constructed a composite network of genomic, phenome, metabolome, and interactome datasets. This network consists of 25,269 nodes and 11,926,113 edges. This network is then passed to MetPriCNet, which calculates their global distance similarity. This method is applied to breast cancer patients, and it is found that the higher rank metabolite in 3 genes, including BARD1, TP53, and AKT1, interact with four seed genes consisting of CDH1, KRAS, CHEK2, CDS1. Fortino et al. [123] proposed fuzzy logic as feature selection, and Random Forest for prioritizing the genes using multi-class Four gene-expression dataset was taken and passed to fuzzy pattern discovery method to select the most relevant and class-specific features (FP). Then the selected feature set (FP) is passed to the random forest, which removes the redundant features and ranks the genes using a Mean decrease accuracy score. The proposed method works well, with an accuracy of 96%. Fan et al. [124] integrate multi-omics data, including genome, epigenome, and transcriptome data, to identify and prioritize the functional Differentially methylated regions (fDMRs). Authors first filter the DMRs, and based on the expression alteration scores, ranks are generated and further aggregated to identify and prioritize the genes. This method identifies ten genes as predictive markers using ranks. Further, classification and survival analysis of identified genes is performed. Chen et al. [125] suggested a BRIDGE method for candidate genes prioritization by integrating gene sequence similarities, protein–protein interaction, gene ontology annotations, gene-expression patterns, and gene pathway memberships. The authors used a regression model with the LASSO penalty to assign a weight to different genes. The test is validated in two case studies, including obesity, and diabetes, from which it is found that eight genes of obesity and 28 genes of diabetes patients lie in the top 100 rank list. Zhang et al. [126] develop a network-based approach to identifying and prioritizing predictive genes by integrating mutation, gene expression, and the PPI dataset. This approach works by identifying the neighbor genes. A relationship between the various differentially co-expressed genes (DCGs) and functional genes is made, and then the weight is calculated to check the impact of DSCs on the functional genes. This procedure is applied to three datasets, including kidney renal clear cell carcinoma (KIRC), thyroid carcinoma (THCA), and head and neck carcinoma (HNSC), to identify the genes. The experiment was performed, and it was found that the proposed method identifies the top five genes, including EGFR, EP300, NRAS, LYN, PTPN11, TP53, PIK3CA, EGFR, EP300, FADD, PBRM1, SETD2, BAP1, SRC and EP300 for THCA, HNSC, and KIRC respectively. Valdeolivas et al. [127] proposed a random walk with a restart method to prioritize the genes on multiplex (RWR-M) and multiplex heterogeneous networks (RWR-MH). First, a graph of the PPI network, pathway interaction, and co-expressed genes is created. The integrated network consists of 17,559 nodes and 1,659,084 edges which are then passed to RWR-M and RWR-MH to explore the different functionalities and associations of the graph. This is applied to Wiedemann Rautenstrauch syndrome patients, identifying three genes (Fig. 4, RNF113A and LMNA) strongly related to the disease. Wei et al. [128] proposed a method for Driver gene discovery with an improved random walk method (Driver_IRW) using transcriptomic and interaction network data integration. A network was created, and then the edge, betweenness, and Katz centralities were found using the constructed network. These scores are integrated and passed to a random walk with an improved method to calculate their rank. Finally, top-ranked genes are selected as the predictive markers. Zeng et al. [129] proposed a tree-based ensemble model called random interaction forest (RIF) to prioritize candidates and generate predictive scores. First, a decision tree is created, and the rank is calculated. The authors identify the top 10 genes and compare the results with other existing methods. Yang et al. [130] proposed a machine learning framework called MapGene to prioritize the candidate genes using high functional modules and gene interactions dataset. First, a PPI network is made of both disease and network interactions, and then module correlation (MC) is calculated using the MapGene algorithm and identifies the top rank genes as predictive markers. The proposed framework is also compared with several base models, and it is found that MapGene outperforms with a precision and recall value of 87% and 90%, respectively. Table 4 shows the work done by various authors on biomarker identification using ML and DL from multi-omics.

5 Biomarker Identification Using Tools

In this, the work done by various authors in biomarker identification using multi-omics data with the help of tools is described. All the tree biomarkers, including prognostic, diagnostic, and predictive markers, are considered here. Singh et al. [131] presented a framework for Data Integration Analysis for Biomarker discovery using Latent components (DIABLO) using a multi-omics dataset. This tool can identify the biomarkers from both simulated and real integrated omics data. mixOmics is used to implement the tool. Kaur et al. [132] developed a web server called HCCpred to identify diagnostic biomarkers and prognostic biomarkers from gene-expression datasets in Hepatocellular Carcinoma (HCC) patients. Raw data were extracted from 30 studies and passed to feature extraction techniques. The extracted genes were then passed to model training which successfully identified three genes (FCN3, CLEC1B, and PRC1). Kaur et al. [133] developed a tool called CancerLSP to identify biomarkers in Liver Hepatocellular Carcinoma (LCC). Genomic and epigenomic data, that is, transcripts and Cpg methylation data, were downloaded from the TCGA portal and passed to machine learning models (SVM, RF, NB, SMO, and J48). These algorithms are implemented in Weka, which successfully identified 21 Cpg sites and 20 transcript profiles related to LCC. Gevaert et al. [134] presented an Imaging-AMARETTO software tool for the identification of biomarkers from multi-omics, clinical, and imaging data fusion. Multi-omics data were downloaded from TCGA, and imaging data were used from Ivy Glioblastoma Atlas Project (IvyGAP). The tool was implemented on glioblastoma multiform (GBM) patients, successfully identifying three key drivers, including STAT3, AHR, and CCR2. Sangaralingam et al. [135] presented O-miner, a powerful online platform for combining and analyzing multi-omics data. The method aids in the discovery of critical pathways and the prioritization of biomarkers in databases that include gene, transcriptome, methylation, clinical and biological data. The pipelines created for the tool use Bioconductor packages and

Table 4 Biomarker identification using ML and DL from multi-omics data

Work [Ref]	Type	ML/DL	Algorithm	Data resource	Result	Future studies
Sinkala et al. [94]	Diagnostic biomarkers	ML	NCA, SVM, KNN	mRNAs, miRNAs and DNA methylation dataset of pan cancer from TCGA and cBioportal	KNN outperforms and accurately identified 50 mRNAs, 49 methylated genes, 14 proteins and 20 miRNAs with 99% accuracy	The identified biomarkers can be used for predicting clinical outcomes, and guiding treatment strategies will need to be assessed
Hamzeh et al. [95]		ML	SVM, RF and NB	mRNA and miRNA data of prostate cancer downloaded from GEO	NB identifies two genes with 95% accuracy and Gleason score of 7 & 6 for PIAS3 and UBE2V2 respectively	Multi-omics model based on different types of genomics data could be examined for disease diagnosis
Xu et al. [96]		ML	PCC, Hybrid feature selection, IG, LR	DNA Methylation and Gene expression dataset of cervical cancer from TCGA and validation dataset from GEO	The proposed model performed well andidentify four diagnostic markers with sensitivity and specificity value of 96.2% and 95.2% respectively	The proposed approach can be applied to the development of new epigenetic therapies
Guo et al. [97]		DL	De-noising autoencoder, k-mean clustering	mRNA-seq, miRNA-seq, CNV data from TCGA and GSE26712, and GSE32062 respectively	DL framework accurately identified 19 biomarkers and 8 KEGG pathways as diagnostic markers	More clinical features can be used to identify genes related to subtypes of ovarian cancer and, transfer learning can be used
Long et al. [98]		ML	RF, Cox regression	Multi-omics dataset from GSE16515 and GSE28735	4 diagnostic markers were identified by RF with 90% accuracy	Integration of multi-omics data in epidemiological context can be done
Long et al. [99]		ML	AUCRF, Boruta, Vita, RF, NB, KNN and LR	Multi-omics dataset collected from GSE83889, GSE44861, GSE41258, GSE8671	The proposed feature selection algorithms identified 8 diagnostic markers and RF performed best with 99% accuracy	–
Feng et al. [100]		ML	KPCA, Spectral Clustering algorithm	Gene expression, DNA methylation and isoform expression data download from GDAC firehose	Proposed method successfully identifies 4 genes of lung and liver cancer respectively as diagnostic marker	To deal with the problem of small samples, large feature of cancer data and to predict the subtype of cancer, ML and DL algorithms can be applied
Kwon et al. [101]		ML	SVM, LOOCV	mRNA and miRNA expression data download from TCGA and GEO	Presented method identify 5 diagnostic markers along with 705 multi-markers accurately	More types of data can be included for better performance
Joshi et al. [102]		DL	DNN with LR	Multi-omics data from TCGA using UCSC Xena repository	SCR-SNN outperforms by accurately identifies 15diagnostic markers with an AUC value of 89%	–

Table 4 (continued)

Work [Ref]	Type	ML/DL	Algorithm	Data resource	Result	Future studies
Cheng et al. [103]		ML	LR, PCC	DNA Methylation, gene expression and clinical data from TCGA	LR outperforms by successfully identifying 6 diagnostic markers with sensitivity of 96%	Investigation of methylation profiles for early prediction of HCC is required
Zhang et al. [104]		ML	BayesNet, Cox Regression	DNA Methylation, Gene expression, and Clinical BRCA data from TCGA p	Proposed framework identify 7 DMSs as diagnostic and prognostic markers with AUC value of 78%	–
Zhang et al. [105]		ML	RF, functional analyses	LncRNAs, miRNAs, mRNAs from TCGA portal	RF outperforms by successfully identifying 25 diagnostic markers with 80% accuracy	–
Liu et al. [52]	Prognostic biomarkers	ML	RF, LASSO, Univariate Cox	Multi-omics dataset of lung cancer patients from TCGA and GEO databases	The proposed framework identified 5 diagnostic and 8 prognostic markers for lung cancer patients	Deep learning methods can be applies in future studies
Zhao et al. [106]		ML	Multi-variate cox model	DNA methylation, Gene expression, somatic CNA and microRNA expression dataset from TCGA	Seven prognostic markers were successfully identified by multi-variate cox with c-index value of 95%	Some non-parametric algorithms can be applied to study biomarkers. This work can be extended across the world
Mishra et al. [107]		ML	LR, Cox Regression	DNA methylation, gene-expression, miRNA and lncRNA data of PDAC patients from TCGA	Five prognostic markers were identified with AUC value of 95% and Hazard ratio of each gene lies in range of 1–2	Artificial intelligence can be employed in future for better performance
Zhuang et al. [108]		ML	Univariate, multi-variate cox,	CNV, mutation, RNA-seq and SNP dataset of AML patients download from TCGA and GEO datasets	Presented framework successfully identified 10 prognostic markers of AML patients with hazard ratio in range of 1–3	–
Dong et al. [109]		ML	Ranger Algorithm, iCluster plus	Clinical, DNA methylation and gene-expression dataset from TCGA	The proposed approach identified 7 prognostic markers with c-index value of 81%	Futures studies involves exploration of biological evidence for identification of markers
Ouyang et al. [110]		ML	SCC,, fisher ratio and DT's	miRNA, somatic mutation, and DNA-methylation of LIHC utilized from TCGA	34 diagnostic markers and one prognostic marker that is p53 were identified with AUC value of 99%	In future, treatment therapies can be guided based on DEG's selected
Peng et al. [111]		DL	CapsNetMMD	DNA Methylation, miRNA expression, and CNA dataset of breast cancer patients from TCGA	CapsNetMMD outperforms various existing ML models and identified top 5% genes with sensitivity and specificity of 90%	In future, the predicted genes with prognostic values in breast cancer may serve as candidates for ecologists and medical scientists

Table 4 (continued)

Work [Ref]	Type	ML/DL	Algorithm	Data resource	Result	Future studies
Lim et al. [53]		DL	ANN, meta LR and SVM	mRNA, DNA methylation, CNV of breast cancer downloaded from TCGA	ANN outperforms Meta-SVM and Meta LR by successfully identifying the prognostic biomarkers	We could use more than two types of data sources to construct a multimodal learning for more accurate prediction
Lai et al. [112]		DL	DNN, step miner	Gene expression and clinical data from GEO	15 prognostic markers were identified by DNN with 70% accuracy	This framework can be employed to predict the survival of other cancers
Cui et al. [113]		DL	u-net architecture	Pathological images downloaded from TCGA	Four prognostic markers have been identified with c-index value of 68%	The results can be improved by integrating multi-omics dataset
Mo et al. [114]		ML	RF, Cox Regression	RNA-Seq, SNP, CNV and clinical information of breast cancer from UCSC	Proposed RF and Cox regression models accurately identified 6 prognostic genes AUC value of 80%	Experimental Validation is required because of limited clinical information present in the current research
Mo et al. [115]		ML	iCluterBayes, Fishers test, ANOVA	DNA Methylation, somatic mutation, RNA seq and CNV data fire browser	6 genes are identified by the proposed method as prognostic markers	Future studies involves the identification of markers from the large sample size datasets
Zhang et al. [54]		ML	ESTIMATE algorithm, LASSO and COX	DNA methylation, Somatic mutation and gene-expression data downloaded from TCGA	Proposed framework identified 6 expression, 5 methylation and 2 mutation genes with 79% and 83% AUC and c-index value	–
Xu et al. [116]		ML	GISTIC 2.0, Mutsig 2.0	RNA seq data from TCGA and gene expression from GSE28735,and GSE62452	The presented method outperforms and select 9 prognostic markers with 87% accuracy	Future Studies include the verification of genes in vivo and in vitro and thorough investigation will be done
Chang et al. [117]		ML	Wilcoxon Rank-sum test, PCC	CNA, RNA-seq and Gene expression data of colorectal cancer patients extracted from TCGA	The proposed pipeline selects 6 prognostic genes having p-value less than 0.05 and R-value greater than or equal to 0.3	Target drugs or treatment methods can be developed in future studies
Yuan et al. [118]		ML	CoCA, GISTIC 2.0, and Mutsig	SNP, DNA copy, DNA methylation, mRNA expression and clinical information	CoCA successfully identify two clusters HX-1 and HX-2 with 2 methylation and 15 mutation genes as prognostic markers	Deep learning can be employed in future and treatment therapies can be provided based on the identified markers
Dimitrakopolos et al. [119]	Predictive biomarkers	ML	NetICS	Genetic, mRNA, miRNA, DNA methylation from TCGA	Top 5% genes from both lung cancer and breast cancer patients are successfully identified by the NetICS with AUC value of 89%	In future, along with the genomic and transcriptomic data, more complex mutational patterns can be integrated for better performance

Table 4 (continued)

Work [Ref]	Type	ML/DL	Algorithm	Data resource	Result	Future studies
Shang et al. [120]		ML	iRank, Constrained Page Rank	DNA Methylation, RNA-Seq, miRNA-seq, CNV of HCC patients from TCGA	iRank outperforms by accurately prioritizing the cancer genes with an accuracy of 81%	–
Guan et al. [121]		ML	PCC, SVM	Gene Expression and DNA methylation from TCGA	The proposed method identifies 10 predictive genes with 80% accuracy	This research can be extended to additional dataset for better results
Yao et al. [122]		ML	MetPriCNet	Genomic, phenome, and metabolome data from STRING,OMIM and TCGA	MetPriCNet prioritize and predict the candidate genes with an AUC value of 91%	It can be used in different fields of biomedicine like disease prediction, drug discovery, and target discovery
Fortino et al. [123]		ML	RF, Fuzzy Logic	Four multi-class gene expression dataset from GEO	RF performed well by successfully prioritizing the genes with 96% accuracy	
Fan et al. [124]		ML	COX	Multi-omics dataset downloaded from GEO	The proposed method identified 10 predictive markers with an 86% AUC	Deep learning can be employed in future for better performance
Chen et al. [125]		ML	BRIDGE, Lasso, regression	PPI, GE, GS, KEGG and GO dataset from TCGA portal	BRIDGE with regression and lasso model identified 8 and 28 genes of obesity and diabetic patients	–
Zhang et al. [126]		ML	Network based approach	PPI, gene expression, and mutation data of KIRC, THCA, and HNSC from TCGA, OMIM and GEO	Proposed method outperforms various existing methods by accurately identifying top 5 genes for each cancer type with 80% accuracy	Gene expression, CNV, and methylation data will be integrated to construct a network
Valdeolivas et al. [127]		ML	RWR-M, RWR-MH	Multi-omics dataset from TCGA, and OMIM	RWR-M and RWR-MH outperforms with an 89% and 82% accuracy	–
Wei et al. [128]		ML	Driver-IRH	Multi-omics data of BRCA, HNSC, KIRC and THCA from TCGA	Driver_IRW successfully identified top 10 genes for each cancer type with 90% precision and recall value	This method can be applied to classify patients in different subtypes of cancer by using the identified genes
Zheng et al. [129]		ML	RF	Clinical data from TCGA	Proposed method outperforms and identified top 10 predictive markers	To adjust the nuisance covariates, regression models can be used
Yang et al. [130]		ML	MapGene	Multi-omics dataset from DisGeNet and String	MapGene outperforms with an 87% precision and 90% recall value	

statistical methods and run in R and Python environment. Abstract et al. [136] developed an open-source ML tool called omics-learn for biomarker discovery. A genomic and proteomics dataset was used for the experiment. Python libraries are used to develop the tool, and it can be downloaded using a local server. This tool used the XGBoost model for training the dataset. The visualization and web interface of omics-learn are built using StreamLit. Leclercq et al. [137] developed a biomarker discovery tool called BioDiscML using multi-omics data comprising genomic, proteomic, and pathological datasets. BioDiscML followed a variety of ML algorithms to identify the optimal set of biomarkers. This tool uses a vast range of ML classifiers within a completely integrated framework that often includes data pre-processing, making it easier for non-machine learning experts to complete their tasks. Song et al. [138] proposed an integrative analysis tool called iProFun for the biomarker identification using Proteomic, CNA, and DNA Methylation datasets. This tool was used on Ovarian cancer patients. The collected data were pre-processed and integrated for further evaluation. Ghannoum et al. [139] presented an open-source pipeline named DIscBIO to identify the genes using transcriptomic data. The authors used two scRNA-seq datasets to demonstrate the pipeline capabilities. All analyses are accessible as notebooks with R coding, explanatory language, output data, and images. The pipeline is implemented in four steps: data pre-processing, cellular clustering, retrieving DEGs, and signature discovery. Netanelly et al. [140] developed a framework Profiler of Multi-omics data (PROMO) for analyzing, pre-processing, clustering, and visualizing the single omics and multi-omics data simultaneously. Further, this tool is also used for biomarker discovery and survival analysis. In this tool, statistical tests are used to identify DEGs, which are further passed to Cox models for survival analysis. Tang et al. [141] developed a web server Gene Expression Profiling Interactive Analysis (GEPIA2), for the biomarker identification using the gene-expression dataset. GEPIA2 works efficiently for 84 cancer subtypes. This tool also helps to classify cancer based on different subtypes. This website is freely accessible and implemented using HTML, javascript, and Php language. Wang et al. [142] developed an online survival web server OScc to validate the prognostic signatures from the gene-expression dataset. This tool is tested on four gene-expression datasets retrieved from GEO and TCGA platforms. This tool will generate a survival curve for p-value, hazard ratio, and log-rank test. Treatment will be provided to the high-risk patients based on the values achieved. Champion et al. [143] software algorithm AMARETTO for discovering cancer genes by incorporating gene expression, DNA methylation, and CNV datasets. Then co-expressed target genes are connected to the driver genes, known as regulatory modules. Then these driver genes are converted into a network to identify cancer genes. AMARETTO is applied to patients from 11 different sites, and it is considered the best tool for identifying cancer genes. Jang et al. [144] developed a web application called Cancer Patient Stratification and Survival Analysis (CAPSAA) to evaluate predictive values of candidate biomarkers by dynamically visualizing the survival stratification for different subgroups of patients. The subgroups are made from gene expression, CNA, and mutation data downloaded from TCGA coherent. Hierarchical clustering is done to divide the patients into subgroups, and this tool is implemented on Lung Cancer patients, which is freely accessible. Xie et al. [145] designed a repository MOBCdb to integrate genetic, clinical, transcriptomic, and epigenomic results. The database was created to enable users to collect data from breast cancer patients' SNV, gene expression, and microRNA. And DNA methylation. An interface is available in MOBCdb for concurrently visualizing multi-omics data from different samples. This data is also subjected to a survival study using MOBCdb's survival module. MOBCdb aids precision medicine by detecting new markers in different subtypes of breast cancer through its comprehensive web interface. Mohammed et al. [146] developed a pipeline named CancerDiscover to predict cancer classes and identify the cancer biomarkers. The tool assists with normalization and offers various function filtering approaches to select the best performing functions. High-throughput raw datasets can be analyzed automatically and reliably with CancerDiscover. CancerDiscover is an open-source platform that is free to download. Chong et al. [147] presented an update to MetaboAnalyst (version 4.0) to analyze metabolomic data. This tool has added four new features to the previous version of MetaboAnalyst, including real-time R command monitoring and show, as well as the introduction of the MetaboAnalystR kit, a Pathway module to predict pathway behavior, Metaanalysis module for comprehensive signature recognition, and a Network explorer which integrates transcriptomic, metagenomics, and metabolomics dataset. Zeng et al. [148] developed Immuno-Oncology Biological Research (IOBR) to identify gene signatures based on a multi-omics dataset. This tool provides batch analysis of the gene markers and their association with lncRNA profiling, clinical phenotypes, genetic characteristics, and the signatures produced from single-cell RNA sequencing data. Moreover, this tool integrates deconvolution methodologies with various signature construction tools to identify gene signatures. This tool is freely available to use, and it is an effective and flexible tool. Liu et al. [149] developed a web server GSCALite to analyze gene sets related to cancer. This tool includes identifying differential expressed genes from mRNA expression, CNV, Methylation, and SNV data and the survival analysis using these genes, detection of genomic variation along with survival analysis, cancer pathway activity analysis, and identification of drug sensitivity related to genes.

Table 5 Existing tools for biomarker identification

[Ref]	Type of data	Tool	Year	Technology	Link
[131]	Genomic + Metabolome	mixOmics	2019	R/Bioconductor	http://mixomics.org/
[132]	Gene-Expression	Web Server	2020	Cloud	https://webs.iiitd.edu.in/raghava/hccpred/
[133]	Genomic + Epienomic	CancerLSP	2019	Weka	http://webs.iiitd.edu.in/raghava/cancerlsp/
[134]	Genomic + Radiology + Clinical	Imaging-AMARETTO	2020	Bioconductor/R Jupiter notebook	http://portals.broadinstitute.org/pochetlab/JCO_CCI_Imaging-AMARETTO/Imaging-AMARETTO_Software_Resources.html
[135]	Transcriptome + genome + Methylation	O-miner	2019	R/Python	http://www.o-miner.org
[136]	Genomic + Proteomic	Omics-Learn	2021	Python	https://omiclearn.com/
[137]	Genomic + Proteomic + pathological	BioDiscML	2019	Java/Weka	https://github.com/mickaellec lercq/BioDiscML
[138]	CNA + DNA Methylation + Proteome	iProFun	2019	R	https://github.com/songxiaoyu/iProFun
[139]	Transcriptomic Data (scRNA-seq)	DIscBIO	2021	R/Jupiter	https://github.com/ocbe-uio/DIscB IO
[140]	Genomic + Transcriptomic + Meabolome	PROMO	2019	Matlab	http://acgt.cs.tau.ac.il/promo/
[141]	Gene-expression + RNA Sequencing	GEPIA2	2018	Javascript/PhP	https://gepia2.cancer-pku.cn/#index
[142]	Gene-expression	OSCC	2019	R/java/cloud	http://bioinfo.henu.edu.cn/CESC/CESCList.jsp
[143]	Gene-expression + CNV + DNA Methylation	AMARETTO	2017	R	https://bitbucket.org/gevaertlab/pancanceramaretto
[144]	CNV + Gene Expression + Somantic Mutation	CAPSAA	2019	Clojure/Fig Wheel	http://capssa.ewha.ac.kr/
[145]	Gene expression + SNV + DNA methylation	MOBCdb	2018	Perl, R, MySQL	http://bigd.big.ac.cn/MOBCd b/
[146]	Gene Expression + Sequencing	CancerDiscover	2017	WEKA, Affy R package	https://github.com/HelikarLab/CancerDiscover
[147]	Metabolome + transcriptome + metagenome	MetaboAnalyst 4.0	2018	R/Google Cloud Server	https://github.com/xia-lab/Metab oAnalystR
[148]	lncRNA + RNA + genomic	IOBR	2020	R	https://github.com/IOBR/IOBR
[149]	mRNA + CNV + SNV + Methylation	GSCALite	2018	R scripts/maftool	http://bioinfo.life.hust.edu.cn/web/GSCALite/
[150]	Gene Expression + DNA Metylation	OSdlbcl	2020	J2EE platform	https://bioinfo.henu.edu.cn/DLBCL/DLBCLList.jsp

Dong et al. [150] developed an Online Survival analysis web server for Diffuse Large Cell Lymphoma (OSdlbcl) to identify prognostic value for some specific gene. Clinical follow-up information and gene expression profiles of 1100 samples were used from TCGA and GEO databases. Moreover, DNA methylation data was also used for prediction purposes. This tool will develop a Kaplan–Meier (KM) plot, which will give the p-value, hazard ratio, and log rank for some specific gene symbol. Table 5 shows the existing work on biomarker identification using multi-omics data with the help of tools used.

6 Discussion

This survey focuses on various methods and techniques for the identification of biomarkers using multi-omics data are described. The most recent and important research papers are analyzed in this survey. The goal of this review is to concentrate on biomarker identification approaches including ML and DL and tools using multi-omics data, as this is anticipated to be a popular topic in the future due to the need for targeted therapy.

The biomarker and its various types including risk, prognostic, diagnostic, predictive, safety, monitoring and response are addressed in answer to the first research problem. Using omics and multi-omics data, a lot of research is being done on identifying diagnostic, prognostic, and predictive markers.

The multi-omics data, the types of omics data (genomics, transcriptomic, proteomic, metabolome, and interactome) and the available databases required for biomarker identification is presented in answer to second research problem.

The answer to third problem is addressed by explaining the feature extraction and selection and their techniques including filter method, wrapper method and embedded methods along with their advantages and disadvantages. In literature, the work done on biomarker identification using feature selection and extraction techniques is described.

The techniques required for biomarker identification including ML and DL for biomarker identification are presented in answer to fourth research problem. The work done by researchers for diagnostic, prognostic and predictive biomarker identification using omics and multi-omics data with the help of ML and are described in literature.

In answer to the fifth research issue, a list of current publicly available tools is discussed, along with their limitations. The link to access them is also provided. The majority of tools are open source, and people can use them to complete their activities. Some tools are built on a cloud network using servers, and packages are made accessible on request.

Finally, the last research problem is addressed by reviewing the challenges of identifying biomarkers using multi-omics data. Recommendations for future research for biomarker discovery are presented based on a systematic analysis of related publications in the literature.

6.1 Challenges in Biomarker Identification

Some problems have been faced while performing the review of existing techniques for biomarker identification using multi-omics data which are shown in Fig. 9 and are described below.

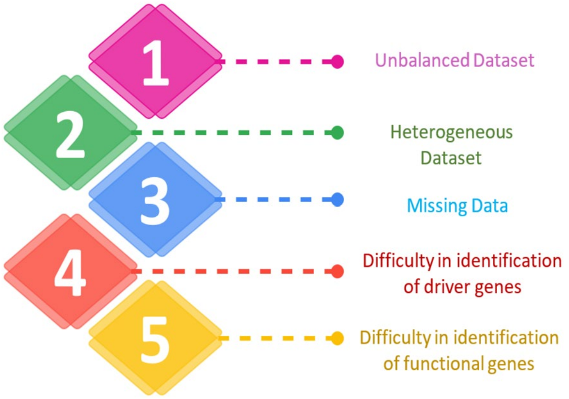

1
Unbalanced Dataset
2
Heterogeneous Dataset
3
Missing Data
4
Difficulty in identification of driver genes
5
Difficulty in identification of functional genes

Fig. 9 Challenges in biomarker identification

- Unbalanced dataset: For biomarker identification, omics data including genome, transcriptome, protein, metabolites, and peptides are used. The available dataset is present in unbalanced form. It means that the variables and attributes are too big than the sample size. This leads to overfitting problem. Therefore, it is very difficult to identify biomarkers using unbalanced dataset. This problem can be eliminated by integrating the different type of dataset and used that integrated dataset for biomarker identification. The feature extraction technique called mRMR can be also be employed to solve this problem.
- Heterogeneous datasets: In biomarker identification, some of the molecular profiles are highly heterogeneous. They can be divided into categorical and continuous and sometimes may be scattered into multiple inputs. It makes the biomarker identification difficult. Therefore, different machine learning algorithm including graph network, clustering approaches and deep learning techniques can be applied to remove heterogeneity.
- Missing Data: In multi-omics biomarker identification, data missing ness is a major challenge. Image noise, batch impacts, and hybridization failures all cause data missing ness in microarray data. Due to this complication, appropriate imputation of missed values based on practice, a mixture of methods, and trial and error is required. One of the most common ML algorithm i.e. KNN is used to impute the missing values. Instead, we can also use median of the attribute and impute that median in place of the missing value.
- Difficulty in identification of driver genes: There are different types of omics data. Sometimes it is not possible to identify driver genes on the basis of single type of data. For example: we identify the genes using the genomic data, but these may not be enough for disease detection. Therefore, another type is also required to identify the cancer genes. Hence, multi-omics is required to identify the driver genes required for disease diagnosis and prognosis.
- Difficulty to identify functional genes: Genomic data focus of DNA data to identify mutations related to cancer. The DNA involves different changes starting from small somatic mutations, several insertions, deletions and large CNV data for the identification of cancer mutations. The mutation further varies in different sub types of cancer. Therefore, it is difficult to identify which function gene is growing the cancer. To solve this challenge,

different deep learning techniques and gene prioritization algorithms are required.

6.2 Future Research Directions

Based on the current literature, the following are potential future directions in this field of study.

- The present research is mainly focused on a single type of dataset. Multi-omics integration is required efficient analysis [74, 78, 80, 96, 110].
- Effective non-parametric methods comprising CHI2, Kruskal–Wallis, Wilcoxon rank-sum test, and Spearman's rank correlation are required for the identification of biomarkers [106].
- Deep learning (DL) algorithms are required for the identification of biomarkers necessary for the prognosis of cancer and provide a more powerful tool for targeted therapy [45, 50, 52, 82, 99, 101].
- A broad sample size dataset is needed to allow for a thorough examination of the disease's progression, diagnosis, and treatment [83, 89, 122].
- AI-based technologies can be used to identify predictive biomarkers which will significantly increase the prediction accuracy [107].
- The present research lacks treatment therapies which can be provided using the identified biomarkers [78, 95, 97, 99, 118].
- Limited methods exist for biomarker identification due to the heterogeneity of omics data sets [98, 113].
- Next Generation Sequencing data analysis can be done for biomarker identification using ML, DL, Quantum Neural Nets and Quantum Computing in future for better performance [151].

7 Conclusion

The collection of different forms of omics data in the post-genomics period allows for the screening of specific markers for accurate diagnosis and prognosis, which is essential in personalized medicine. Unfortunately, identifying biomarkers from a large volume of omics data, particularly when there are complex interactions between molecules, is a difficult task. In this article, different existing approaches, feature extraction/selection techniques, tools and technologies for the identification of diagnostic, prognostic, and predictive biomarkers using omics and multi-omics data have been studied. Their comparative study has been performed by analyzing the ML and DL approaches used by the authors. From the research, it is found that single type of data is not enough for identification of genes in patients. Therefore, multi-omics data is required for accurate discovery of markers and to guide treatment therapies based on the identified markers. We hope that by conducting this survey, researchers will be able to learn which algorithms can be used to identify the biomarkers and how to apply specific techniques including ML and DL, and tools to precision medicine.

Funding The authors have no funding to report.

Declarations

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

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

References

1. Collins FS, Varmus H (2015) A new initiative on precision medicine. N Engl J Med 372:793–795
2. Cagney DN, Sul J, Huang RY et al (2017) The FDA NIH Biomarkers, EnfpointS, and other Tools (BEST) Resource in Neurology. Neuro-Oncology 20:1162–1172. https://doi.org/10.1093/neuonc/nox242
3. Zhu K, Zhan H, Peng Y et al (2020) Plasma hsa_circ_0027089 is a diagnostic biomarker for hepatitis B virus-related hepatocellurar carcinoma. Carcinogenesis 41:296–302. https://doi.org/10.1093/carcin/bgz154
4. Fattahi S, Kosari-Monfared M, Golpour M et al (2020) LncRNAs as potential diagnostic and prognostic biomarkers in gastric cancer: a novel approach to personalized medicine. J Cell Physiol 235:3189–3206. https://doi.org/10.1002/jcp.29260
5. Marquardt JU, Galle PR, Teufel A (2012) Molecular diagnosis and therapy of hepatocellular carcinoma (HCC): an emerging field for advanced technologies. J Hepatol 56:267–275. https://doi.org/10.1016/j.jhep.2011.07.007
6. The Cancer Genome Atlas Program. https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga. Accessed 20 Jan 2021
7. (2021) TARGET. https://ocg.cancer.gov/programs/target/overview. Accessed 20 Feb 2021
8. (2021) ICGC Data Portal. https://dcc.icgc.org/. Accessed 28 Feb 2021
9. Cao H, Schwarz E (2019) Opportunities and challenges of ML approaches for biomarker signature identification in psychiatry. Elsevier Inc., Amsterdam
10. Kaur P, Singh A, Chana I (2021) Computational techniques and tools for omics data analysis: state-of-the-art, challenges, and future directions. Arch Comput Methods Eng. https://doi.org/10.1007/s11831-021-09547-0
11. Zhang ZY (2015) Healthcare engineering defined: a white paper. J Healthc Eng 6(4):635–648. https://doi.org/10.1260/2040-2295.6.4.635
12. Swan AL, Mobasheri A, Allaway D et al (2013) Application of ML to proteomics data: classification and biomarker identification in postgenomics biology. OMICS 17:595–610. https://doi.org/10.1089/omi.2013.0017

13. Qin G, Zhao XM (2014) A survey on computational approaches to identifying disease biomarkers based on molecular networks. J Theor Biol 362:9–16. https://doi.org/10.1016/j.jtbi.2014.06.007
14. Jagga Z, Gupta D (2015) ML for biomarker identification in cancer research developments toward its clinical application. Pers Med 12:371–387. https://doi.org/10.2217/PME.15.5
15. Dragani TA, Matarese V, Colombo F (2020) Biomarkers for early cancer diagnosis: prospects for success through the lens of tumor genetics. BioEssays 42:1–6. https://doi.org/10.1002/bies.201900122
16. Shi K, Lin W, Zhao X (2020) Identifying molecular biomarkers for diseases with ML based on integrative omics. IEEE/ACM Trans Comput Biol Bioinform 5963:1–1. https://doi.org/10.1109/tcbb.2020.2986387
17. Kaur H, Kumar R, Lathwal A, Raghava GPS (2021) Computational resources for identification of cancer biomarkers from omics data. Brief Funct Genomics 00:1–10. https://doi.org/10.1093/bfgp/elab021
18. (2021) What are biomarkers. https://www.mycancer.com/resources/what-are-biomarkers/. Accessed 25 Jan 2021.
19. Khan TK (2016) Introduction to Alzheimer's disease biomarkers. Biomarkers Alzheimers Dis. https://doi.org/10.1016/b978-0-12-804832-0.00001-8
20. Sechidis K, Papangelou K, Metcalfe PD et al (2018) Distinguishing prognostic and predictive biomarkers: an information theoretic approach. Bioinformatics 34:3365–3376. https://doi.org/10.1093/bioinformatics/bty357
21. Pezo RC, Bedard PL (2015) Definition: translational and personalised medicine, biomarkers, pharmacodynamics. https://oncologypro.esmo.org/content/download/67864/1221489/1/2015-ESMO-Handbook-Translational-Research-Chapter-1.pdf
22. Matheis K, Laurie D, Andriamandroso C et al (2011) A generic operational strategy to qualify translational safety biomarkers. Drug Discov Today 16:600–608. https://doi.org/10.1016/j.drudis.2011.04.011
23. Jones K, Nourse JP, Keane C et al (2014) Plasma microRNA are disease response biomarkers in classical Hodgkin lymphoma. Clin Cancer Res 20:253–264. https://doi.org/10.1158/1078-0432.CCR-13-1024
24. Ibraheem O, Adigun RO, Olatunji IT (2018) Omics technologies in unraveling plant stress responses; using Sorghum as a model crop, how far have we gone? Int J Plant Res 31:1–18. https://doi.org/10.4172/2229-4473.1000405
25. Bravo-Merodio L, Williams JA, Gkoutos GV, Acharjee A (2019) Omics biomarker identification pipeline for translational medicine. J Transl Med 17(1):1–10. https://doi.org/10.1186/s12967-019-1912-5
26. Subramanian I, Verma S, Kumar S et al (2020) Multi-omics data integration, interpretation, and its application. Bioinform Biol Insights 14:7–9. https://doi.org/10.1177/1177932219899051
27. Husi H, Albalat A (2014) Proteomics. Handb Pharm Stratif Med 147–179. https://doi.org/10.1016/b978-0-12-386882-4.00009-8
28. Mestrovic T (2020) Proteomics uses. https://www.news-medical.net/life-sciences/Proteomics-Uses.aspx. Accessed 28 Jan 2020
29. Kim M, Tagkopoulos I (2018) Data integration and predictive modeling methods for multi-omics datasets. Mol Omics 14(1):8–25. https://doi.org/10.1039/c7mo00051k
30. Hasin Y, Seldin M, Lusis A (2017) Multi-omics approaches to disease. Genome Biol 18(1):1–15. https://doi.org/10.1186/s13059-017-1215-1
31. Cortese-Krott MM, Santolini J, Wootton SA et al (2019) The reactive species interactome. Elsevier Inc., Amsterdam
32. 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
33. Dhillon A, Singh A (2020) EBreCaP: extreme learning-based model for BRCA survival prediction. IET Syst Biol 14(3):160–169. https://doi.org/10.1049/iet-syb.2019.0087
34. Jollife IT, Cadima J (2016) Principal component analysis: a review and recent developments. Philos Trans R Soc A. https://doi.org/10.1098/rsta.2015.0202
35. Izenman AJ (2013) Linear discriminant analysis. Springer, New York
36. Gillis N (2020) Nonnegative matrix factorization. Society for Industrial and Applied Mathematics, Philadelphia
37. Schölkopf B, Smola A, Müller KR (1998) Nonlinear component analysis as a kernel eigenvalue problem. Neural Comput 10(5):1299–1319. https://doi.org/10.1162/089976698300017467
38. De Ridder D, Kouropteva O, Okun O et al (2003) Supervised locally linear embedding. Lect Notes Comput Sci 2714:333–341. https://doi.org/10.1007/3-540-44989-2_40
39. Van Der Maaten L, Hinton G (2008) Visualizing data using t-SNE. J Mach Learn Res 9(11):2579–2625
40. Wang Y, Yao H, Zhao S (2016) Auto-encoder based dimensionality reduction. Neurocomputing 184:232–242. https://doi.org/10.1016/j.neucom.2015.08.104
41. Ding H (2016) Visualization and integrative analysis of cancer multi-omics data. The Ohio State University, Columbus
42. Bommert A, Sun X, Bischl B et al (2020) Benchmark for filter methods for feature selection in high-dimensional classification data. Comput Stat Data Anal 143:106839. https://doi.org/10.1016/j.csda.2019.106839
43. Xie Y, Meng W-Y, Li R-Z et al (2021) Early lung cancer diagnostic biomarker discovery by ML methods. Transl Oncol 14(1):100907. https://doi.org/10.1016/j.tranon.2020.100907
44. Khatri I, Bhasin MK (2020) A transcriptomics-based meta-analysis combined with ML approach identifies a secretory biomarker panel for diagnosis of pancreatic adenocarcinoma. medRxiv. https://doi.org/10.1101/2020.04.16.20061515
45. Liu B, Liu Y, Pan X et al (2019) DM markers for pan-cancer prediction by DL. Genes (Basel). https://doi.org/10.3390/genes10100778
46. Senthil Kumar P, Lopez D (2016) A review on feature selection methods for high dimensional data. Int J Eng Technol 8(2):669–672
47. Darst BF, Malecki KC, Engelman CD (2018) Using recursive feature elimination in random forest to account for correlated variables in high dimensional data. BMC Genet 19(1):1–6. https://doi.org/10.1186/s12863-018-0633-8
48. Aha DW, Bankert RL (1996) A comparative evaluation of sequential feature selection algorithms. In: Fisher D, Lenz HJ (eds) Learning from data. Springer, New York, pp 56–63
49. Mirjalili S (2019) Genetic algorithm. Evol Algorithms Neural Netw 780:43–55. https://doi.org/10.1007/978-3-319-93025-1_4
50. Yu J, Zhu M, Lv M et al (2019) Characterization of a five-microRNA signature as a prognostic biomarker for esophageal squamous cell carcinoma. Sci Rep 9(1):1–11. https://doi.org/10.1038/s41598-019-56367-1
51. Lal TN, Chapelle O, Weston J (2006) Embedded methods. Study Fuzziness Soft Comput 165:137–165
52. Liu P, Tian W (2020) Identification of DM patterns and biomarkers for clear-cell renal cell carcinoma by multi-omics data analysis. PeerJ 8:1–31. https://doi.org/10.7717/peerj.9654
53. Lim J, Bang S, Kim J et al (2019) Integrative DL for identifying differentially expressed (DE) biomarkers. Comput Math Methods Med. https://doi.org/10.1155/2019/8418760
54. Zhang Y, Yang M, Ng DM et al (2020) Multi-omics data analyses construct TME and identify the immune-related prognosis signatures in human LUAD. Mol Ther Nucleic Acids 21:860–873. https://doi.org/10.1016/j.omtn.2020.07.024

55. Dhillon A, Singh A (2019) ML in healthcare data analysis: a survey. J Biol Todays World 8(6):1–10
56. Hastie T, Tibshirani R, Friedman J (2009) Overview of supervised learning. Elem Stat Learn 27(2):83–85. https://doi.org/10.1007/b94608
57. Quinlan JR (1993) C4.5: programs for ML. Morgan Kaufman Publishers, San Francisco
58. Ghahramani Z (2004) Unsupervised learning. Mach Learn. https://doi.org/10.1007/978-3-540-28650-9_5
59. Goldberg AB, Zhu X (2009) Introduction to semi-supervised learning. Morgan & Claypool, San Rafael
60. Esteva A, Robicquet A, Ramsundar B et al (2019) A guide to DL in healthcare. Nat Med 25(1):24–29. https://doi.org/10.1038/s41591-018-0316-z
61. Chung NC et al (2019) Unsupervised classification of multi-omics data during cardiac remodeling using DL. Methods 166:66–73
62. Kamel HFM, Al-Amodi HSB (2015) Cancer biomarkers role. Biomarkers Med 45:1–32. https://doi.org/10.5772/62421
63. George ED, Sadovsky R (1999) Multiple myeloma: recognition and management. Am Fam Physician 59(7):1885–1892
64. Biomarker.en.wikipedia.org/wiki/Biomarker. Accessed 28 Jan 2021
65. Chatterjee SK, Zetter BR (2005) Cancer biomarkers: knowing the present and predicting the future. Futur Oncol 1(1):37–50. https://doi.org/10.1517/14796694.1.1.37
66. Kitchenham B, Brereton O, Budgen B, Turner M, Bailey J, Linkman S (2009) Systematic literature reviews in software engineering—a systematic literature review. Inf Softw Technol 51(1):7–15. https://doi.org/10.1016/j.infsof.2008.09.009
67. 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
68. Fujita N, Mizuarai S, Murakami K, Nakai K (2018) Biomarker discovery by integrated joint non-negative matrix factorization and pathway signature analyses. Sci Rep 8(1):1–10. https://doi.org/10.1038/s41598-018-28066-w
69. Jia Y, Shen M, Zhou Y, Liu H (2020) Development of a 12-biomarkers-based prognostic model for pancreatic cancer using multi-omics integrated analysis. Acta Biochim Pol 67(4):501–508. https://doi.org/10.18388/ABP.2020_5225
70. Southekal S, Mishra NK, Guda C (2021) Pan-cancer analysis of human kinome gene expression and promoter DNA methylation identifies dark kinase biomarkers in multiple cancers. Cancers (Basel) 13:1189. https://doi.org/10.3390/cancers13061189
71. Moon M, Nakai K (2018) Integrative analysis of gene expression and DM using unsupervised feature extraction for detecting candidate cancer biomarkers. J Bioinform Comput Biol 16(2):1850006. https://doi.org/10.1142/S0219720018500063
72. Hamzeh O, Rueda L (2019) A gene-disease-based ML approach to identify prostate cancer biomarkers. In: ACM-BCB 2019—proceedings of the 10th ACM international conference on bioinformatics, computational biology and health informatics. pp 633–638
73. Zhao X, Dou J, Cao J et al (2020) Uncovering the potential differentially expressed miRNAs as diagnostic biomarkers for hepatocellular carcinoma based on ML in the Cancer Genome Atlas database. Oncol Rep 43(6):1771–1784. https://doi.org/10.3892/or.2020.7551
74. Kloten V, Becker B, Winner K et al (2013) Promoter hypermethylation of the tumor-suppressor genes ITIH5, DKK3, and RASSF1A as novel biomarkers for blood-based BRCA screening. BRCA Res 15(1):1–11. https://doi.org/10.1186/bcr3375
75. Rehman O, Zhuang H, Ali AM et al (2019) Validation of miRNAs as BRCA biomarkers with a ML approach. Cancers (Basel) 11(3):1–10. https://doi.org/10.3390/cancers11030431
76. Alkhateeb A, Rezaeian I, Singireddy S et al (2019) Transcriptomics signature from next-generation sequencing data reveals new transcriptomic biomarkers related to prostate cancer. Cancer Inform. https://doi.org/10.1177/1176935119835522
77. Jin T, Talos FM, Wang D (2019) ECMarker: interpretable ML model identifies gene expression biomarkers predicting clinical outcomes and reveals molecular mechanisms of human disease in early stages. bioRxiv. https://doi.org/10.1101/825414
78. Tyanova S, Albrechtsen R, Kronqvist P et al (2016) Proteomic maps of BRCA subtypes. Nat Commun 7(1):1–11. https://doi.org/10.1038/ncomms10259
79. Muazzam F (2020) Multi-class cancer classification and biomarker identification using DL. bioRxiv. https://doi.org/10.1101/2020.12.24.424317
80. Toth R, Schiffmann H, Hube-Magg C et al (2019) Random forest-based modelling to detect biomarkers for prostate cancer progression. Clin Epigenet 11(1):148–163. https://doi.org/10.1101/602334
81. Ma B, Geng Y, Meng F et al (2020) Identification of a sixteen-gene prognostic biomarker for lung adenocarcinoma using a ML method. J Cancer 11(5):1288–1298. https://doi.org/10.7150/jca.34585
82. Hossain MA, Saiful Islam SM, Quinn JMW et al (2019) ML and bioinformatics models to identify gene expression patterns of ovarian cancer associated with disease progression and mortality. J Biomed Inform 100:103313. https://doi.org/10.1016/j.jbi.2019.103313
83. Cai J, Li B, Zhu Y et al (2017) Prognostic biomarker identification through integrating the gene signatures of hepatocellular carcinoma properties. EBioMedicine 19:18–30. https://doi.org/10.1016/j.ebiom.2017.04.014
84. Ghosal S, Das S, Pang Y et al (2020) Long intergenic noncoding RNA profiles of pheochromocytoma and paraganglioma: a novel prognostic biomarker. Int J Cancer 146(8):2326–2335. https://doi.org/10.1002/ijc.32654
85. Li Y, Lu S, Lu S et al (2020) A prognostic nomogram integrating novel biomarkers identified by ML for cervical squamous cell carcinoma. J Transl Med 18(1):1–12. https://doi.org/10.1186/s12967-020-02387-9
86. Liu F, Xing L, Zhang X, Zhang X (2019) A four-pseudogene classifier identified by ML serves as a novel prognostic marker for survival of osteosarcoma. Genes (Basel) 10(6):414. https://doi.org/10.3390/genes10060414
87. Xing L, Zhang X, Zhang X, Tong D (2020) Expression scoring of a small-nucleolar-RNA signature identified by ML serves as a prognostic predictor for head and neck cancer. J Cell Physiol 235(11):8071–8084. https://doi.org/10.1002/jcp.29462
88. Long NP, Jung KH, Yoon SJ et al (2017) Systematic assessment of cervical cancer initiation and progression uncovers genetic panels for DL-based early diagnosis and proposes novel diagnostic and prognostic biomarkers. Oncotarget 8(65):109436–109456. https://doi.org/10.18632/oncotarget.22689
89. Wong KK, Rostomily R, Wong STC (2019) Prognostic gene discovery in glioblastoma patients using DL. Cancers (Basel) 11(1):1–15. https://doi.org/10.3390/cancers11010053
90. Nam Y, Jhee JH, Cho J et al (2019) Disease gene identification based on generic and disease-specific genome networks. Bioinformatics 35(11):1923–1930. https://doi.org/10.1093/bioinformatics/bty882
91. Zhao T, Hu Y, Peng J, Cheng L (2020) GCN-CNN A novel DL method for prioritizing lncRNA target genes. Bioinformatics 36(16):4466–4472. https://doi.org/10.1093/bioinformatics/btaa428
92. Zhang Y, Chen Y, Hu T (2020) PANDA: prioritization of autism-genes using network-based deep-learning approach. Genet Epidemiol 44(4):382–394. https://doi.org/10.1002/gepi.22282

93. Jiang X, Zhao J, Qian W et al (2020) A generative adversarial network model for disease gene prediction with RNA-seq data. IEEE Access 8:37352–37360. https://doi.org/10.1109/ACCESS.2020.2975585
94. Sinkala M, Mulder N, Martin D (2020) ML and network analyses reveal disease subtypes of pancreatic cancer and their molecular characteristics. Sci Rep 10(1):1–14. https://doi.org/10.1038/s41598-020-58290-2
95. Hamzeh O, Alkhateeb A, Zheng JZ et al (2019) A hierarchical ML model to discover Gleason grade-specific biomarkers in prostate cancer. Diagnostics. https://doi.org/10.3390/diagnostics9040219
96. Xu W, Xu M, Wang L et al (2019) Integrative analysis of DM and gene expression identified cervical cancer-specific diagnostic biomarkers. Signal Transduct Target Ther 4(1):1–11. https://doi.org/10.1038/s41392-019-0081-6
97. Guo LY, Wu AH, Wang YX et al (2020) DL-based ovarian cancer subtypes identification using multi-omics data. BioData Min 13(1):1–12. https://doi.org/10.1186/s13040-020-00222-x
98. 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 (Basel) 11(2):155. https://doi.org/10.3390/cancers11020155
99. 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
100. Feng J, Jiang L, Li S et al (2021) Multi-omics data fusion via a joint kernel learning model for cancer subtype discovery and essential gene identification. Front Genet 12:1–10. https://doi.org/10.3389/fgene.2021.647141
101. Kwon MS, Kim Y, Lee S et al (2017) Erratum: integrative analysis of multi-omics data for identifying multi-markers for diagnosing pancreatic cancer. [BMC Genomics. (2015), 16, Suppl 9: (S4)]. BMC Genomics 18(1):1–10. https://doi.org/10.1186/s12864-016-3464-x
102. Joshi P, Jeong S, Park T (2020) Sparse superlayered neural network-based multi-omics cancer subtype classification. Int J Data Min Bioinform 24(1):58–73. https://doi.org/10.1504/IJDMB.2020.109500
103. Cheng J, Wei D, Ji Y et al (2018) Integrative analysis of DM and gene expression reveals hepatocellular carcinoma-specific diagnostic biomarkers. Genome Med 10(1):1–11. https://doi.org/10.1186/s13073-018-0548-z
104. Zhang M, Wang Y, Wang Y et al (2020) Integrative analysis of DM and gene expression to determine specific diagnostic biomarkers and prognostic biomarkers of BRCA. Front Cell Dev Biol 8:1–16. https://doi.org/10.3389/fcell.2020.529386
105. Zhang M, Cheng L, Zhang Y (2020) Characterization of dysregulated lncRNA-ASSOCIATED ceRNA network reveals novel lncRNAs With ceRNA activity as epigenetic diagnostic biomarkers for osteoporosis risk. Front Cell Dev Biol 8:1–9. https://doi.org/10.3389/fcell.2020.00184
106. Zhao N, Guo M, Wang K et al (2020) Identification of pan-cancer prognostic biomarkers through integration of multi-omics data. Front Bioeng Biotechnol 8:1–15. https://doi.org/10.3389/fbioe.2020.00268
107. Mishra NK, Southekal S, Guda C (2019) Survival analysis of multi-omics data identifies potential prognostic markers of pancreatic ductal adenocarcinoma. Front Genet 10:1–18. https://doi.org/10.3389/fgene.2019.00624
108. Zhuang H, Chen Y, Sheng X et al (2020) Searching for a signature involving 10 genes to predict the survival of patients with acute myelocytic leukemia through a combined multi-omics analysis. PeerJ 8(6):e9437. https://doi.org/10.7717/peerj.9437
109. Dong X, Zhang R, He J et al (2019) Trans-omics biomarker model improves prognostic prediction accuracy for early-stage lung adenocarcinoma. Aging (Albany NY) 11(16):6312–6335. https://doi.org/10.18632/aging.102189
110. Ouyang X, Fan Q, Ling G et al (2020) Identification of diagnostic biomarkers and subtypes of liver hepatocellular carcinoma by multi-omics data analysis. Genes (Basel) 11(9):1–18. https://doi.org/10.3390/genes11091051
111. Peng C, Zheng Y, Huang DS (2020) Capsule network based modeling of multi-omics data for discovery of BRCA-related genes. IEEE/ACM Trans Comput Biol Bioinform 17(5):1605–1612. https://doi.org/10.1109/TCBB.2019.2909905
112. Lai YH, Chen WN, Hsu TC et al (2020) Overall survival prediction of non-small cell lung cancer by integrating microarray and clinical data with DL. Sci Rep 10(1):1–11. https://doi.org/10.1038/s41598-020-61588-w
113. Cui L, Li H, Hui W et al (2020) A DL-based framework for lung cancer survival analysis with biomarker interpretation. BMC Bioinform 21(1):1–14. https://doi.org/10.1186/s12859-020-3431-z
114. Mo W, Ding Y, Zhao S et al (2020) Identification of a 6-gene signature for the survival prediction of BRCA patients based on integrated multi-omics data analysis. PLoS ONE 15(11):1–18. https://doi.org/10.1371/journal.pone.0241924
115. Mo Q, Li R, Adeegbe DO et al (2020) Integrative multi-omics analysis of muscle-invasive bladder cancer identifies prognostic biomarkers for frontline chemotherapy and immunotherapy. Commun Biol 3(1):1–14. https://doi.org/10.1038/s42003-020-01491-2
116. Xu D, Wang Y, Liu X et al (2021) Development and clinical validation of a novel 9-gene prognostic model based on multi-omics in pancreatic adenocarcinoma. Pharmacol Res 164:105370. https://doi.org/10.1016/j.phrs.2020.105370
117. Chang Z, Miao X, Zhao W (2020) Identification of prognostic dosage-sensitive genes in colorectal cancer based on multi-omics. Front Genet 10:1–8. https://doi.org/10.3389/fgene.2019.01310
118. Yuan Y, Qi P, Xiang W et al (2020) Multi-omics analysis reveals novel subtypes and driver genes in glioblastoma. Front Genet 11:1–9. https://doi.org/10.3389/fgene.2020.565341
119. 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
120. Shang H, Liu ZP (2020) Network-based prioritization of cancer genes by integrative ranks from multi-omics data. Comput Biol Med 119:103692. https://doi.org/10.1016/j.compbiomed.2020.103692
121. Guan Y, Li T, Zhang H et al (2018) Prioritizing predictive biomarkers for gene essentiality in cancer cells with mRNA expression data and DNA copy number profile. Bioinformatics 34(23):3975–3982. https://doi.org/10.1093/bioinformatics/bty467
122. Yao Q, Xu Y, Yang H et al (2015) Global prioritization of disease candidate metabolites based on a multi-omics composite network. Sci Rep 5(1):1–14. https://doi.org/10.1038/srep17201
123. Fortino V, Kinaret P, Fyhrquist N et al (2014) A robust and accurate method for feature selection and prioritization from multiclass OMICs data. PLoS ONE 9(9):e107801. https://doi.org/10.1371/journal.pone.0107801
124. Fan H, Zhao H, Pang L et al (2015) Systematically prioritizing functional differentially methylated regions (fDMRs) by integrating multi-omics data in colorectal cancer. Sci Rep 5(1):1–16. https://doi.org/10.1038/srep12789
125. Chen Y, Wu X, Jiang R (2013) Integrating human omics data to prioritize candidate genes. BMC Med Genomics. https://doi.org/10.1186/1755-8794-6-57

126. Zhang T, Zhang D (2017) Integrating omics data and protein interaction networks to prioritize driver genes in cancer. Oncotarget 8(35):58050–58060. https://doi.org/10.18632/oncotarget.19481
127. Valdeolivas A, Tichit L, Navarro C et al (2019) Random walk with restart on multiplex and heterogeneous biological networks. Bioinformatics 35(3):497–505. https://doi.org/10.1093/bioinformatics/bty637
128. Wei PJ, Wu FX, Xia J et al (2020) Prioritizing cancer genes based on an improved random walk method. Front Genet 11:1–10. https://doi.org/10.3389/fgene.2020.00377
129. Zeng Z, Lu Y, Shen J et al (2019) A random interaction forest for prioritizing predictive biomarkers. arXiv. https://doi.org/10.48550/arXiv.1910.01786
130. Yang K, Lu K, Wu Y et al (2021) A network-based machine-learning framework to identify both functional modules and disease genes. Hum Genet. https://doi.org/10.1007/s00439-020-02253-0
131. Singh A, Shannon CP, Gautier B et al (2019) DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays. Bioinformatics 35(17):3055–3062. https://doi.org/10.1093/bioinformatics/bty1054
132. Kaur H, Dhall A, Kumar R, Raghava GPS (2020) Identification of platform-independent diagnostic biomarker panel for hepatocellular carcinoma using large-scale transcriptomics data. Front Genet 10:1–16. https://doi.org/10.3389/fgene.2019.01306
133. Kaur H, Bhalla S, Raghava GPS (2019) Classification of early and late stage liver hepatocellular carcinoma patients from their genomics and epigenomics profiles. PLoS ONE 14(9):e0221476. https://doi.org/10.1371/journal.pone.0221476
134. Gevaert O, Nabian M, Bakr S et al (2020) Imaging-AMARETTO: an imaging genomics software tool to interrogate multiomics networks for relevance to radiography and histopathology imaging biomarkers of clinical outcomes. JCO Clin Cancer Inform 4(4):421–435. https://doi.org/10.1200/cci.19.00125
135. 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
136. Abstract G, Torun FM, Virreira Winter S et al (2021) Transparent exploration of ML for biomarker discovery from proteomics and omics data. bioRxiv. https://doi.org/10.1101/2021.03.05.434053
137. Leclercq M, Vittrant B, Martin-Magniette ML et al (2019) Large-scale automatic feature selection for biomarker discovery in high-dimensional omics data. Front Genet 10:452. https://doi.org/10.3389/fgene.2019.00452
138. Song X, Ji J, Gleason KJ et al (2018) Insights into impact of DNA copy number alteration and methylation on the proteogenomic landscape of human ovarian cancer via a multi-omics integrative analysis. bioRxiv. https://doi.org/10.1101/488833
139. Ghannoum S, Netto WL, Fantini D et al (2021) Discbio: a user-friendly pipeline for biomarker discovery in single-cell transcriptomics. Int J Mol Sci 22(3):1–19. https://doi.org/10.3390/ijms22031399
140. Netanely D, Stern N, Laufer I, Shamir R (2019) PROMO: an interactive tool for analyzing clinically-labeled multi-omic cancer datasets. BMC Bioinform 20(1):1–10. https://doi.org/10.1186/s12859-019-3142-5
141. Tang Z, Kang B, Li C et al (2019) GEPIA2: an enhanced web server for large-scale expression profiling and interactive analysis. Nucleic Acids Res 47(W1):W556–W560. https://doi.org/10.1093/nar/gkz430
142. Wang Q, Zhang L, Yan Z et al (2019) OScc: an online survival analysis web server to evaluate the prognostic value of biomarkers in cervical cancer. Futur Oncol 15(32):3693–3699. https://doi.org/10.2217/fon-2019-0412
143. Champion M, Brennan K, Croonenborghs T et al (2018) Module analysis captures pancancer genetically and epigenetically deregulated cancer driver genes for smoking and antiviral response. EBioMedicine 27:156–166. https://doi.org/10.1016/j.ebiom.2017.11.028
144. Jang Y, Seo J, Jang I et al (2019) CaPSSA: visual evaluation of cancer biomarker genes for patient stratification and survival analysis using mutation and expression data. Bioinformatics 35(24):5341–5343. https://doi.org/10.1093/bioinformatics/btz516
145. Xie B, Yuan Z, Yang Y et al (2018) MOBCdb: a comprehensive database integrating multi-omics data on BRCA for precision medicine. BRCA Res Treat 169(3):625–632. https://doi.org/10.1007/s10549-018-4708-z
146. Mohammed A, Biegert G, Adamec J, Helikar T (2018) Cancer-Discover: an integrative pipeline for cancer biomarker and cancer class prediction from high-throughput sequencing data. Oncotarget 9(2):2565–2573. https://doi.org/10.18632/oncotarget.23511
147. Chong J, Soufan O, Li C et al (2018) MetaboAnalyst 4.0: towards more transparent and integrative metabolomics analysis. Nucleic Acids Res 46(W1):W486–W494. https://doi.org/10.1093/nar/gky310
148. Zeng D, Ye Z, Yu G et al (2020) IOBR: multi-omics immuno-oncology biological research to decode tumor microenvironment and signatures. bioRxiv. https://doi.org/10.1101/2020.12.14.422647
149. Liu CJ, Hu FF, Xia MX et al (2018) GSCALite: a web server for gene set cancer analysis. Bioinformatics 34(21):3771–3772. https://doi.org/10.1093/bioinformatics/bty411
150. Dong H, Wang Q, Zhang G et al (2020) OSdlbcl: an online consensus survival analysis web server based on gene expression profiles of diffuse large B-cell lymphoma. Cancer Med 9(5):1790–1797. https://doi.org/10.1002/cam4.2829
151. Gill S, Xu M, Ottaviani C et al (2022) AI for next generation computing: emerging trends and future directions. Internet Things 19:100514. https://doi.org/10.1016/j.iot.2022.100514

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.