frontiers | Frontiers in Genetics

TYPE Review
PUBLISHED 28 October 2025
DOI 10.3389/fgene.2025.1667325

OPEN ACCESS

EDITED BY
Valentina Silvestri,
Sapienza University of Rome, Italy

REVIEWED BY
Shixiang Wang,
Central South University, China
Asim Waqas,
Moffitt Cancer Center, United States

*CORRESPONDENCE
Caili Fang,
fangleheart@henu.edu.cn

RECEIVED 16 July 2025
ACCEPTED 16 October 2025
PUBLISHED 28 October 2025

CITATION
Wang J, Zhang J, Dai X, Yan C and Fang C (2025) Computational models for pan-cancer classification based on multi-omics data. *Front. Genet.* 16:1667325.
doi: 10.3389/fgene.2025.1667325

Computational models for pan-cancer classification based on multi-omics data

Jianlin Wang, Jiao Zhang, Xuebing Dai, Chaokun Yan and Caili Fang*

School of Computer and Information Engineering, Henan University, Kaifeng, Henan, China

Tumor heterogeneity presents a significant challenge in cancer treatment, limiting the ability of clinicians to achieve accurate early-stage diagnoses and develop customized therapeutic strategies. Early diagnosis is crucial for effective intervention, yet current methods lack robust solutions to overcome this challenge. The Pan-Cancer Atlas has emerged as a pivotal framework to investigate cancer heterogeneity by integrating multi-omics data (genomics, transcriptomics, proteomics) across tumor types. This initiative systematically maps inter- and intratumor variations, providing insight for clinical decision making. However, such frameworks often struggle to integrate dynamic temporal changes and spatial heterogeneity within tumors, limiting their real-time clinical applicability. In this review, we first summarize the available multi-omics data and public biomedical databases used in pan-cancer research. Then, we examine current pan-cancer classification approaches based on the computational models they employed, including machine learning and deep learning. We also provide a comparison of these classification methods to explore their advantages and limitations. Finally, we conclude by discussing the key challenges in pan-cancer research and suggesting potential directions for future studies.

KEYWORDS

pan-cancer classification, multi-omics data, deep learning algorithm, convolutional neural network, tumor heterogeneity

1 Background

Cancer, a heterogeneous group of diseases that affect various tissues and organs, constitutes a major global health burden. Despite advances in prevention, detection, and therapeutic interventions, global cancer incidence and mortality rates continue to increase (Santucci et al., 2020; Bray et al., 2024). A key limitation of current clinical practices is their reliance on molecularly insensitive tools, which often detect cancer only at intermediate or advanced stages, preventing early diagnosis (Wei et al., 2022). This delay is critical, as early detection is directly related to patient outcomes. For example, the 5-year survival rate for early-stage prostate cancer is 98%, and early breast cancer has a cure rate exceeding 95% (Siegel et al., 2020). However, tumor heterogeneity and similarity complicate early and accurate diagnosis, as well as treatment planning. Tumor heterogeneity manifests itself through genomic, transcriptomic, and proteomic differences between tumor cells, driving variations in morphology, proliferation, and metastatic potential (Zheng et al., 2022). Furthermore, even within the same tumor, cancer cells exhibit phenotypic and morphological heterogeneity during progression (Zhang et al., 2025). For example, lung cancer cells can differentiate into the subtypes of small cell lung cancer, lung squamous cell

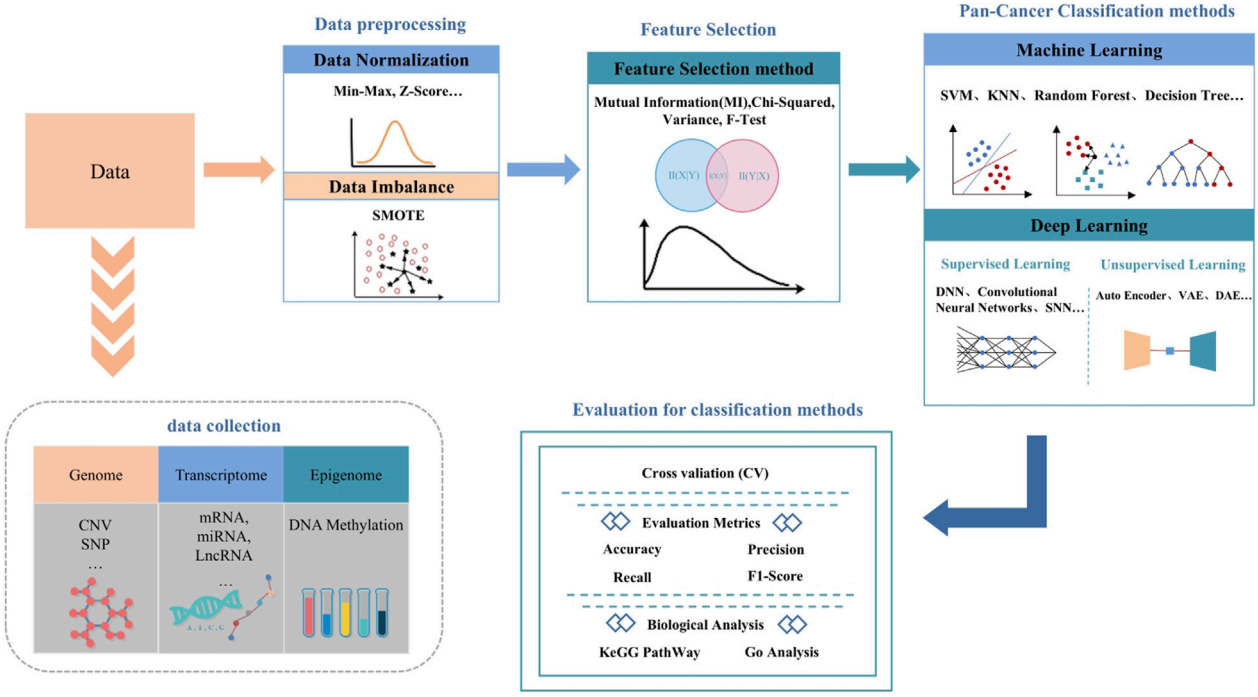

FIGURE 1
The workflow of pan-cancer classification model.

carcinoma, and lung adenocarcinoma (Yang and Fan, 2024). Each type and subtype of cancer has unique characteristics, leading to various clinical treatment approaches, and this heterogeneity poses significant challenges to diagnosis and treatment (Capper et al., 2018). The similarity of tumors is reflected in the finding that, at a molecular level, tumors in different parts of the body can be more similar than tumors of the same type (Sinha et al., 2021).

To address these challenges, The Cancer Genome Atlas (TCGA) launched the Pan-Cancer Project in 2012 (Weinstein et al., 2013), integrating omics data from more than 11,000 tumor samples to identify shared and unique oncogenic drivers. Pan-cancer aims to describe and identify the commonalities and differences between different types of cancer in order to find the key factors that may trigger cancer and thus guide clinical diagnosis, which is important to improve the cure rate of cancer. Many institutions have launched pan-cancer studies and developed public databases that collect data from various cancer-related researches. For example, the UCSC Genome Browser, that developed and maintained by the University of California, Santa Cruz (UCSC), is a comprehensive multi-omics database. Integrates various types of molecular data including copy number variations, methylation profiles, gene and protein expression levels, and mutation records. Furthermore, the platform supports efficient data analysis and visualization through user-friendly tools. The Gene Expression Omnibus (GEO), developed and maintained by the National Center for Biotechnology Information (NCBI), serves as a public repository for gene expression data. This database systematically integrates diverse cancer-related datasets, including high-throughput gene expression profiles and microarray data. Analysis of these pan-cancer datasets enables researchers to identify unique features of individual cancer types and explore shared or distinct molecular patterns across cancers. Such insights support the accurate classification of cancer subtypes and the development of targeted therapies. These research efforts form the foundation for the advancement of precision cancer and remain a central focus in contemporary cancer studies.

Traditional pan-cancer studies relied on cluster analysis, network modeling, and pathway enrichment to identify histological similarities. However, these methods lack the resolution required for early diagnosis. Rapid advancements in sequencing technologies have exponentially increased the scale and complexity of omics data, necessitating advanced computational approaches. Machine learning (ML) and deep learning (DL) methods now offer scalable solutions to analyze these high-dimensional datasets. For example, Li et al. (2017) achieved 90% precision in classifying 31 tumor types using genetic algorithms (GA) and K closest neighbors (KNN), while Lyu and Haque (2018) leveraged convolutional neural networks to classify 33 cancers with 95. 59% precision, identification of biomarkers via guided Grad-CAM. Overall, classification studies of pan-cancer datasets are important for improving the cure rate of cancer. Figure 1 shows the standardized workflow for pan-cancer classification models utilizing machine learning and deep learning frameworks.

Initially, researchers must collect and curate data from diverse publicly accessible biomedical databases relevant to the onset and progression of cancer. These data are critical for identifying oncogenic drivers underlying tumorigenesis. With advances in computer technology, a variety of feature dimension reduction and classification algorithms have been developed. These tools

are instrumental in constructing models that can accurately discriminate between different cancer types. Once developed, the performance of these methodologies should be assessed against state-of-the-art approaches. This involves comparing them across various metrics and prediction tasks using both standard and supplementary test datasets. Lastly, conducting relevant biological analyses and validations is vital to ensure the reliability and applicability of the findings.

Despite the existence of numerous classification methods for pan-cancer studies, there is a lack of comprehensive literature reviewing the data and methodologies employed. We addresses this gap by providing a thorough analysis of recent pan-cancer classification methods based on diverse models. We begin by exploring the data types commonly used in pancancer research and curating biomedical databases. This process improves our understanding of cancer heterogeneity and similarities and helps to validate research findings. We then examine prevalent classification approaches utilizing machine learning and deep learning models. Finally, we analyze standard datasets and evaluation metrics used in pan-cancer classification and provide a concise comparison of various methods. This comparison aims to assess the strengths and limitations of each approach.

2 Data and databases

2.1 Available data

With the conclusion of the Human Genome Project and the onset of the post-genomic era, innovative sequencing technologies have emerged (Waterman, 2021). Currently, gene microarray technology and transcriptome sequencing are the primary methods for acquiring cancer multi-omics data. Gene microarray technology, also called DNA microarray, detects both qualitative and quantitative information of DNA or RNA within a sample (Karakach et al., 2010). Transcriptome sequencing (RNA-Seq), also known as second-generation sequencing, offers greater accuracy and sensitivity in gene expression detection compared to microarray technology (Wang et al., 2009). Advancements in sequencing technologies have generated vast multi-omics datasets encompassing genomic, transcriptomic, and proteomic profiles. These multi-omics datasets serve as foundational resources for systematic exploration of oncogenic mechanisms across genomic, transcriptomic, and proteomic dimensions. Subsequently, we provide a detailed description of the multi-omics data closely related to pan-cancer research.

2.1.1 mRNA expression data

mRNA is a single-stranded RNA molecule that carries genetic information transcribed from DNA. It plays a crucial regulatory role in protein synthesis within the cell (Qin et al., 2022). mRNA expression data provide insights into gene function and activity. Investigating fluctuations in gene expression levels can elucidate disease development mechanisms. In cancer research, mRNA expression profiling has emerged as an essential element in elucidating cancer progression mechanisms. Studies show that dysregulation of specific genes can result in uncontrolled cell proliferation, a major factor in cancer development (Leibovitch and Topisirovic, 2018). For example, Li et al. (2017) used GA with a KNN classifier to classify mRNA data from 9,096 tumor samples of 31 types with 90% precision. Similarly, Kim et al. (2020) identified key genes that accurately distinguish 21 types of tumors by using ANOVA tests on mRNA data from cancer and normal samples. Therefore, studying mRNA expression data to find oncogenes helps in early cancer diagnosis and more accurate classification, improving treatment.

2.1.2 miRNA expression data

miRNAs are small noncoding RNAs present in plants and animals, typically 20 to 24 nucleotides long. They play a critical role in the regulation of cellular processes (Cui et al., 2025). miRNA controls oncogenes and tumor suppressor gene expression by degrading mRNAs or inhibiting their translation (Tang et al., 2021; Galagali, 2020). For example, in non-small cell lung cancer, high let-7 expression reduced lung cancer cell growth and inhibited differentiation (Pop-Bica et al., 2020). In gastric cancer, certain miRNAs inhibit the expression of the phosphatase and tensin homolog (PTEN) gene and promote cancer cell growth and invasion (Ashrafizadeh et al., 2020). Wang et al. (2019) combined GA with random forest (RF) for pan-cancer classification of miRNA data from 32 tumor types, achieving 92% sensitivity. To more robust and reliable set of miRNA features capable of distinguishing different types of tumor, Lopez-Rincon et al. (2019). developed an integrated feature selection algorithm for an accfor ante classification of 28 types otypes of tumorsth reliable miRNA features. Therefore, studying miRNA functions is vital for accurate cancer classification and early diagnosis, significantly impacting treatment and prognosis.

2.1.3 lncRNA expression data

lncRNAs are RNA molecules with transcript sequences of more than 200 nucleotides. Although they do not encode proteins, they regulate biological processes such as gene expression, development, and differentiation (Chen et al., 2021). Initially considered as genomic noise, lncRNAs are now recognized as important in cancer development. Changes in their expression can serve as diagnostic markers (Nandwani et al., 2021; Fang and Fullwood, 2016). Analyzing lncRNA data has identified potential biomarkers and distinguished between tumor types (Al Mamun and Mondal, 2019a; Al Mamun and Mondal, 2019b; Al Mamun et al., 2020). Therefore, understanding the roles of lncRNAs is crucial for early cancer diagnosis and treatment.

2.1.4 Copy number variation (CNV)

CNV refers to the variation in the number of copies of a particular gene present in an individual's genome (Pös et al., 2021). Genes such as BRCA1, CHEK2, ATM, and BRCA2 have strong associations with cancers like breast cancer (Hu et al., 2018). Zhang et al. (2016) proposed using a Dagging classifier to categorize CNV data from six cancer types, highlighting key features for accurate classification. Therefore, studying CNV helps explore cancer pathogenesis, aiding early diagnosis and treatment selection.

2.1.5 DNA methylation

DNA methylation, an epigenetic modification, involves adding a methyl group to DNA, usually suppressing gene expression (Liu

TABLE 1 Description of common data types of pan-cancer. The dimensions presented are the feature counts derived from the TCGA Pan-Cancer Atlas dataset.

Data type	Data description	Dimension
mRNA	The real-time product of gene expression, which controls protein synthesis, and abnormal expression can lead to the development of cancer	20,531
miRNA	Key molecules in the regulation of transcription and translation of oncogenes and tumor suppressor genes. Aberrant expression regulates tumor cell growth, proliferation and apoptosis	1,882
lncRNA	An RNA molecule that does not have protein-coding ability and is involved in the development of cancer, and changes in its expression level can be used as a marker for the diagnosis of cancer	19,166
CNV	Caused by genomic rearrangements, occurring in genes 1-kb or longer in length that are implicated in the development and progression of human cancers	24,174
DNA Methylation	DNA methylation usually inhibits the expression of genes in cells and plays an important regulatory role, and abnormal expression silences tumor suppressor genes leading to the development of cancer	48,578

TABLE 2 Overview of the cancer database.

Database	Brief description	Links
TCGA (Weinstein et al., 2013)	Collected multiple omics data of 33 tumor types, the largest human tumor sequencing database in the world	https://www.cancergenome.nih.gov/
EGA (Lappalainen et al., 2015)	Collection of over 800 medical studies of all types of sequencing data and typing data	https://ega-archive.org/
CGHub (Wilks et al., 2014)	Sequencing data of 25 different types of cancers from TCGA, TARGET, and CCLE were collected and organized	https://cghub.ucsc.edu/
ICGC (Consortium et al., 2010)	Collecting omics data from many different types of cancers, and comprehensively described the genomic changes of many cancer	https://dcc.icgc.org/
COSMIC (Forbes et al., 2015)	Collecting omics data on many types of cancer, it is the world's largest and most comprehensive database of somatic mutations	https://cancer.sanger.ac.uk/cosmic
cBioPortal (Gao et al., 2013)	Collects genomic data on many different types of cancer, providing visual analysis tools across genes, samples and data types	http://www.cbioportal.org/
UCSC Xena (Navarro Gonzalez et al., 2021)	Collecting data from several large cancer research projects, and provides convenient data analysis and visualization capabilities	http://genome.ucsc.edu/
arrayMap (Cai et al., 2015)	Provide pre-processed tumor genome microarray data and CNA atlas	http://www.arraymap.org/
BioMuta (Wu et al., 2014)	26 different types of cancers were collected SNV-data	https://hive.biochemistry.gwu.edu/home/
GEO (Barrett et al., 2012)	Collection and organization of high-throughput gene expression data submitted by research institutions around the world	https://www.ncbi.nlm.nih.gov/geo/
ArrayExpress (Kolesnikov et al., 2015)	Collected and organized microarray chip-based and high-throughput sequencing of experimental genomics data	https://www.ebi.ac.uk/arrayexpress/
OncomiRDB (Wang et al., 2014)	Collection and annotation of experimentally validated miRNAs with promotive or inhibitory effects on different cancer types	http://www.oncomir.org/
miRCancer (Xie et al., 2013)	A comprehensive collection of miRNA expression profiles in various human cancers	http://mircancer.ecu.edu/
SomaMiR (Bhattacharya et al., 2013)	Collecting data on miRNAs and mutations on their targets	https://compbio.uthsc.edu/SomaMiR/
ChiTaRS (Frenkel-Morgenstern et al., 2015)	Cancer genome sequence breakpoints were collected along with expression level data of the corresponding chimeric transcripts	https://chitars.bioinfo.cnio.es/
MethylCancer (He et al., 2007)	Collected tumor DNA methylation, cancer-related genes, mutations, CpG islands, and cancer information	http://methylcancer.psych.ac.cn/
MethHC (Huang et al., 2015)	Organized DNA methylation, mRNA/miRNA gene expression, miRNA methylation, and association between methylation and gene expression levels from TCGA	http://methhc.mbc.nctu.edu.tw/
CGC (Subramanian et al., 2021)	NCI-funded cloud platform co-localizing large datasets, and compute power for secure, collaborative multi-omics analysis	https://www.cancergenomicscloud.org/
CPTAC (Mesri et al., 2024)	CPTAC provides a rich source of public data, serving as a critical resource for researchers studying pan-cancer proteomics	https://cptac-data-portal.georgetown.edu/cptac/

et al., 2016). It is crucial for normal cellular functions and implicated in cell differentiation and tumorigenesis. Dysregulated methylation, such as hypermethylation of CpG islands in promoter regions, can silence tumor suppressor genes or reduce oncogenic miRNA transcription, increasing cancer risk (Formosa et al., 2013). Liu et al. (Liu et al., 2019) used methylation data from 27 cancers types and proposed machine learning and deep learning strategies for accurate cancer differentiation. Therefore, DNA methylation is closely related to the occurrence and development of cancer, and the analysis and study of methylation is very important in the field of cancer diagnosis.

2.1.6 Multi-omics

The development of cancer is a very complex process that is not simply caused by the occurrence of abnormalities in one type of data, but often involves multiple histological pathological processes. Therefore, data mining analysis based on single omic data has certain one-sidedness and limitations. In recent years, with the rapid development of next-generation genomic technologies, a large amount of genomic data of different types of cancers has been accumulated, and more and more researchers have started to integrate multiple omic data to conduct systematic and complete analysis of the mechanisms of cancer occurrence, and cancer research is developing from single omic to multi-omics. Integrated multi-omics analysis can make up for the lack of information in single-omics data and provide a comprehensive view of patients, and enable researchers to explore the relationship between cancer and genes from multiple perspectives, so as to perform early cancer diagnosis more accurately.

Table 1 summarizes the characteristics of common pan-cancer data types, including mRNA, miRNA, and DNA methylation.

2.2 Biomedical database

With the rapid development of high-throughput sequencing technology, a large amount of tumor-related histological data has been accumulated, and meanwhile, various public medical databases have emerged continuously. These public databases can be classified into comprehensive databases, genomic, transcriptomic, epigenomic databases, etc. according to the research areas or data types. Table 2 summarizes some cancer-related databases and provides brief descriptions and access links.

Next, we provide a detailed description of the most commonly used databases in pan-cancer research.

2.2.1 TCGA

TCGA is the largest human tumor genome sequencing database globally (Weinstein et al., 2013). Jointly sponsored by the National Human Genome Research Institute (NHGRI) and the National Cancer Institute (NCI), this major research project was officially launched in 2005. TCGA has sequenced 33 common cancers and over 11,000 tumor samples, using genomic analysis technology to enhance understanding of tumor mechanisms and improve cancer diagnosis and treatment capabilities (Tomczak et al., 2015). TCGA currently provides mRNA expression data, miRNA expression data, DNA methylation data, CNV data, and other high-throughput sequencing data. Researchers can access these datasets through the Genomic Data Commons (GDC) Data Portal, the primary data source for many cancer researchers.

2.2.2 GEO

GEO is a subdatabase of the National Center for Biotechnology Information (NCBI). This free and publicly accessible repository houses biological data from gene chips, second-generation sequencing, and other high-throughput functional genomics experiments. It includes submissions from over 16,000 laboratories and research teams worldwide, featuring 175,825 datasets with 5,069,606 data samples. GEO supports data download capabilities, enabling users to obtain samples or datasets of interest. Additionally, it offers tools to discover genes of interest and their expression profiles, as well as to identify genes with similar expression patterns.

2.2.3 UCSC Xena

UCSC Xena is a cancer genomics data analysis platform developed by the UCSC Cancer Genome Browser (Navarro Gonzalez et al., 2021). This platform collects and standardizes data from several major cancer research projects such as TCGA, ICGC, and TARGET, facilitating subsequent analysis (Consortium et al., 2010). UCSC Xena encompasses multiple levels of data, including copy number, methylation, gene expression, protein expression, and mutation data. It provides user-friendly data analysis and visualization tools. Researchers can easily analyze or download organized data with link clicks and can also upload their data for analysis. This flexibility considerably aids in the advancement of genomic research.

2.2.4 CPTAC

The Clinical Proteomic Tumor Analysis Consortium (CPTAC) is a comprehensive proteomic and genomic research program initiated by the National Cancer Institute (NCI) that aims to accelerate the understanding of cancer biology through the integration of large-scale proteomic and genomic analysis (Mesri et al., 2024). The consortium systematically identifies, quantifies, and analyzes proteins from cancer biospecimens characterized by genomic data to improve cancer prevention, early diagnosis, treatment, and prognosis. CPTAC provides a rich source of public data, serving as a critical resource for researchers studying pan-cancer proteomics. Its data, which includes protein abundance, post-translational modifications, and mass spectrometry data, is often used in combination with genomic data to provide a multi-layered view of tumors, enabling the discovery of new biomarkers and therapeutic targets.

2.2.5 CGC

The Cancer Genomics Cloud (CGC), an NCI-funded resource powered by Seven Bridges, is a secure and scalable cloud-based platform designed to overcome the challenges associated with accessing, sharing, and analyzing massive, diverse multi-omics datasets (Subramanian et al., 2021). The platform achieves this by co-localizing three essential components within the cloud: major cancer datasets like The Cancer Genome Atlas (TCGA) and Clinical Proteomic Tumor Analysis Consortium (CPTAC); over 400 bioinformatics tools and best-practice workflows; and

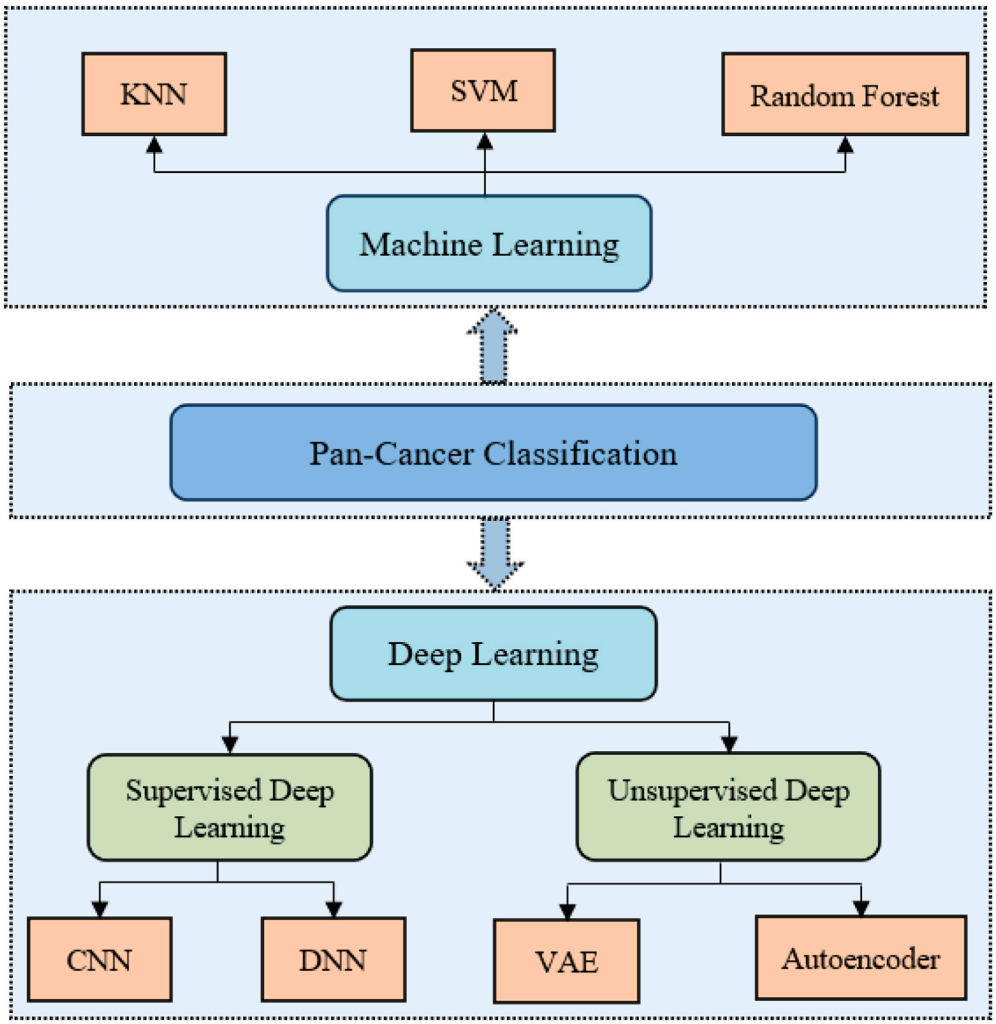

FIGURE 2
Pan-Cancer classification methods based on various models.

the high-performance computational capabilities for large-scale analysis. The CGC simplifies the user experience by enabling researchers to browse, query, and filter datasets, run their entire analysis workflow on the platform, and even integrate their own private tools and data.

Building on the data sources described above, the following section reviews computational methods for pan-cancer classification.

3 Methods

Advances in biotechnology have significantly expanded the application of gene sequencing in pan-cancer studies. The proliferation of high-throughput sequencing data offers a critical foundation for research. However, a key challenge lies in developing efficient computational algorithms to extract biologically meaningful insights from these complex datasets. Current methodologies for pan-cancer analysis are broadly categorized into two frameworks: classical machine learning and deep learning. As illustrated in Figure 2 deep learning models can be further subdivided into supervised and unsupervised approaches, depending on the utilization of labeled data.

3.1 Pan-cancer classification model based on machine learning

Feature selection innovations and model optimization strategies in machine learning have significantly advanced pan-cancer classification accuracy. To balance feature relevance and parsimony, Kim et al. (2020) implemented a two-stage gene selection strategy: ANOVA-based F-statistic ranking identified top genes across 21 cancers, followed by frequency-based filtering. Neural networks trained on 300 selected genes achieved peak accuracy (90%), outperforming other classifiers. Mahin et al. (2022) refined this approach by retaining only genes consistently expressed across all 21 cancers and incorporating data smoothing/oversampling, enhancing model robustness. Luo et al. (2023) developed an ML approach to predict cancer prognosis

considering 32 cancer types from TCGA.Initially, the approach was applied to hepatocellular carcinoma and then extended to other types of tumors.

Beyond conventional methods, researchers have explored hybrid and multi-algorithm frameworks. Khadirnaikar et al. (2023) analyzed mRNA, miRNA, DNA methylation, and protein of 33 different types of cancer from TCGA. Firstly, multi-omics data was combined by concatenating the features for each sample, and then the autoencoder was used to reduce the dimension of data. Novel subtypes of cancer samples were identified by clustering k-means. Further exploring the efficacy of the classifier, Elsadek et al. (2019) employed a machine learning approach using gene CNV data across six types of tumor. Their approach utilized an information gain algorithm for gene selection and evaluated various classifiers, with LR achieving superior performance, underscoring machine learning's role in cancer classification. Liu (2022) analyzed the association with a correlation test of epi-driver CpG sites between DNA methylation and gene expression profiles. XGBoost and SHAP algorithms identified the best biomarkers in five genes and used them as features for the generation of a random forest model to identify cancer subtypes. Finally, Cheerla and Gevaert (2017) and Al Mamun and Mondal (2019a) both explored two-stage feature selection approaches. Cheerla's team reduced miRNA features using correlation and recursive elimination, achieving the best classification with SVM radial among 21 tumor types. Mamun's approach selected common features for classifiers, finding SVM provided the best accuracy for eight different cancers. Collectively, these innovations underscore machine learning's adaptability in addressing omics complexity while balancing feature parsimony and accuracy.

3.2 Pan-cancer classification model based on deep learning

Although machine learning methods have been widely used to study pan-cancer classification problems and achieved good results, with the development of deep learning and the high performance shown on classification tasks, more and more researchers have started to use deep learning to improve the performance of tumor classification models. In the field of deep learning, deep learning methods can be classified into two categories based on whether the models use the labels of the data, namely, supervised learning and unsupervised learning (Alzubaidi et al., 2021).

3.2.1 Supervised classification models

Recent advancements in supervised deep learning have demonstrated remarkable efficacy in pan-cancer classification through tailored architectural innovations. Sun et al., 2018) introduced GeneCT, an artificial neural network (ANN) framework designed to classify 11 tumor types using raw mRNA expression data without feature engineering, achieving 98.2% accuracy and underscoring the potential of end-to-end learning in omics analysis. Complementing this approach (Cava et al., 2023), applied principal component analysis (PCA) to reduce data dimensionality before deploying the model. The neural network achieved a mean accuracy of 84%, the random forest reached 86%, and XGBoost achieved the highest performance with a mean accuracy of 90%. To address the challenges of limited sample sizes in specific cancer types (Cho et al., 2023), proposed a meta-learning method that integrates multi-omics data (transcriptomics, proteomics, and clinical data from TCGA) to create predictive models using survival information from 17 cancer types. Their approach requires fewer samples than conventional deep learning models, effectively mitigating data scarcity issues. Expanding this paradigm (Divate et al., 2022), employed deep neural networks (DNNs) to classify 33 cancer types. Their methodology integrated expression-based gene screening with SHAP (Shapley Additive exPlanations) interpretability, identifying critical biomarkers and achieving superior performance in distinguishing cancers from healthy controls.

To address high-dimensional data challenges (Wu et al., 2024) developed DeepMoIC, a method combining deep graph convolutional networks (GCNs) with autoencoders for cancer subtype classification. By constructing a patient similarity network (PSN) and leveraging GCNs, DeepMoIC outperformed existing models on multi-omics datasets, highlighting its potential for precision oncology. (Li et al., 2025) introduced DGHNN, a deep graph and hypergraph neural network for pan-cancer related gene prediction that takes biological pathways into consideration. This method applies a deep graph and hypergraph neural network to encode higher-order information in protein interaction networks and biological pathways. This approach, along with the introduction of skip residual connections and a feature tokenizer with a transformer for classification, demonstrates how advanced network architectures can capture the multi-level complexity of biological systems, setting a new standard for performance. (Li et al., 2020) tackled CNV sparsity by coupling Monte Carlo feature selection (MCFS), which evaluates feature stability via randomized sampling, with self-normalizing neural networks (SNNs) to enhance training robustness. Their framework achieved 79.8% accuracy in classifying four cancer types. These studies collectively highlight how supervised architectures can be customized to diverse omics modalities while balancing performance and biological interpretability.

In recent years, due to the excellent performance of convolutional neural networks (CNNs) on image classification tasks, more and more researchers have started to apply these networks to the classification problem of pan-cancer. For instance (Ameen et al., 2025) proposed a stacked deep learning ensemble model for multi-omics cancer type classification, demonstrating that deep learning can be effectively applied to high-dimensional biological data. Similarly (Lyu and Haque, 2018), firstly proposed the use of a convolutional neural network to classify mRNA expression data by embedding high-dimensional gene expression data into a two-dimensional image as the input of the convolutional neural network to classify 33 different types of tumors. Building on this, Mostavi et al. (Mostavi et al., 2020) systematically compared CNN architectures (e.g., Inception modules, residual connections), revealing that deeper networks achieved 95. 82% precision on 33-class tasks that highlight the impact of structural optimization. Addressing computational inefficiency Khalifa et al., 2020), applied binary particle swarm optimization (BPSO) to reduce the dimensionality of mRNA from 20,531 to 512 features before CNN training, achieving accuracy of 96. 9% on five types of tumors. Hybrid models also

emerged as a promising frontier: (Huynh et al., 2019) combined deep CNNs (DCNN) with SVM classifiers, where DCNNs extracted high-order features and SVMs performed classification, reaching 76. 33% precision for 25 cancers. (Abdullahi et al., 2020) further demonstrated the efficiency of fine-tuning pre-trained AlexNet models on mRNA data, reaching 98.1% accuracy for five cancers with minimal computational overhead. Beyond expression data (Ye et al., 2021) encoded somatic mutation profiles into heatmap-like "mutation maps," enabling ResNet-50 and Inception-v3 models to outperform traditional methods (89.7% vs. SVM's 72.3%). Finally (AlShibli and Mathkour, 2019) validated CNNs' versatility in CNV analysis, showing that a six-layer residual network (ResCNN6) surpassed standard CNNs and VGG-16 (86% accuracy for six cancers), underscoring the efficacy of residual connections in combating gradient vanishing. These innovations exemplify CNNs' adaptability to multi-omics integration through data transformation, architectural refinement, and cross-domain transfer learning.

3.2.2 Unsupervised classification models

Unsupervised deep learning techniques have emerged as powerful tools for pan-cancer classification, particularly in scenarios with limited labeled data. Rong et al. (Rong et al., 2022) proposed a computational approach, multi-omics clustering variational autoencoders (Mcluster-VAEs), based on a new probabilistic model of a deep learning method consisting of clustering algorithm for multi-omics data to estimate posterior cancer subtypes. Building on this (Al Mamun et al., 2020) introduced the Concrete Autoencoder (CAE), an unsupervised framework for identifying discriminative lncRNAs. The CAE outperformed supervised methods (Lasso, RF, SVM-RFE) in classifying 33 tumors, achieving 93% accuracy. To address feature instability across CAE iterations (Al Mamun et al., 2021) later proposed the multi-run CAE (mrCAE), which aggregated high-frequency lncRNAs from 100 CAE runs to derive a stable subset of 69 markers. This refined set enabled accurate classification of 12 cancers, resolving reproducibility challenges inherent to stochastic deep learning models. Expanding to multi-omics integration (Zhang et al., 2019) developed OmiVAE, an end-to-end model combining VAEs with a classification network. OmiVAE first compressed the mRNA and DNA methylation data into low-dimensional embeddings, then predicted 33 tumor types using a three-layer neural network, achieving precision of 97. 49%. Finally (Albaradei et al., 2021) designed MetaCancer, which used convolutional VAE to extract features from mRNA, miRNA and methylation data. When fed into a deep neural network (DNN), this multi-omics integration classified 11 cancers with 88.85% accuracy-surpassing mRNA-only approaches by 14.2%. (Li et al., 2024) proposed AVBAE-MODFR, a two-phase framework that combines adversarial variational Bayes autoencoder for multi-omics embedding with a dual-net feature ranking module. Tested on TCGA pan-cancer data, AVBAE-MODFR outperformed four state-of-the-art methods, highlighting its robustness in representation learning and biomarker discovery. Compared with earlier VAE-based models such as OmiVAE and MetaCancer, AVBAE-MODFR not only integrates heterogeneous omics but also incorporates an explicit feature ranking mechanism, thereby enhancing interpretability and facilitating the identification of biologically meaningful markers. These innovations underscore unsupervised learning's potential to uncover robust biomarkers and integrate heterogeneous omics data without reliance on labeled datasets.

Figure 3 illustrates the growing prominence of deep learning in pan-cancer research. It shows the percentage of all pan-cancer-related articles that used deep learning methods for classification over the past few years. A systematic review of papers published on the PubMed and Web of Science platforms using search terms "pan-cancer classification", "deep learning" and "machine learning" from 2018-2024 revealed a steady increase in this ratio from 2018 to 2024. To summarize the current landscape of pan-cancer classification, we present an overview of relevant studies in recent years in Table 3. This table highlights the variety of machine learning and deep learning approaches, as well as the multi-omics data they employ.

3.3 Integration strategies

The integration of multi-omics data is a critical step in pan-cancer research, as it provides a more comprehensive view of cancer's molecular mechanisms by combining information from multiple platforms. Integration strategies are typically categorized by the stage at which the data is combined. For instance, an early integration approach, where mRNA and CNV data are simply concatenated, may be easy to implement but can lead to a high-dimensional feature space and potentially introduce noise (Zhao et al., 2024). In contrast, an intermediate integration approach using a variational autoencoder (VAE) to create a joint latent space can handle the high dimensionality and may reveal more complex, underlying relationships between omics types, but the learned features are often less interpretable.

To better evaluate the performance of these pan-cancer classification models, researchers are developing new benchmarks. These include integrating multi-omics data from large consortia, assessing cross-cohort generalization, and shifting the focus to more specific clinical endpoints beyond simple cancer type classification. For example, integrating genomics from TCGA with proteomics from CPTAC offers a more comprehensive understanding of cancer's molecular mechanisms, as proteins are the functional molecules that execute cellular processes. A related large-scale multi-omics benchmark, CMOB, integrates data from the TCGA platform, providing an accessible and usable resource for machine learning research (Yang et al., 2024).

Beyond these comprehensive datasets, evaluating a model's generalization ability across different patient cohorts is essential for validating its robustness and reliability in diverse clinical settings. In addition, new benchmarks are moving beyond the simple classification of cancer types to include more refined clinical endpoints such as subtype classification, stage prediction, survival analysis, and prediction of response to treatment. These more granular predictions are crucial for personalized medicine, as they inform specific patient care strategies. Several recent case studies highlight these advances. AVBAE-MODFR is a deep learning framework that integrates multi-omics data through embedding and feature selection, showing potential clinical applications in tumor diagnosis and precision medicine (Li et al., 2024). TMO-Net is another model that is pre-trained on multi-

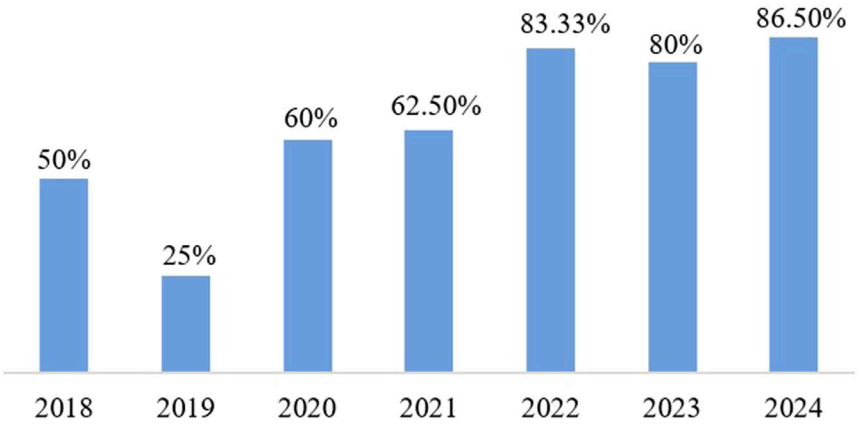

FIGURE 3
The ratio of pan-cancer research using deep learning technologies.A systematic review of the relevant literature shows a steady increase in the use of deep learning in pan-cancer research in recent years.

omics pan-cancer datasets to facilitate cross-omics interactions and enable joint representation learning and inference on incomplete omics data, thereby supporting various downstream oncology tasks (Wang et al., 2024).

Future research is also expanding to incorporate new data types and modalities that offer a more holistic view of tumor biology. Single-cell multi-omics (e.g., scRNA-seq, scATAC-seq) provides an unprecedented resolution of tumor heterogeneity at the cellular level, capturing differences between individual cells that are lost in bulk omics data. In addition, integration of radiology and pathology images with molecular data is a promising area. This represents a different data modality (unstructured images) that requires specialized models such as CNNs. Combining these visual cues with molecular data can provide a more comprehensive view of the tumor, bridging the gap between molecular mechanisms and the morphological features observed in clinical practice.

4 Evaluation and discussion

4.1 Selection criteria

We systematically reviewed papers published on the Ovid and Web of Science platforms. Our search criteria focused on machine learning and multi-omics data for pan-cancer studies. We only included full-text, English-language papers from peer-reviewed journals that used artificial intelligence to analyze multi-omics data on cancer samples. We excluded any papers that only applied machine learning to a single cancer type or data type, did not use cancer samples, or were themselves reviews or proceedings.

4.2 Classification evaluation metrics

Classification performance evaluation metrics are essential to objectively assess the effectiveness of classification models. Selecting a high-performing classifier relies on using rigorous evaluation criteria. Accuracy is a common metric for evaluating overall model performance in classification tasks. However, in pan-cancer classification, sample size imbalance is a prevalent issue, where some cancer types have many samples while others have few. In such cases, the majority class can disproportionately influence overall accuracy, diminishing its evaluative significance. For example, a model trained on an imbalanced dataset might achieve a deceptively high accuracy simply by correctly classifying all samples from the majority class, while failing to identify samples from the rarer, minority classes. Thus, relying solely on accuracy is insufficient.

Therefore, it is necessary to consider other metrics that provide a more complete picture of a model's performance on multi-label, imbalanced datasets. We analyze several evaluation metrics as reported in the reviewed literature, including Precision (PR), Recall (RC), F1-score, Area Under the Receiver Operating Characteristic Curve (AUC), and Matthews Correlation Coefficient (MCC). Precision measures the proportion of true positive predictions among all positive predictions, while recall measures the proportion of true positives correctly identified from all actual positives. The F1-score provides a single value that balances both precision and recall, making it particularly useful for evaluating models on imbalanced data. The AUC and MCC are also important for assessing overall performance, with MCC providing a balanced measure that accounts for all four values in a confusion matrix, regardless of class size.

4.3 Data sets

For pan-cancer classification research, multiple of the following 33 cancer types are commonly used for analysis. The types and sample information of these 33 cancers are shown in Table 4.

Next, the analysis performed in terms of datasets employed by the distinct research works is elaborated. Figure 4 depicts several datasets utilized for pan-cancer classification. BRCA is the most frequently utilized dataset in pan-cancer classification research. In addition, the most commonly used datasets in pan-

TABLE 3 Overview of pan-cancer classification methods.

References	Method type(s)	Data type(s)	Data source	Cancer types	Code link
Kim et al. (2020)	ML (SVM,KNN)	mRNA	–	21	–
Mahin et al. (2022)	ML (KNN)	mRNA	TCGA	22	https://github.com/Zwei-inc/panclassif
Luo et al. (2023)	ML (SVM)	Gene expression	TCGA	32	–
Khadirmaikar et al. (2023)	ML (SVM)	mRNA, miRNA,DNA Methylation	GDC	33	https://github.com/seemark11/Pancancer-subgroup-identification
Cheerla and Gevaert (2017)	ML (SVM)	miRNA	–	21	–
Al Mamun and Mondal, (2019a)	ML	lncRNA	–	8	–
Elsadek et al. (2019)	ML (SVM,Random forest)	CNV	–	6	–
Liu (2022)	ML (Random forest)	DNA Methylation, Gene expression	–	11	–
Sun et al. (2018)	SDL	mRNA	–	11	http://sunlab.cpy.cuhk.edu.hk/GeneCT/
Cava et al. (2023)	SDL (Neural Network)	Gene expression	–	16	https://github.com/claudiacava/Applied-Sciences
Cho et al. (2023)	SDL (Neural Network)	Gene expression	TCGA	17	https://github.com/berkuva/TCGA-omics-integration
Mostavi et al. (2020)	SDL (CNN)	mRNA	–	33	https://github.com/chenlabgccri/CancerTypePrediction
Khalifa et al. (2020)	SDL (CNN)	mRNA	–	5	–
Huynh et al. (2019)	SDL (DCNN)	mRNA	–	25	–
Abdullahi et al. (2020)	SDL (CNN)	mRNA	–	5	–
Li et al. (2020)	SDL (SNN)	CNV	–	4	https://github.com/KohTseh/CancerClassification
Rong et al. (2022)	DL	miRNA,DNA methylation,CNV	UCSC	32	https://github.com/luyiyun/MCluster-VAEs
Albaradei et al. (2021)	UDL (CVAE)	mRNA, miRNA,DNA Methylation	–	11	https://github.com/SomayahAlbaradei/MetaCancer
Al Mamun et al. (2020)	UDL (CAE)	lncRNA	–	33	–
Al Mamun et al. (2021)	UDL (CAE)	lncRNA	–	12	–
Zhang et al. (2019)	UDL (VAE)	mRNA,DNA Methylation	UCSC	33	https://github.com/zhangxiaoyu11/OmiVAE
Li et al. (2024)	UDL (VAE/CVAE)	mRNA, miRNA,DNA Methylation	–	33	https://github.com/zhanglabNKU/AVBAE-MODFR

ML: machine learning; SDL: supervised deep learning; UDL: unsupervised deep learning; CNN: convolutional neural network; SNN: self-normalizing neural network; CVAE: convolutional variational autoencoder; CAE: concrete autoencoder; VAE: variational autoencoder.

cancer classification also include KIRC, LUAD, COAD, KIRP, LIHC, etc.

4.4 Comparison and analysis

As reported in the reviewed literature, a performance comparison of various pan-cancer classification methods on the mRNA gene expression dataset for 33 cancer types reveals that deep learning models generally achieve higher classification accuracies than traditional machine learning methods. For instance (Lyu and Haque, 2018) reported a 95.59% accuracy using a convolutional neural network, a performance that surpasses many of the reported accuracies of traditional machine learning algorithms on similar tasks. This qualitative comparison of architectures suggests that deep learning models are often more capable of distinguishing between 33 different cancer types due to their ability to learn complex, hierarchical features from high-dimensional data.

Next, the classifiers used in different research works are elaborated and analyzed. Figure 5 illustrates several common classifiers utilized for pan-cancer classification. This figure was generated by counting the primary classifiers used in the reviewed articles. A classifier was

TABLE 4 Types of cancer and number of samples.

No.	Cancer name	Code	Cases
1	Adeno-cortical carcinoma	ACC	79
2	Bladder-Urothelial-Carcinoma	BLCA	408
3	Breast-invasive carcinoma	BRCA	1093
4	Cervical and endocervical cancers	CESC	304
5	Cholangiocarcinoma	CHOL	36
6	Colon-adenocarcinoma	COAD	457
7	Lymphoid-Neoplasm-Diffuse-Large B-cell-Lymphoma	DLBCL	48
8	Esophageal carcinoma	ESCA	184
9	Glioblastoma multiforme	GBM	160
10	Head and Neck squamous cell carcinoma	HNSC	520
11	Kidney-Chromophobe	KICH	66
12	Kidney renal clear cell carcinoma	KIRC	533
13	Kidney renal papillary cell carcinoma	KIRP	290
14	Acute-Myeloid Leukemia	LAML	179
15	Brain Lower-Grade Glioma	LGG	516
16	Liver-hepatocellular carcinoma	LIHC	371
17	Lung adenocarcinoma	LUAD	515
18	Lung squamous cell carcinoma	LUSC	501
19	Mesothelioma	MESO	87
20	Ovarian serous cystadenocarcinoma	OV	304
21	Pancreatic adenocarcinoma	PAAD	178
22	Pheochromocytoma and Paraganglioma	PCPG	179
23	Prostate-adenocarcinoma	PRAD	497
24	Rectum-adenocarcinoma	READ	166
25	Sarcoma	SARC	259
26	Skin Cutaneous Melanoma	SKCM	469
27	Stomach adenocarcinoma	STAD	415
28	Testicular Germ Cell Tumors	TGCT	150
29	Thyroid carcinoma	THCA	501
30	Thymoma	THYM	120
31	Uterine Corpus Endometrial Carcinoma	UCEC	545
32	Uterine Carcinosarcoma	UCS	57
33	Uveal Melanoma	UVM	80

counted if it was the main model for the classification task. The raw counts were then converted to percentages to show the proportion of each classifier type. As shown in the figure, the most frequently used machine learning classifiers in pan-cancer classification studies are SVM, RF, ANN, and KNN, respectively. Meanwhile, among deep learning classifiers, CNNs and fully connected deep neural networks (DNNs, e.g., multilayer perceptrons) were the most frequently used.

4.5 Discussion

In our review, we have summarized the diverse ML and DL algorithms applied to pan-cancer multi-omics analysis. In many cases, proposed methods were evaluated against existing algorithms, often showing comparable levels of performance. However, no systematic comparison of different approaches on a common

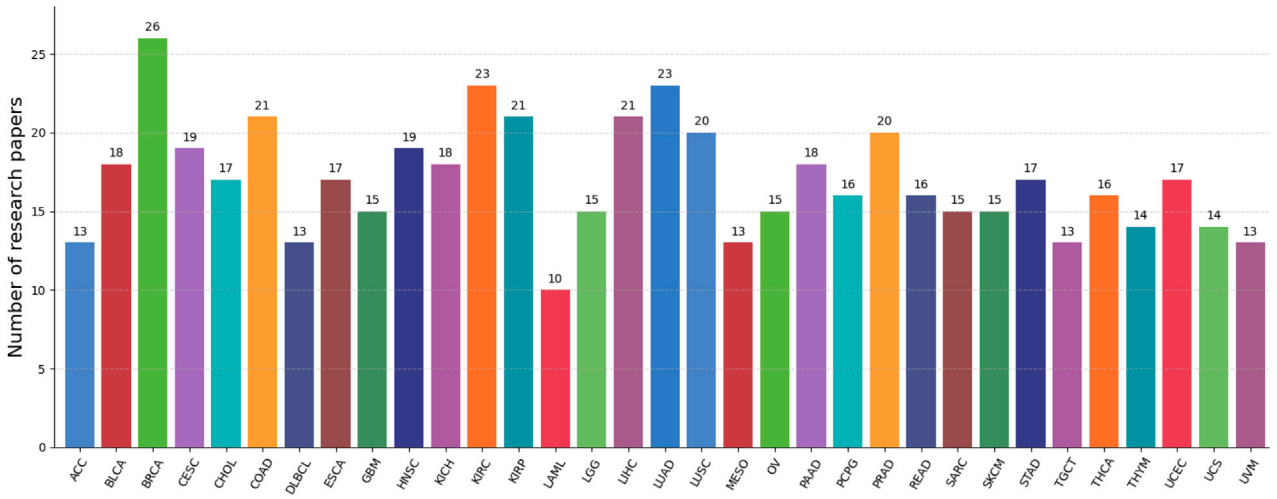

FIGURE 4
Frequency of cancer types used in pan-cancer classification studies reviewed in this paper. The x-axis indicates the specific cancer types, while the y-axis shows the number of research papers that utilized each cancer type's dataset. The data presented here is based on a statistical analysis of the literature reviewed in this manuscript.

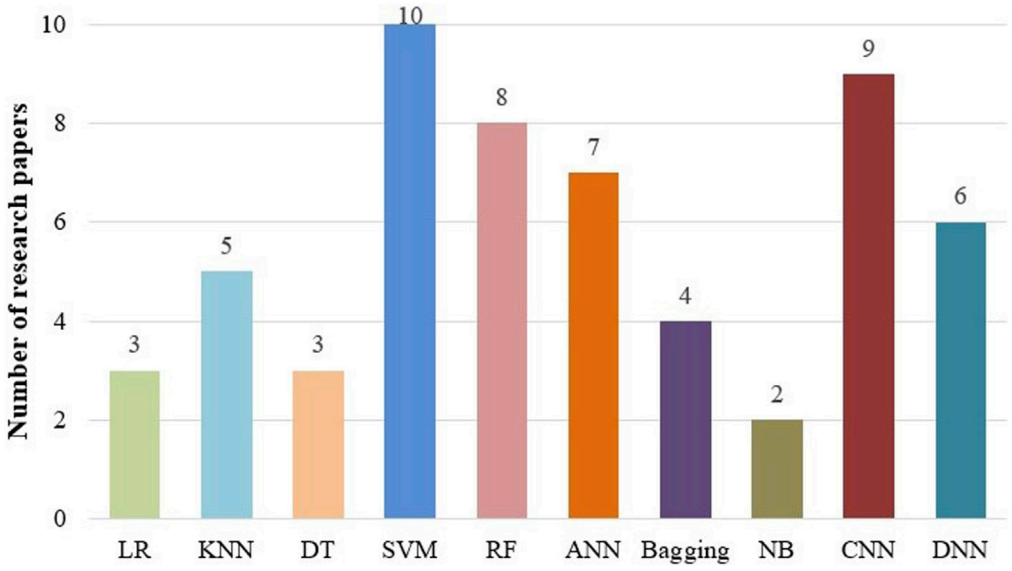

FIGURE 5
The frequency of different classifiers used in the pan-cancer classification research reviewed in this paper. Here, DNN refers to fully connected architectures (e.g., multilayer perceptrons), excluding convolutional neural networks (CNNs).

dataset has yet been conducted. Despite the variety of methods, there is still no standardized framework applicable in clinical practice. A major challenge remains the difficulty of generalizing results across studies and ensuring reproducibility. To address this, automatic and standardized methodologies that can be readily applied by non-expert users should be developed to better support clinical decision-making.

The application of ML and DL to multi-omics data also presents significant challenges. As multi-omics data derived from different platforms have varying distributions, this must be carefully considered before data integration (Reel et al., 2021). Furthermore, the integration of multiple omics datasets can generate noise and introduce redundant information. New algorithms must also be designed to effectively handle missing observations, as samples may be absent in one or more omics datasets (Leng et al., 2022).

In addition, class imbalance and overfitting are commonly reported issues in biomedical datasets. A training set composed of imbalanced classes can negatively influence the accuracy of a classifier, necessitating the use of statistical techniques such as

under- or oversampling (Misra et al., 2019). Moreover, the high-dimensional nature of multi-omics features can impact a classifier's performance, as correlated features introduce redundant information. To address this, optimal feature selection algorithms should be applied to select a limited, yet representative, subset of features.

5 Challenges and future work

Current pan-cancer classification methods leverage diverse data types and models to improve cancer type differentiation and inform clinical decision-making. This review systematically summarizes the methodologies, data sets, and evaluation metrics used in pan-cancer research, highlighting the progress in utilizing genomics, transcriptomics, and epigenomics to analyze tumor heterogeneity. We reviewed current pan-cancer classification methods, categorizing them based on the models used and assessing their performance across different data types.

Despite these advancements, challenges persist. Many models heavily depend on labeled data, overlooking the potential of abundant unlabeled data. Pan-cancer studies often focus on molecular features, neglecting clinical correlations with diagnosis and treatment. Additionally, data imbalance and the underrepresentation of some tumor types lead to unstable models.

Moreover, a lack of standardized benchmarks, limited cross-cohort validation, and a need for uncertainty quantification and calibration remain significant obstacles for the field. The absence of standardized and reproducible benchmarks hampers fair comparison across methods. We encourage the community to establish unified benchmark datasets with consistent splitting protocols—such as 5-fold stratified cross-validation (CV) standardized in TCGA-33 mRNA data with fixed preprocessing steps (e.g., gene filtering, normalization, and batch-effect correction) to facilitate transparent and reproducible evaluation. In addition, the use of common baseline models (e.g., logistic regression, random forest, standard deep neural networks) alongside more advanced architectures will help future studies assess genuine performance gains. Data imbalance, especially the underrepresentation of rare cancers, further restricts the generalizability of the model, calling for strategies such as data augmentation, few-shot learning, or federated learning to mitigate scarcity.

Future studies should prioritize semi-supervised learning (SSL) frameworks to leverage both annotated and unannotated datasets, thereby addressing data scarcity challenges. Self-supervised pre-training on large-scale unlabeled datasets could uncover tumor heterogeneity and enhance downstream classification tasks. Incorporating multi-modal data fusion—combining genomics, proteomics, and normal tissue data—could bridge the gap between molecular research and clinical applications.Beyond general cancer classification, future research must pivot toward more granular, clinically actionable predictions. This includes predicting cancer subtypes, disease stage, patient survival rates, and response to specific treatments, which directly informs personalized medicine.

In conclusion, addressing data limitations, imbalance, and clinical integration using advanced techniques such as SSL and multimodal fusion will enable more robust pan-cancer classification models, improving cancer prediction, diagnosis, and treatment for better patient outcomes.

6 Clinical translation and ethics

Developing robust pan-cancer models is the first step; translating them into effective clinical tools requires addressing a second set of critical challenges related to translation, generalizability, and ethics. Although a model may perform well on a single curated dataset, its utility in real-world clinical practice depends on its performance in diverse patient populations and healthcare systems.

Currently Available vs. Necessary Validation. Pancancer models are mainly in the research and development stages. Models that can now be used are typically those integrated into established platforms (like the CGC) for secondary research analysis, offering broad tumor type classification or basic survival predictions on standardized datasets (e.g., TCGA, CPTAC). However, most high-performing models require rigorous, multi-center external validation before they can influence patient care. To ensure external validity, models must be evaluated in data from multiple centers, reducing batch effects and acquisition bias that can arise when trained in the data set of a single institution (Cen et al., 2025). Batch effects, often stemming from variations in sequencing platforms or laboratory protocols across different institutions, can introduce confounding signals that a model may mistakenly learn as biological features. Similarly, acquisition bias can occur if certain rare cancer subtypes or patient demographics are disproportionately represented in the training data from a single center, limiting the model's ability to generalize to a broader patient cohort.

Equally important is equitable performance across diverse demographic groups. The precision of a model must remain consistent regardless of the race, sex, or age of the patient, to ensure fair clinical outcomes and prevent health disparities from being exacerbated (Desai et al., 2022). These validation efforts must be accompanied by strict attention to data privacy and informed consent, particularly given the reliance of pan-cancer studies on large-scale, sensitive patient data. Concurrently, the increasing complexity of deep learning models highlights a critical need for interpretability, enabling clinicians to understand model predictions and extract meaningful biomarkers that inform clinical decision-making with confidence (Su et al., 2024). Going beyond simply identifying individual genes, interpretable models can provide pathway-level attribution, linking predictions to entire biological processes (e.g., the p53 signaling pathway), which offers more clinically actionable and biologically meaningful insights.

To be reliable for high-stakes clinical decisions, a model must also provide more than a single prediction. It is crucial for models to offer uncertainty estimation, which allows clinicians to gauge the confidence of the model in its prediction. A well-calibrated model, for example, will have its predicted probability (e.g., a 90% chance of a certain tumor type) accurately reflect its true correctness. Such reliability measures are essential to build trust and ensure the safe deployment of these models in patient care. Furthermore, potential regulatory considerations are paramount; any model intended for diagnostic or prognostic use must undergo rigorous review by

regulatory bodies (such as the FDA) to ensure safety, efficacy, and clinical benefit.

In conclusion, the path from a pan-cancer model to a clinical tool is complex. It requires a holistic approach that moves beyond technical performance metrics to embrace the crucial factors of external validation, cost-effectiveness, and ethical responsibility. This comprehensive perspective is essential for developing models that are not only accurate in a research setting but are also robust, trustworthy, and beneficial in real-world clinical applications.

Author contributions

JW: Conceptualization, Supervision, Writing – review and editing, Investigation. JZ: Writing – review and editing, Validation, Methodology, Writing – original draft. XD: Writing – original draft, Investigation, Conceptualization, Writing – review and editing. CY: Formal Analysis, Methodology, Writing – review and editing, Supervision. CF: Funding acquisition, Resources, Writing – review and editing.

Funding

The author(s) declare that no financial support was received for the research and/or publication of this article.

Acknowledgments

JW conceived and designed the algorithm and analysis. JW and JZ gathered all the data, designed the study, conducted experiments, and drafted the manuscript. JW, JZ, XD, CF, and CY contributed to results analysis and discussions, and gave the final approval of the version to be published. CF and CY supervised the study and revised the manuscript. We thank LetPub (www.letpub.com) for its linguistic assistance during the preparation of this manuscript.

Conflict of interest

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

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

References

Abdullahi, A., Bawazeer, K., Alotaibai, S., Almoaither, E., Al-Otaibi, M., Alaskar, H., et al. (2020). "Pretrained convolutional neural networks for cancer genome classification," in *2020 3rd international conference on computer applications and information security (ICCAIS)* (IEEE), 1–5.

Al Mamun, A., and Mondal, A. M. (2019a). "Feature selection and classification reveal key lncRNAs for multiple cancers," in *2019 IEEE international conference on bioinformatics and biomedicine (BIBM)* (IEEE), 2825–2831.

Al Mamun, A., and Mondal, A. M. (2019b). "Long non-coding RNA based cancer classification using deep neural networks," in *Proceedings of the 10th ACM international conference on bioinformatics, computational biology and health informatics*, 541.

Al Mamun, A., Duan, W., and Mondal, A. M. (2020). "Pan-cancer feature selection and classification reveals important long non-coding RNAs," in *2020 IEEE international conference on bioinformatics and biomedicine (BIBM)* (IEEE), 2417–2424.

Al Mamun, A., Tanvir, R. B., Sobhan, M., Mathee, K., Narasimhan, G., Holt, G. E., et al. (2021). Multi-run concrete autoencoder to identify prognostic lncRNAs for 12 cancers. *Int. J. Mol. Sci.* 22, 11919. doi:10.3390/ijms222111919

Albaradei, S., Napolitano, F., Thafar, M. A., Gojobori, T., Essack, M., and Gao, X. (2021). Metacancer: a deep learning-based pan-cancer metastasis prediction model developed using multi-omics data. *Comput. Struct. Biotechnol. J.* 19, 4404–4411. doi:10.1016/j.csbj.2021.08.006

AlShibli, A., and Mathkour, H. (2019). A shallow convolutional learning network for classification of cancers based on copy number variations. *Sensors* 19, 4207. doi:10.3390/s19194207

Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., et al. (2021). Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *J. big Data* 8, 53–74. doi:10.1186/s40537-021-00444-8

Ameen, A., Alganmi, N., and Bajnaid, N. (2025). Stacked deep learning ensemble for multiomics cancer type classification: development and validation study. *JMIR Bioinforma. Biotechnol.* 6, e70709. doi:10.2196/70709

Ashrafizadeh, M., Najafi, M., Ang, H. L., Moghadam, E. R., Mahabady, M. K., Zabolian, A., et al. (2020). PTEN, a barrier for proliferation and metastasis of gastric cancer cells: from molecular pathways to targeting and regulation. *Biomedicines* 8, 264. doi:10.3390/biomedicines8080264

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., et al. (2012). NCBI GEO: archive for functional genomics data Sets—Update. *Nucleic acids Res.* 41, D991–D995. doi:10.1093/nar/gks1193

Bhattacharya, A., Ziebarth, J. D., and Cui, Y. (2013). SomamiR: a database for somatic mutations impacting microRNA function in cancer. *Nucleic acids Res.* 41, D977–D982. doi:10.1093/nar/gks1138

Bray, F., Laversanne, M., Sung, H., Ferlay, J., Siegel, R. L., Soerjomataram, I., et al. (2024). Global cancer statistics 2022: globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA a cancer J. Clin.* 74, 229–263. doi:10.3322/caac.21834

Cai, H., Gupta, S., Rath, P., Ai, N., and Baudis, M. (2015). arrayMap 2014: an updated cancer genome resource. *Nucleic acids Res.* 43, D825–D830. doi:10.1093/nar/gku1123

Capper, D., Jones, D. T., Sill, M., Hovestadt, V., Schrimpf, D., Sturm, D., et al. (2018). DNA methylation-based classification of central nervous system tumours. *Nature* 555, 469–474. doi:10.1038/nature26000

Cava, C., Salvatore, C., and Castiglioni, I. (2023). Pan-cancer classification of gene expression data based on artificial neural network model. *Appl. Sci.* 13, 7355. doi:10.3390/app13137355

Cen, X., Lan, Y., Zou, J., Chen, R., Hu, C., Tong, Y., et al. (2025). Pan-cancer analysis shapes the understanding of cancer biology and medicine. *Cancer Commun.* 45, 728–746. doi:10.1002/cac2.70008

Cheerla, N., and Gevaert, O. (2017). MicroRNA based pan-cancer diagnosis and treatment recommendation. *BMC Bioinforma.* 18, 32–11. doi:10.1186/s12859-016-1421-y

Chen, Y., Li, Z., Chen, X., and Zhang, S. (2021). Long non-coding RNAs: from disease code to drug role. *Acta Pharm. Sin. B* 11, 340–354. doi:10.1016/j.apsb.2020.10.001

Cho, H. J., Shu, M., Bekiranov, S., Zang, C., and Zhang, A. (2023). Interpretable meta-learning of multi-omics data for survival analysis and pathway enrichment. *Bioinformatics* 39, btad113. doi:10.1093/bioinformatics/btad113

Consortium, I. C. G., Hudson, T. J., Anderson, W., Artez, A., Barker, A. D., Bell, C., et al. (2010). International network of cancer genome projects. *Nature* 464, 993–998. doi:10.1038/nature08987

Cui, S., Yu, S., Huang, H.-Y., Lin, Y.-C.-D., Huang, Y., Zhang, B., et al. (2025). miRTarBase 2025: updates to the collection of experimentally validated microRNA–target interactions. *Nucleic Acids Res.* 53, D147–D156. doi:10.1093/nar/gkae1072

Desai, D., Hlaing, S. S., Goyal, A., and Keogh, A. (2022). Racial disparities in oncology clinical trials, 40, 356, doi:10.1200/jco.2022.40.28_suppl.356

Divate, M., Tyagi, A., Richard, D. J., Prasad, P. A., Gowda, H., and Nagaraj, S. H. (2022). Deep learning-based pan-cancer classification model reveals tissue-of-origin specific gene expression signatures. *Cancers* 14, 1185. doi:10.3390/cancers14051185

Elsadek, S. F. A., Makhlouf, M. A. A., and Aldeen, M. A. (2019). "Supervised classification of cancers based on copy number variation," in *Proceedings of the International Conference on Advanced Intelligent Systems and Informatics 2018 4* (Springer), 198–207.

Fang, Y., and Fullwood, M. J. (2016). Roles, functions, and mechanisms of long non-coding RNAs in cancer. *Genomics, proteomics & Bioinforma.* 14, 42–54. doi:10.1016/j.gpb.2015.09.006

Forbes, S. A., Beare, D., Gunasekaran, P., Leung, K., Bindal, N., Boutselakis, H., et al. (2015). COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic acids Res.* 43, D805–D811. doi:10.1093/nar/gku1075

Formosa, A., Lena, A., Markert, E., Cortelli, S., Miano, R., Mauriello, A., et al. (2013). DNA methylation silences mir-132 in prostate cancer. *Oncogene* 32, 127–134. doi:10.1038/onc.2012.14

Frenkel-Morgenstern, M., Gorohovski, A., Vucenovic, D., Maestre, L., and Valencia, A. (2015). ChiTaRS 2.1—an improved database of the chimeric transcripts and RNA-seq data with novel sense–antisense chimeric RNA transcripts. *Nucleic acids Res.* 43, D68–D75. doi:10.1093/nar/gku1199

Galagali, H., and Kim, J. K. (2020). The multifaceted roles of microRNAs in differentiation. *Curr. Opin. Cell Biol.* 67, 118–140. doi:10.1016/j.ceb.2020.08.015

Gao, J., Aksoy, B. A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S. O., et al. (2013). Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.* 6, pl1. doi:10.1126/scisignal.2004088

He, X., Chang, S., Zhang, J., Zhao, Q., Xiang, H., Kusonmano, K., et al. (2007). MethyCancer: the database of human DNA methylation and cancer. *Nucleic acids Res.* 36, D836–D841. doi:10.1093/nar/gkm730

Hu, L., Yao, X., Huang, H., Guo, Z., Cheng, X., Xu, Y., et al. (2018). Clinical significance of germline copy number variation in susceptibility of human diseases. *J. Genet. Genomics* 45, 3–12. doi:10.1016/j.jgg.2018.01.001

Huang, W.-Y., Hsu, S.-D., Huang, H.-Y., Sun, Y.-M., Chou, C.-H., Weng, S.-L., et al. (2015). MethHC: a database of DNA methylation and gene expression in human cancer. *Nucleic acids Res.* 43, D856–D861. doi:10.1093/nar/gku1151

Huynh, P.-H., Nguyen, V.-H., and Do, T.-N. (2019). Novel hybrid DCNN–SVM model for classifying RNA-sequencing gene expression data. *J. Inf. Telecommun.* 3, 533–547. doi:10.1080/24751839.2019.1660845

Karakach, T. K., Flight, R. M., Douglas, S. E., and Wentzell, P. D. (2010). An introduction to DNA microarrays for gene expression analysis. *Chemom. Intelligent Laboratory Syst.* 104, 28–52. doi:10.1016/j.chemolab.2010.04.003

Khadirnaikar, S., Shukla, S., and Prasanna, S. (2023). Integration of pan-cancer multi-omics data for novel mixed subgroup identification using machine learning methods. *Plos one* 18, e0287176. doi:10.1371/journal.pone.0287176

Khalifa, N. E. M., Taha, M. H. N., Ali, D. E., Slowik, A., and Hassanien, A. E. (2020). Artificial intelligence technique for gene expression by tumor RNA-Seq data: a novel optimized deep learning approach. *IEEE Access* 8, 22874–22883. doi:10.1109/access.2020.2970210

Kim, B.-H., Yu, K., and Lee, P. C. (2020). Cancer classification of single-cell gene expression data by neural network. *Bioinformatics* 36, 1360–1366. doi:10.1093/bioinformatics/btz772

Kolesnikov, N., Hastings, E., Keays, M., Melnichuk, O., Tang, Y. A., Williams, E., et al. (2015). ArrayExpress update—simplifying data submissions. *Nucleic acids Res.* 43, D1113–D1116. doi:10.1093/nar/gku1057

Lappalainen, I., Almeida-King, J., Kumanduri, V., Senf, A., Spalding, J. D., Ur-Rehman, S., et al. (2015). The European Genome-phenome Archive of human data consented for biomedical research. *Nat. Genet.* 47, 692–695. doi:10.1038/ng.3312

Leibovitch, M., and Topisirovic, I. (2018). Dysregulation of mRNA translation and energy metabolism in cancer. *Adv. Biol. Regul.* 67, 30–39. doi:10.1016/j.jbior.2017.11.001

Leng, D., Zheng, L., Wen, Y., Zhang, Y., Wu, L., Wang, J., et al. (2022). A benchmark study of deep learning-based multi-omics data fusion methods for cancer. *Genome Biol.* 23, 171. doi:10.1186/s13059-022-02739-2

Li, Y., Kang, K., Krahn, J. M., Croutwater, N., Lee, K., Umbach, D. M., et al. (2017). A comprehensive genomic pan-cancer classification using the Cancer Genome Atlas gene expression data. *BMC genomics* 18, 508–513. doi:10.1186/s12864-017-3906-0

Li, J., Xu, Q., Wu, M., Huang, T., and Wang, Y. (2020). Pan-cancer classification based on self-normalizing neural networks and feature selection. *Front. Bioeng. Biotechnol.* 8, 766. doi:10.3389/fbioe.2020.00766

Li, M., Guo, H., Wang, K., Kang, C., Yin, Y., and Zhang, H. (2024). AVBAE-MODFR: a novel deep learning framework of embedding and feature selection on multi-omics data for pan-cancer classification. *Comput. Biol. Med.* 177, 108614. doi:10.1016/j.compbiomed.2024.108614

Li, B., Xiao, X., Zhang, C., Xiao, M., and Zhang, L. (2025). DGHNN: a deep graph and hypergraph neural network for pan-cancer related gene prediction. *Bioinformatics* btaf379. doi:10.1093/bioinformatics/btaf379

Liu, P. (2022). Pan-cancer DNA methylation analysis and tumor origin identification of carcinoma of unknown primary site based on multi-omics. *Front. Genet.* 12, 798748. doi:10.3389/fgene.2021.798748

Liu, X. S., Wu, H., Ji, X., Stelzer, Y., Wu, X., Czauderna, S., et al. (2016). Editing DNA methylation in the mammalian genome. *Cell* 167, 233–247.e17. doi:10.1016/j.cell.2016.08.056

Liu, B., Liu, Y., Pan, X., Li, M., Yang, S., and Li, S. C. (2019). DNA methylation markers for pan-cancer prediction by deep learning. *Genes* 10, 778. doi:10.3390/genes10100778

Lopez-Rincon, A., Martinez-Archundia, M., Martinez-Ruiz, G. U., Schoenhuth, A., and Tonda, A. (2019). Automatic discovery of 100-miRNA signature for cancer classification using ensemble feature selection. *BMC Bioinforma.* 20, 480–17. doi:10.1186/s12859-019-3050-8

Luo, K., Qian, Z., Jiang, Y., Lv, D., Zhu, K., Shao, J., et al. (2023). Characterization of the metabolic alteration-modulated tumor microenvironment mediated by TP53 mutation and hypoxia. *Comput. Biol. Med.* 163, 107078. doi:10.1016/j.compbiomed.2023.107078

Lyu, B., and Haque, A. (2018). "Deep learning based tumor type classification using gene expression data," in *Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics*, 89–96.

Mahin, K. F., Robiuddin, M., Islam, M., Ashraf, S., Yeasmin, F., and Shatabda, S. (2022). PanClassif: improving pan cancer classification of single cell RNA-seq gene expression data using machine learning. *Genomics* 114, 110264. doi:10.1016/j.ygeno.2022.01.001

Mesri, M., An, E., Zhang, X., Bavarva, J., Robles, A. I., Hiltke, T., et al. (2024). Abstract 1852: nci's Clinical Proteomic Tumor analysis Consortium: a proteogenomic cancer analysis program. *Cancer Res.* 84, 1852. doi:10.1158/1538-7445.am2024-1852

Misra, B. B., Langefeld, C., Olivier, M., and Cox, L. A. (2019). Integrated omics: tools, advances and future approaches. *J. Mol. Endocrinol.* 62, R21–R45. doi:10.1530/JME-18-0055

Mostavi, M., Chiu, Y.-C., Huang, Y., and Chen, Y. (2020). Convolutional neural network models for cancer type prediction based on gene expression. *BMC Med. genomics* 13, 44–13. doi:10.1186/s12920-020-0677-2

Nandwani, A., Rathore, S., and Datta, M. (2021). LncRNAs in cancer: regulatory and therapeutic implications. *Cancer Lett.* 501, 162–171. doi:10.1016/j.canlet.2020.11.048

Navarro Gonzalez, J., Zweig, A. S., Speir, M. L., Schmelter, D., Rosenbloom, K. R., Raney, B. J., et al. (2021). The UCSC genome browser database: 2021 update. *Nucleic acids Res.* 49, D1046–D1057. doi:10.1093/nar/gkaa1070

Pop-Bica, C., Pintea, S., Magdo, L., Cojocneanu, R., Gulei, D., Ferracin, M., et al. (2020). The clinical utility of mir-21 and let-7 in non-small cell lung cancer (NSCLC). A systematic review and meta-analysis. *Front. Oncol.* 10, 516850. doi:10.3389/fonc.2020.516850

Pös, O., Radvanszky, J., Buglyó, G., Pös, Z., Rusnakova, D., Nagy, B., et al. (2021). DNA copy number variation: main characteristics, evolutionary significance, and pathological aspects. *Biomed. J.* 44, 548–559. doi:10.1016/j.bj.2021.02.003

Qin, S., Tang, X., Chen, Y., Chen, K. K., Fan, N., Xiao, W., et al. (2022). mRNA-based therapeutics: powerful and versatile tools to combat diseases. *Signal Transduct. Target. Ther.* 7, 166. doi:10.1038/s41392-022-01007-w

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

Rong, Z., Liu, Z., Song, J., Cao, L., Yu, Y., Qiu, M., et al. (2022). MCluster-VAEs: an end-to-end variational deep learning-based clustering method for subtype discovery

using multi-omics data. *Comput. Biol. Med.* 150, 106085. doi:10.1016/j.compbiomed.2022.106085

Santucci, C., Carioli, G., Bertuccio, P., Malvezzi, M., Pastorino, U., Boffetta, P., et al. (2020). Progress in cancer mortality, incidence, and survival: a global overview. *Eur. J. Cancer Prev.* 29, 367–381. doi:10.1097/CEJ.0000000000000594

Siegel, R. L., Miller, K. D., Goding Sauer, A., Fedewa, S. A., Butterly, L. F., Anderson, J. C., et al. (2020). Colorectal cancer statistics, 2020. *CA a cancer J. Clin.* 70, 145–164. doi:10.3322/caac.21601

Sinha, R., Luna, A., Schultz, N., and Sander, C. (2021). A pan-cancer survey of cell line tumor similarity by feature-weighted molecular profiles. *Cell Rep. Methods* 1, 100039. doi:10.1016/j.crmeth.2021.100039

Su, L., Hounye, A. H., Pan, Q., Miao, K., Wang, J., Hou, M., et al. (2024). Explainable cancer factors discovery: shapley additive explanation for machine learning models demonstrates the best practices in the case of pancreatic cancer. *Pancreatology* 24, 404–423. doi:10.1016/j.pan.2024.02.002

Subramanian, S. L., Ray, M., DiGiovanna, J., Radenkovic, J., Tosic, M., Mirkovic, N., et al. (2021). Abstract 253: the Cancer genomics cloud: a secure and scalable cloud-based platform to access, share and analyze multi-omics datasets. *Cancer Res.* 81, 253. doi:10.1158/1538-7445.am2021-253

Sun, K., Wang, J., Wang, H., and Sun, H. (2018). GeneCT: a generalizable cancerous status and tissue origin classifier for pan-cancer biopsies. *Bioinformatics* 34, 4129–4130. doi:10.1093/bioinformatics/bty524

Tang, S., Li, S., Liu, T., He, Y., Hu, H., Zhu, Y., et al. (2021). MicroRNAs: emerging oncogenic and tumor-suppressive regulators, biomarkers and therapeutic targets in lung cancer. *Cancer Lett.* 502, 71–83. doi:10.1016/j.canlet.2020.12.040

Tomczak, K., Czerwińska, P., and Wiznerowicz, M. (2015). Review The Cancer genome atlas (TCGA): an immeasurable source of knowledge. *Contemp. Oncology/ Współczesna Onkol.* 2015, 68–77. doi:10.5114/wo.2014.47136

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* 10, 57–63. doi:10.1038/nrg2484

Wang, D., Gu, J., Wang, T., and Ding, Z. (2014). OncomiRDB: a database for the experimentally verified oncogenic and tumor-suppressive microRNAs. *Bioinformatics* 30, 2237–2238. doi:10.1093/bioinformatics/btu155

Wang, S., Zheng, Z., Chen, P., and Wu, M. (2019). Tumor classification and biomarker discovery based on the 5'isomir expression level. *BMC Cancer* 19, 127. doi:10.1186/s12885-019-5340-y

Wang, F.-a., Zhuang, Z., Gao, F., He, R., Zhang, S., Wang, L., et al. (2024). TMO-Net: an explainable pretrained multi-omics model for multi-task learning in oncology. *Genome Biol.* 25, 149. doi:10.1186/s13059-024-03293-9

Waterman, S. (2021). The Human genome Project: the beginning of the beginning. *Quant. Biol.* 9, 4–7. doi:10.15302/j-qb-021-0243

Wei, Q., Zhou, H., Hou, X., Liu, X., Chen, S., Huang, X., et al. (2022). Current status of and barriers to the treatment of advanced-stage liver cancer in China: a questionnaire-based study from the perspective of doctors. *BMC Gastroenterol.* 22, 351. doi:10.1186/s12876-022-02425-4

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R., Ozenberger, B. A., Ellrott, K., et al. (2013). The cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* 45, 1113–1120. doi:10.1038/ng.2764

Wilks, C., Cline, M. S., Weiler, E., Diehkans, M., Craft, B., Martin, C., et al. (2014). The Cancer Genomics Hub (CGHub): overcoming cancer through the power of torrential data. *Database* 2014, bau093. doi:10.1093/database/bau093

Wu, T.-J., Shamsaddini, A., Pan, Y., Smith, K., Crichton, D. J., Simonyan, V., et al. (2014). A framework for organizing cancer-related variations from existing databases, publications and NGS data using a High-performance Integrated Virtual environment (HIVE). *Database.* 2014, bau022. doi:10.1093/database/bau022

Wu, J., Chen, Z., Xiao, S., Liu, G., Wu, W., and Wang, S. (2024). DeepMoIC: multi-omics data integration *via* deep graph convolutional networks for cancer subtype classification. *BMC genomics* 25, 1209–1213. doi:10.1186/s12864-024-11112-5

Xie, B., Ding, Q., Han, H., and Wu, D. (2013). miRCancer: a microRNA–cancer association database constructed by text mining on literature. *Bioinformatics* 29, 638–644. doi:10.1093/bioinformatics/btt014

Yang, Y., and Fan, S. (2024). Small cell lung cancer transformations from non-small cell lung cancer: biological mechanism and clinical relevance. *Chin. Med. J. Pulm. Crit. Care Med.* 2, 42–47. doi:10.1016/j.pccm.2023.10.005

Yang, Z., Kotoge, R., Chen, Z., Piao, X., Matsubara, Y., and Sakurai, Y. (2024). Cmob: large-Scale cancer multi-omics benchmark with open datasets, tasks, and baselines. *arXiv e-prints*. doi:10.48550/arXiv.2409.02143

Ye, T., Li, S., and Zhang, Y. (2021). Genomic pan-cancer classification using image-based deep learning. *Comput. Struct. Biotechnol. J.* 19, 835–846. doi:10.1016/j.csbj.2021.01.010

Zhang, N., Wang, M., Zhang, P., and Huang, T. (2016). Classification of cancers based on copy number variation landscapes. *Biochimica Biophysica Acta (BBA)-General Subj.* 1860, 2750–2755. doi:10.1016/j.bbagen.2016.06.003

Zhang, X., Zhang, J., Sun, K., Yang, X., Dai, C., and Guo, Y. (2019). "Integrated multi-omics analysis using variational autoencoders: application to pan-cancer classification," in *2019 IEEE international conference on bioinformatics and biomedicine (BIBM)* (IEEE), 765–769.

Zhang, X., Xie, J., Yang, Z., Yu, C. K. W., Hu, Y., and Qin, J. (2025). Tumour heterogeneity and personalized treatment screening based on single-cell transcriptomics. *Comput. Struct. Biotechnol. J.* 27, 307–320. doi:10.1016/j.csbj.2024.12.020

Zhao, Y., Li, X., Zhou, C., Peng, H., Zheng, Z., Chen, J., et al. (2024). A review of cancer data fusion methods based on deep learning. *Inf. Fusion* 108, 102361. doi:10.1016/j.inffus.2024.102361

Zheng, R., Zhang, S., Zeng, H., Wang, S., Sun, K., Chen, R., et al. (2022). Cancer incidence and mortality in China, 2016. *J. Natl. cancer Cent.* 2, 1–9. doi:10.1016/j.jncc.2022.02.002