International Journal of Mathematics, Statistics, and Computer Science
Vol. 2, 2024
ISSN: 2704-1077 eISSN 2704-1069, DOI: 10.59543/ijmscs.v2i.8703

A Comprehensive Review of Artificial Intelligence Approaches in Omics Data Processing: Evaluating Progress and Challenges

Ali Mahmoud Ali[1], Mazin Abed Mohammed[2]
[1]Iraqi Commission for Computer and Informatics (ICCI), Informatics Institute for Postgraduate Studies (IIPS), Baghdad, Iraq; ms202210711@iips.edu.iq
[2]Artificial Intelligence Department, College of Computer Science and Information Technology, University of Anbar, Anbar, Iraq; mazinalshujeary@uoanbar.edu.iq

ABSTRACT: The primary objective of this study is to review and assess the best available research on omics-related artificial intelligence (AI) methods. Furthermore, it seeks to demonstrate the promise of AI approaches in omics data analysis and identify the critical problems that must be solved to achieve this potential fully. There are many moving parts when trying to make sense of a plethora of research through a literature review. Essential components include, for instance, clinical applications and literature collections. Other researchers have faced challenges, and the existing literature highlights them. Using a systematic strategy, we searched all relevant articles on omics and AI utilizing multiple keyword variations. We also seek additional research, such as guidelines, studies of comparison, and review studies. Challenges with AI, preprocessing, datasets, validation of models, and testbed applications arose when AI was used to analyze omics data. To solve these problems, several pertinent investigations were carried out. Our work offers unique insights into the intersection of omics and AI model fields, setting it apart from prior review articles. We anticipate that practitioners seeking an all-encompassing perspective on using AI in omics data processing would find this study's findings invaluable.

Keywords: Artificial intelligence, cancer diagnosis, Machine Learning, Deep Learning, Quantum Computing, Multi-Omics, Omics dataset.

1. INTRODUCTION

The word "omics" is commonly used in the bioinformatics and biology communities when discussing large-scale, all-encompassing methods for investigating and evaluating different parts or features of biological systems. Omics is a common suffix, and many forms of omics examine various facets of biological data. Omics is a term used to describe a branch of biological sciences that focuses on studying several -omics disciplines [1]. Integrating data from many omics methodologies allows researchers to get a more complete picture of intricate biological systems since each omics area offers a holistic perspective on a different component of biological information. Systematic biology is an approach to studying biological systems that incorporates omics data to describe and evaluate these systems' relationships and networks.

To get a more complete and holistic knowledge of biological systems, "multi-omics" means to combine and analyze data from many omics domains. Genomic, proteome, transcriptome, metabolic, epigenetic, lipid, and glycomic omics are among the many subfields that study different aspects of biological data. However, multi-omics methods broadly view a biological system's molecular composition and function by merging data from many omics fields. One of the main goals of multi-omics is to learn more about the interplay and connections between various molecular parts of the body. Systems biology, which aims to model and understand the many pathways and networks inside biological systems, is strongly related to this data integration. Scientists want to better understand biological processes by combining data from many omics fields [2].

The interplay of metabolites, proteins, genes, and other components in a living system may be better understood using multi-omics methods. This can pave the way for discoveries in personalized medicine, biomarker research, and illness processes. High-throughput technologies have allowed a lot of data to be collected in many omics fields. This has led to the development of efficient computational and analytical methods for combining and interpreting multi-omics data. Many areas of biomedical research use multi-omics techniques, including cancer studies, customized therapy, and studying complicated disorders. Targeted medicines and interventions can be developed using the knowledge gathered from multi-omics investigations [3]. Data quality, platform integration, and the requirement to handle computational and analytical complexity are some obstacles the sector

must overcome. Research in the field of multi-omics is always looking for new ways to help us better comprehend the complex molecular environment of living systems.

There are several hurdles to using ML and DL for cancer diagnosis using omics data. The huge size and complexity of omics data is one obstacle. Large volumes of data are required to train ML and DL algorithms, and they can be computationally expensive to run. Another issue is the lack of standardization in the gathering and interpreting of omics data. This can make comparing results from different studies challenging [4]. Despite these obstacles, using ML and DL for cancer diagnosis using omics data can transform cancer care. ML and DL algorithms can be utilized to create novel non-invasive cancer detection and diagnosis techniques. They can also be used to develop personalized cancer treatment strategies for patients. A wide range of computational approaches have emerged due to the exponential growth of computing power. Many believe the emergence of big data and AI technology will modernize traditional Chinese medicine. The advancements in technology, such as (AI) and fast computing, have boosted the prominence of machine learning, especially deep learning, in computational biology. This approach has shown impressive achievements in several areas of biology. DL approaches have led to notable advancements in several fields, including genetic variant identification, DNA methylation, and picture analysis. Applying DL to varied omics data shows exciting results. Current approaches still have some limits despite their potential. Due to the expanding diversity of accessible data types, many multi-omics techniques concentrate on a small range of data types, usually two or three, and have a limited scope involving specific samples or patients. Also, when using unsupervised learning, expected findings or classifier labels can be challenging to understand.

Continuing the preceding discussion, this research provides an exhaustive analysis of omics data combined with AI methods, covering every angle. We take a medical and technological approach to our study, recognizing the opportunities and threats that the deluge of highly complex and unstructured omics data presents to healthcare research. We combed through research publications to get a feel for the terrain, and our review study contributions and unique features are laid forth below:

- In this review, we focus on developmental research that has used AI methods to improve medical procedures; these studies provide new insights into how medical technology is evolving.
- We offer a synopsis of significant accomplishments made by studies in response to the needs of omics data, shedding light on the development of research in the field.
- To demonstrate the revolutionary effect of AI in this field, we emphasise the tangible benefits of analysing omics data using DL models.
- This study highlights the difficulties AI methods encounter when used with omics data. These challenges include problems such as a large number of variables, the scarcity of data, the absence of labeled data, and the need for reliable assessment tools.
- We present a review that uses AI methods to classify the vast amount of published information, outlining several paths in omics research. We hope that other scholars may find this organized, helpful framework.
- Our study investigates the latest advancements in artificial intelligence (AI) methods to handle omics data. Specifically, our main emphasis is on using multi-omics data to improve the precision of illness prognosis and prediction of outcomes.

2. MACHINE LEARNING FOR MULTI-OMICS DATA.

Machine learning (ML) is an effective and adaptable method in multi-omics data processing, contributing much to discovering new insights, developing accurate predictions, and advancing our knowledge of complex biological processes and disorders. The heterogeneous data integration: Many kinds of data comprise multi-omics data, from genomics to metabolomics. Machine learning is crucial in bridging these disparate data sets, allowing scientists to do comprehensive analyses and reveal hidden connections between variables that could go unnoticed [4].

Numerous features are often included in multi-omics datasets. Feature selection, or determining which biomarkers or variables are most important, is a strong suit of machine learning algorithms. This method helps researchers focus on relevant patterns by reducing background noise and improving data interpretability—identification of complex patterns and correlations in multi-omics data (pattern discovery). Unsupervised machine learning approaches, such as clustering and dimensionality reduction, are beneficial for unearthing previously concealed ways. By clustering samples with similar characteristics, researchers can get insight into the underlying biological phenomena [5].

Using supervised machine learning models, researchers may make predictions using multi-omics data. Diseases may be categorized, patients can be stratified, treatment outcomes can be predicted, and prognoses can be made. Machine learning models help with clinical decision-making by generating predictions on fresh,

unlabeled samples based on what they've learned from previously labeled data. ML plays a crucial role in customized medicine. The guide provides a personalized treatment plan based on the patient's unique molecular profile. This method improves medicinal efficacy while reducing collateral damage. By identifying significant pathways, genes, or molecular processes related to a given biological phenomenon or illness, machine learning techniques aid in interpreting multi-omics data. Scientists can use this information to form hypotheses and plan specific studies [6].

It is common practice to impute missing values, normalize, and reduce noise in multi-omics of the data before analysis. The use of machine learning methods streamlines these processes while guaranteeing the integrity of the data collected. Big and complicated multi-omics datasets provide a challenge to scalability. Researchers can quickly and easily evaluate massive datasets with the help of machine learning techniques. Visualization: Insightful visual representations of multi-omics data may be generated by integrating ML with data visualization techniques. Researchers may successfully explain their conclusions from a large, complicated dataset using simple language. Data preparation, model training and assessment, and other steps in studying multi-omics data are all automated by machine learning processes. This automation helps researchers save time and eliminates errors caused by human error [7].

Data Integration Across Research: Machine learning assists in harmonizing multi-omics data from multiple research sources. This capacity enables meta-analyses, letting researchers identify higher-quality biomarkers or relationships using a more extensive variety of data. Machine learning algorithms can learn and adapt to new data, making them well-suited to fast-paced settings like those in multi-omics research. Because of its flexibility, analysis can always be applied to new situations [8].

In this section, our study focused on ML approach review studies that only considered the multi-omics type of common omics data. Artificial intelligence's fast and expanding growth in recent years has made its way into healthcare, leading to an urgent need for accurate diagnosis to detect many diseases in their early stages before the patient’s condition deteriorates. The most dangerous of these diseases is cancer, according to biomedical data. Medical studies and machine learning have achieved promising results by demonstrating the interaction between affected cell levels. This facilitates disease detection and highly accurate prediction of the patient’s survival period. Machine learning using multi-omics has also enabled researchers to develop treatments for diseases targeting cancer cells. It is more accurate and practical than traditional treatments, such as chemotherapy, which destroys the body’s immunity and kills infected and healthy cells.

This study reviewed previous studies on using multi-omics with automated learning methods to diagnose disease and predict the patient's life, which has achieved promising results in this field. According to [4] They proposed a study to analyze multi-omics and take advantage of the ability of machine learning to detect cancer and classify its types and the possibility of using a cloud processing system to perform this complex task, which has become possible through reinforcement learning and based SARSA on-policy on learning a workload. They have developed schemes for the study, and these new schemes are hybrid for detecting cancer. Another study by [5] was presented to classify cancer as one of the main reasons for death, the world's solution. This study dealt, in particular, with breast cancer in people, The study aims to assess the second most prevalent cause of mortality among women globally by using two methods: percentage division and cross-validation. These techniques will be used to evaluate and compare the effectiveness of SVM and KNN classification models. These are considered evaluation techniques, and they reached the result that the level of classification accuracy of the SVM is higher than that of KNN when using cross-validation. The accuracy was 95.7081%, an advantage of the SVM classifier. However, when using percentage split, the KNN classifier achieved higher accuracy than the SVM classifier, with a high accuracy of 95.4220%. Also, new research by [6] evaluated the efficacy of the Support Vector Machine (SVM) and Random Forest (RF) classification methods in classifying breast cancer (BC). Achieved a success rate of 95.45% for SVM & 90.90% for RF. Through the above results, both methods demonstrated the possibility of using them to classify breast cancer due to their excellent results and high accuracy.

The new framework suggested by [7] conducted research where they developed a prediction framework using ML. This framework integrated multi-omics data with information related to cancer lncRNAs. This study presents a novel ML technique for forecasting disease-related long non-coding RNAs (lncRNAs) using multi-omics data and a neural network to consolidate neighborhood information. Furthermore,

Another study by [8] focused on heterogeneous diseases at the clinical, histological, and molecular levels. In particular, this study addressed primary breast cancer (PBC) using unsupervised machine learning techniques. Unsupervised clustering methods were used. These methods have proven effective in improving prediction through machine learning algorithms and have enabled the understanding of the relationships between components and heterogeneous clinical signs of the disease. Improving the detection of (PBC) cancer may be necessary for identifying disease subgroups to aid treatment and accurate diagnosis. Another study by [9] Their study

demonstrated the feasibility of identifying and discovering the subtype of breast cancer through the audio spectrum that is analyzed using machine learning algorithms at the biomolecular level. The PLS-DA method distinguishes the molecular subtypes of breast cancer based on the APSD spectra. This method achieved an accuracy of 84%.

The researchers in the study [10] applied machine learning techniques to a data set on miRNAs, which are fluids spread throughout the body, and obtaining them does not require significant surgical intervention. They presented a novel approach to reducing the dimensions of this data so that it can be handled and interpreted using different classification algorithms to detect cancer types and subtypes. The researchers in [11] suggested a two-step methodology for examining the prognosis of cancer of the breast. Using breast cancer next-generation sequencing (NGS) data, researchers successfully identified the most critical miRNA biomarkers linked to the illness. This technology, known as integrated feature selection techniques for machine learning (ML), has been successfully utilized in clinical settings by researchers. Most current multi-view data clustering methods are imperfect and have several problems. The existing methods contain Not less than one of the following issues:

1. Common points between the data may be ignored.

2-Original information representation is not exploited well in multi-view.

3 - Incomplete data and incorrect entries cannot be categorized.

The study proposed A new method [12] to avoid these limitations. This is a novel consensus learning approach to incomplete multi-view clustering (CLIMC). This method effectively employs clustering-based machine learning in the study [13]. The researchers suggested a two-stage procedure that uses particular classifiers for identifying subtypes and classifying tumors according to their health. Utilizing the TGCA dataset for training and the GSE68085 dataset for testing. A digital medical system for cancer detection still faces challenges related to privacy, processing quickly, and enhancing the reliability of cancer prediction, as seen in previous studies in another studies [14]. They introduced a new cancer detection paradigm based on heterogeneous cloud computing nodes with improved accuracy, speedier processing, and more security.

Laboratory (MCMOCL) approaches for predicting multiple cancers with many classifications are based on a multi-cancer clinical dataset and include conjoint learning, autoencoder, and XGBoost techniques. Based on the study[15]The authors proposed a classification model using extreme gradient boosting (XGBoost) and complex multi-omics data to differentiate between initial and late-stage cancers. The XGBoost model was used to analyse four categories of cancer data obtained from TCGA. Its efficacy was assessed in comparison to other well-recognised ML methodologies. Experimental findings demonstrate the strategy produced statistically comparable or noticeably superior prediction accuracy. In addition, the findings of their research indicate that using an autoencoder to integrate diverse omics data may enhance the precision of cancer stage classification. Lung cancer likelihood in plasma (Lung-CLiP) is an ML approach by [16] created and verified that this method may effectively distinguish early-stage lung cancer patients from risk-matched controls. This method produced performance on par with tumor-informed ctDNA detection and allows for fine-tuning test specificity to support various clinical applications. The outcomes of this approach illustrate the value of risk-matched patients and controls in cfDNA-based screening trials and demonstrate the potential of cfDNA for lung cancer screening. Another research, as referenced by the citation [17], used multi-omics data and classifiers such as (SVMs) and random forests (RFs) to show the potential of ML and classification approaches in predicting the occurrence of cancer and early stages of infection. This study showed accuracy in the test results. Many cancer diseases, such as colon, breast, lung, etc., achieved high accuracy between 85.29% and 100%.

In another study by [18], Data from the TCGA, which includes seven multi-omics indicators and other clinical markers, was used to investigate four different forms of cancer. They also embraced cancer integration using the MultIkernel LeaRning (CIMLR) clustering approach. They created a uniform pipeline for raw data preprocessing to achieve better quality data by a released study[19] were they created a new data mining method for precise prediction of breast cancer (BC), a major killer of women globally. Developing a more accurate computerized system of experts (ES) for BC diagnosis is their top priority. Artificial Neural Networks (ANNs) and SVM were employed to analyze BC data. A new model based on low-rank approximations was suggested by [20] as a quick way to find the main subspace that different types of data share. The probabilistic model's convexity of the low-rank regularized likelihood function makes it possible for the model to fit well and reliably. Hundreds of cancer samples can be clustered unsupervised to find potential molecular subgroups in the reduced low-dimensional subspace. The LRA cluster (low-rank-based multi-omics data clustering) did better at clustering test datasets than the previous method. Next, the LRA cluster was applied to the TCGA's massive multi-omics dataset on cancer. The findings of the pan-cancer study show that tumors originating from diverse types of tissues are typically grouped, except squamous-like carcinomas.

The omics data offers various subtyping capabilities for different cancer types, as seen when looking at only one form of cancer. There was a study that they gave [21] the clinical results for TNBC are dismal, and there

is currently no approved targeted treatment for this aggressive subtype of breast cancer. Ten studies that examined miRNA profiling in TNBC were pooled together in this meta-analysis. They trained a Naive Bayes classifier on miRNA signatures. They used the robust rank aggregation method to correctly tell the difference between TNBC and non-TNBC samples in a test dataset. This showed that it was very good at diagnosing both types of cancer. The study's analysis showed that the discovered miRNAs have a crucial role in improving TNBC diagnosis and treatment. One of the top killers of females is breast cancer. Most known survival analyses center on the factors' relationships to patients' five-year survival rates. To this day, there is still no definitive solution to the individualized question of breast cancer survival rates. The research by [22] aims to predict the individual survival times of breast cancer patients. Two machine-learning issues are derived from the customized question. The first problem is using a binary categorization to determine if a patient has a five-year survival rate. The second thing to do is construct a regression model to predict how long the patient will live over the following five years. One way to predict the likelihood of breast cancer is by analyzing a patient's methylome. This regression model [22] uses Crystall, a novel method, to identify methylomic characteristics. In both cases, their models successfully predict the five-year survival rate of breast cancer patients with an MAE of about one month. The identified biomarker genes are highly associated with breast cancer.

Furthermore, regarding bioinformatics, ML, and pattern classification, medical data categorization based on microarray gene expression has consistently been among the most challenging research areas. The study [23] presented two versions of kernel ridge regression (KRR)—radial basis kernel ridge regression (RKRR) and wavelet kernel ridge regression (WKRR)—for microarray medical dataset categorization. Ignorant or duplicated genes are to blame for the enormous complexity and small sample sizes of microarray medical datasets. Microarray datasets overcame the curse of dimensionality using modified cat swarm optimization (MCSO). An evolutionary algorithm inspired by nature chooses the most essential traits from the datasets. They provide four examples from the binary and multiclass microarray medical datasets to demonstrate the suitable classifiers. Databases for breast cancer, prostate cancer, colon tumors, and leukemia fall under the first group, whereas datasets for leukemia1, leukemia2, SRBCT, and brain tumor1 go under the second. The experimental findings reveal that WKRR performs better than RKRR and that KRR is the best model overall, regardless of the dataset. After comparing the results of binary and multiclass datasets, this study concludes that, across all models, the binary class yields somewhat superior outcomes than the multiclass.

They [24] introduced Cat Swarm Optimization (CSO) as an innovative swarm intelligence method. CSO, which is trained by seeing cats in action, consists of two sub-models that mimic the actions of cats: tracing mode and seeking mode. Using six different test functions, the experimental findings reveal that CSO performs better than PSO. Many more influential people are members of the CSO. The CSO has a lot of power. According to the experimental data, 1-PSO with a factor of weighting typically outperforms pure PSO in terms of speed to better solutions, and the CSO is even better. 2-To make it realistic, the mixing ratio (MR) must remain minimal so the cats may spend most of their time searching. 3- It's easy to put into operation. 4- CSO is easily parallelizable, allowing the utilization of many cat swarms simultaneously to search the solution space more effectively. This can result in faster convergence and improved solutions.

Furthermore, when interacting with the CSO, any method that may be vulnerable should be used with prudence. The shortcomings of CSO:

1- Parameter Tuning: CSO, like many optimization algorithms, necessitates the adjustment of numerous parameters, including the number of cats, the maximum number of iterations, and control parameters affecting cat movement. Finding the best parameter values for a given situation can be difficult and time-consuming. 2- CSO may take longer to converge than other optimization techniques, particularly for high-dimensional and complex problems. A huge number of iterations may be required to obtain an ideal solution. 3- The performance of CSO can be affected by the initial placements of the cats in the search space. Inadequate initialization can result in inferior results. 4- It may not be the ideal solution for all problems. Its efficiency varies based on the characteristics of the problem.

The purpose of this study by [25] is to enhance the multiclassification capability of machine learning models used for the subtyping of cancer. In their systematic approach, they utilized data optimization and machine learning approaches. Within an interpretable framework, this novel method for data optimization integrates several data integration methodologies with multi-stage feature selection, paving the way for further improvements. Machine learning classifiers' actions were clarified by applying the SHAP theory.

As the power of computing has increased, several computational methodologies have emerged. Collecting various types of genome-wide data has become more affordable due to recent technical breakthroughs. Compiling this information into a complete picture of a particular illness or biological process requires

computational methods. This problem can be solved by similarity network fusion (SNF), which combines networks with samples (like patients) for all the available data types into a single network that includes all the data types. For instance, SNF leverages the complementary nature of their data sets to provide a more complete picture of a disease in a cohort of patients by computing and merging patient similarity networks. According to [26], SNF combined data on microRNA (miRNA) expression, DNA methylation, and mRNA expression for five cancer datasets. Compared to established integrated approaches and single-data type analysis, SNF is superior at recognizing cancer types and determining survival. This approach can capture continuous phenotypes, an improvement over conventional subtyping approach.

There is a wide variety of breast cancers. Subtyping the illness and finding the genetic markers that drive these subtypes are essential for precision oncology in breast cancer through study [27]—investigating the possibility of developing a novel computational approach to subtyping breast cancer. The Cancer Genome Atlas assessed 762 breast cancer patients using Bayesian tensor factorization (BTF), a method for integrating multi-omics data on breast cancer. This data includes RNA-sequencing expression patterns, the number of copy variations, and DNA methylation. Researchers employed a consensus clustering method to determine breast cancer subtypes using BTF's factorized latent features. Kaplan-Meier (KM) survival algorithms evaluated breast cancer patients' survival habits according to subtype. Modern, state-of-the-art methods for cancer subtyping were compared to the suggested approach. They used the 17 optimized latent elements of the BTF-subtyping approach to identify six main breast cancer subtypes. No other options showed survival rates ($p = 0.05$) different from the suggested method. There is statistical significance in the patterns of the identified clusters, according to the tests. The results showed that the proposed approach might be an excellent way to use publicly accessible multi-omics data to identify breast cancer subgroups successfully.

In 2017, lung cancer was the most common malignancy among men, followed by prostate cancer (PCa). The illness kills more than half of its victims in Peru despite the availability of several treatments. Tissue samples and the Gleason grading system are the conventional methods for diagnosing and grading PCa. Multiple molecular subtypes of PCa have been suggested for diagnosis and prognosis in the paper [28]. They conducted research to demonstrate their strategy. The study's primary objective is to identify the genes that significantly impact the prediction of a patient's time without disease using genomic expression and to develop a tool to do so. The NN, in particular, has contributed to expanding and improving current data categorization methods. The Local Interpretable Model-Agnostic Explanations (LIME) method is a good base for the ANN-based automated genomic classification technique because it lets the network pick the best features to tell the difference. As a proof of concept, 242 recurrence-related genes from 499 PCa genomes were used to build the NNS. With a loss of less than 5% and an accuracy of 96.9%, the neural network generated from genomic expression can forecast a recurrent time within three months. The final network, an ANN rather than a conventional, completely connected layer, provided the anticipated survival rate or time to recurrence. The study's findings, derived from the LIME approach, demonstrate that this collection of genes is highly predictive and instructive regarding recurrence.

ES-JSS, a new static ensemble selection approach, was developed by [29]. It combines structural sparsity with joint spectral clustering. This approach integrates structural sparsity and spectral clustering into a unified framework. The resulting ensemble selection is both space-efficient and resistant to test instances. This research made use of twenty-five datasets collected by KEEL and UCI. Primary learners are selected dynamically, case by case, using dynamic ensemble selection. Complexity is not much enhanced. So far, no ensemble selection method that enhances space complexity and is resilient to test examples has been developed. The output of the ensemble selection process is space-efficient and resistant to test cases.

Another approach a multi-view hierarchical ensemble clustering algorithm, was suggested by [30] for finding disease subtypes it was tested on actual patient data that included seven types of cancer. This method is derived from the Greek word "parea," which denotes a gathering of friends who exchange stories, beliefs, and ideas. Across all six cancer types tested, the Parea ACGT dataset beats the current state-of-the-art. Incorporating several fusion and clustering approaches allows for the flexible and uncomplicated building of ensemble operations. While Parea*hc*2 allows for hierarchical clustering, Parea*hc*1 is restricted to employing only two hierarchical clustering methods, making it an inflexible mechanism for building clustering ensembles of any complexity. Additional research is needed to attain clinically meaningful results due to the study's constraint of 606 participants, which is modest compared to the other clusters.

In [31], the authors presented a new ensemble clustering approach called PA. Two more steps were added to the goal function of the famous fuzzy k-means algorithm. Incorporating a penalty term is the first step toward making the algorithm immune to the initialization of cluster centroids. The second one is to alter feature weights in a clustering process programmatically. On datasets including actual cancer gene expression profiles, the

suggested algorithm (PA) achieves better results than the majority of cutting-edge clustering algorithms. On the other hand, this method will automatically cluster something other than financial, streaming, or category data.

Presented by [32] kESVR, kESVR the drug response values of each patient are predicted using a k-means Ensemble Support Vector Regression (kESVR) model specifically designed for gene expression data in cell lines. The kESVR uses a data-driven approach by integrating supervised and unsupervised learning approaches. The system employs embedded clustering techniques such as PCA and k-means clustering and local regression using SVR to predict drug reactions and capture the overall pattern. It is designed to handle missing data and filter out noise caused by outliers. Once the issue of overfitting was resolved, this model surpassed its previous versions in terms of prediction speed and precision. This model integrates supervised and unsupervised learning elements, making it robust and based on empirical data. There are two limitations associated with this approach. Firstly, feature selection has not yet been achieved. Secondly, kESVR only relies on gene expression data. The most recent studies using ML on multi omics data summarized in Table 1.

Table 1. The summary on utilization of ML techniques on Omics data

Author(s)/ Year	Techniques used	The summary	Dataset used	The advantages	The limitaions
1-M.A. Mohammed et al (2023) [4]	machine learning methods	Machine learning techniques are very beneficial when dealing with cancer datasets in practical applications. Reinforcement learning, supervised and unsupervised learning, deep learning (DL), (SVM), (RF), and (ANN) are all proposed as potential methods for analyzing cancer datasets in practical applications.	clinical data	1- Train the various models by assigning weights based on the layers.	1- Only single datasets were used to examine the accuracy of cancer data set.
2-Anita Desiani. et. al. (2022)[5]	1-SVM classifications	This study focuses on incorporating a levy flight (LF) approach to enhance the fruit fly optimisation (FO) technique. Additionally, a levy flight and fruit fly-based (LFFO-SVM) is developed.[33]	https://www.openml.org/search?type=data&sort=runs&id=15	Achieving 93.83% accuracy, 91.22% recall, and 96.53% specificity.	When employing the % split evaluation technique, accuracy is reduced.
	2-SVM classification	SVM classification method and used the Radial Basis Function (RBF)	1-Wisconsin Breast Cancer (WBC) dataset 2- Wisconsin Diagnostic Breast Cancer dataset (WDBC).	1-accuracy of (WBC) =96.58%. 2- accuracy of (WDBC)= 95.91%.	1-When employing the % split evaluation technique, accuracy is reduced.

Author(s)/ Year	Techniques used	The summary	Dataset used	The advantages	The limitaions
	3-(KNN) classification	Used approach of the K-Nearest Neighbor for classification data		1- By using the percentage split assessment technique, 1 achieves a higher accuracy rate of 97.8571% compared to the SVM. 2-Both methods achieve excellence with performance metrics over 90% in accuracy, recall, and F1-score. Breast cancer classification proves that SVM and KNN are solid and practical algorithms.	1- When applying the cross-validation assessment technique, SVM achieves a higher accuracy of 95,7081%.
3- Yuan et al. (2021) [7]	The LGDLDA algorithm is a network-based method for predicting LncRNA-gene-disease associations.	Their suggested LGDLDA gathers information from neural network neighborhoods, multi-omics data, and machine learning algorithms to guess how lncRNAs might be linked to diseases.	from three databases LncRNADisease v2.0, http://www.rnanut.net/lncrnadisease/ Lnc2Cancer, and MNDR v2.0 databases	Compared to other methods, LGDLDA performs better in stability tests. Concerning the AUC value, LGDLDA fared better than the four different approaches. In comparison to IDHI-MIRW, NCPLDA, LncDisAP, and NCPHLDA, LGDLDA has (AUC) of 0.935, which is 0.067, 0.134, 0.205, and 0.131 greater, respectively.	The sparse lncRNA-disease connection simulated network that supports our hypothesis. Incomplete data impacts LGDLDA.
4-Ferro, S. et al (2021) [8]	Machine Learning-Based Unsupervised Cluster	Machine learning algorithms may help us better comprehend the links between clinical disorders. This research aims to show that unsupervised learning methods may be helpful in better primary breast cancer (PBC) classification.	712 women with PBC were studied in this study.	Clustering methods are adaptable and may be applied to various data types, such as categorical, numerical, or mixed data, making them useful for a wide range of applications.	Some clustering techniques, especially hierarchical and density-based methods, are highly computational and may not scale well to huge datasets or high-

Author(s)/ Year	Techniques used	The summary	Dataset used	The advantages	The limitaions
					dimensional data. The results of clustering are not always interpretable.
	1- K-means	• Probably the most widely used clustering method • A measure of similarity that is based on the distance between two points on Earth • Cluster number must be given pre-specified	712 women with PBC were studied in this study.	1-Easy implementation 2-Implemented in a variety of software 3-Reduced computational costs	1-It may not be resistant to outliers and skewed distributions. 2-Difficulty identifying complicated links in data 3-Only handles continuous variables 4- Clustering based on determinism.
	2- Self-organizing maps (SOM)	• A limited variant of K-means based on (ANN) • The number of clusters needs to be communicated beforehand.	712 women with PBC were studied in this study.	1-Perfect for high-dimensional data. 2-Easy to execute and adapt to a variety of situations.	1- Clustering based on determinism 2-Only handles continuous variables 3-Can be unstable when identifying clusters is challenging.

Author(s)/ Year	Techniques used	The summary	Dataset used	The advantages	The limitaions
	3-Hierarchical agglomerative clustering (HAC)	• Cluster with an organizational structure • Shown visually as a tree structure with nested levels • Unspecified number of clusters.	712 women with PBC were studied in this study.	1. Easy to implement. 2-Easy to understand. 3-Ideal for identifying patient subgroups.	1-Extremely high computing costs due to sophisticated clustering structures 2- Finding the optimal number of clusters becomes a real challenge when working with massive datasets.
	4- Gaussian mixture model (GMM)	• Clustering approach based on models. • Data collected from samples distributed normally. • The quantity of clusters must be specified beforehand.	712 women with PBC were studied in this study.	1. Easy to implement. 2-Implemented in a variety of software. 3-Estimates the likelihood of belonging to each cluster.	1-Could be susceptible to distributions that are skewed and outliers. 2-Only works with straight lines. Thirdly, working with huge datasets and complex clustering structures results in extremely high computational costs.

Author(s)/ Year	Techniques used	The summary	Dataset used	The advantages	The limitaions
5- J. Li et al. (2023)[9]	photoacoustic spectral analysis (PASA) using machine learning (ML)	This research made use of photoacoustic spectrum analysis (PASA) in conjunction with the (PLS-DA) in order to determine the molecular subtypes of (BC) at the bio-macromolecular level in vivo.	Data will be provided upon request. The data used to draw these results has yet to be available to the public, but it is available upon reasonable request from the author.	1- PASA was able to obtain an accuracy of 84%, with AUC, values of 0.93350 for luminal, 0.87500 for triple negative, and 0.82000 for HER2.	1- There were some discrepancies seen among the PA identification and histopathology staining findings in the specified areas.
6- Lopez-Rincon et al. (2020) [10]	1- Feature Selection (an ensemble recursive feature selection algorithm)	When the accuracy of classification falls below a specified threshold or when a certain number of characteristics is reached, the procedure is restarted using the remaining data. Subsequently, the characteristics with the lowest scores are eliminated.	gene expression omnibus (GEO) datasets.	1-The unique technique proposed does not need users to define threshold levels randomly. 2- To guarantee a classifier with excellent performance (>90% average accuracy in classifying) and to decrease the initial set of 253 miRNAs to just 5 for each case study, a recursive ensemble feature selection technique was used.	1-This method works, but it's still dependent on a classification algorithm, which might be biased when scoring features.

Author(s)/ Year	Techniques used	The summary	Dataset used	The advantages	The limitaions
	2-the classifier random forest (RF)	ten-fold cross-validation was performed on the discovered 5-miRNA signature for tumour categorization.	clinical studies from the GEO database	1-reliability of 97.61% 2- Because testing just five miRNA levels is more straightforward and less likely to result in inaccurate measurements than testing every potential miRNA level.	1-Training a random forest model may be computationally demanding, particularly for huge datasets.
7-Sarkar et al. (2021) [11]	ML Integrated Ensemble of Feature Selection Techniques is a two-phase technique.	The researchers developed a new set of feature selection methods that work with machine learning to find the most essential miRNA markers for different types of breast cancer. A Cox regression-based survival analysis would follow this. They selected the one with the highest classification accuracy using all attributes out of seven machine learning algorithms.	TCGA – with the aid of breast cancer NGS data	1-The best ML approach (in this case, RF) was chosen. 2-It is clear from the findings that RF delivers more accuracy (76.5761, ±0.33) % than the other six approaches. SVM, ANN, KNN, DT, NB, and DISCR yielded outputs of (74.9094, ±0.48) %, (74.9094, ±0.37) %, (67.1014, ±0.35) %, (64.4565, ±0.46) %, (70.5978, ±0.35) %, and (73.1884, ±0.35) %, respectively.	1-It should be emphasized that the TCGA data amount used is quite tiny.

Author(s)/ Year	Techniques used	The summary	Dataset used	The advantages	The limitaions
8- J. Liu, S. Teng, L. Fei et al. (2021)[12]	a novel Consensus Learning method to Incomplete Multi-view Clustering (CLIMC)	The CLIMC method is a new way to solve the complex IMC problem. It combines graph Laplacian regularization, consensus similarity graph learning, and low-dimensional consensus representation learning.	Several multi-view datasets. http://erdos.ucd.ie/datasets/3sources.html. https://github.com/GPMVCDummy/GPMVC/tree/master/partialMV/PVC/ recreateResults/data. http://erdos.ucd.ie/datasets/segment.html. http://mlg.ucd.ie/aggregation/index.html.	1. Extensive research on many multi-view datasets reveals that CLIMC improves benchmarking methods. 2. There is at least one problem with most existing approaches of clustering partial multi-view data: a) The common relationships between data items are disregarded across all representations. b) The original data representation's complementing multi-view information is underutilised. c) Data with negative entries or incomplete conditions cannot be dealt indiscriminately. In order to overcome these restrictions, a pioneering (CLIMC) was developed.	1- They have an interest in enhancing the technique for modelling nonlinear interactions in future investigations, since the analysed linkages among the points of data in CLIMC are linear. 2-Feature selection and supervised multi-view learning are not considered within the purview of 2-CLIMC.
9- P. Andreini, S. Bonechi, M. Bianchini et al. (2022)[13].	1. The initial step was differentiating tumor and healthy samples using the SVM classifier. 2. Two- the RF Because of	They use two improvised classifiers to provide a two-stage approach for distinguishing between tumour and healthy samples and identifying specific subtypes. The study used two distinct datasets, TGCA for the purpose of training and GSE68085 for the purpose of testing, in order to evaluate the outcomes.	datasets: TGCA, GSE68085	1- The two-stage classification strategy allows for the specialization of a distinct model for each step, allowing tremendous flexibility. To classify tumors as benign or malignant, they	1-For training and assessment, the study employed a limited sample size. This might make the findings less applicable to

Author(s)/ Year	Techniques used	The summary	Dataset used	The advantages	The limitaions
	this, RFs can deal with feature spaces with many dimensions.			employed the SVM and used the RF to identify tumor subtypes.	different groups.
10-M.A. Mohammed et al. (2023)[14]	MCMOCL Hybrid (federated learning, auto-encoder, and XGBoost)	In this study, they build the MCMOCL method framework, which uses the system's shortest processing time and highest cancer prediction accuracy to train and evaluate multi-omics datasets for multi-cancer applications. In the study, data was trained and tested on distinct nodes using learning-enabled federated auto-encoder techniques. Then, the aggregated node chosen for decision execution received the data.	(MCMOCL) https://www.mirbase.org/ { HYPERLINK "https://github.com/al2na/methylKit/treemaster/data "}	1-MCMOCL It is distinguished by having the lowest training, testing, and classification times among all available machine learning schemes: auto-encoder has 1500.0 (ms), the XGBoost has 1300.0 (ms), and MCMOCL has 600.0 (ms). They used these disparate traits to properly diagnose cancer inside patients, achieving a nearly 99 percent validity accuracy.	1- AES provides more security but has higher computational and memory needs. Lightweight algorithms value efficiency over security. 2- Federated learning is the practice of training machine learning models using decentralized data sources, which necessitates frequent communication among devices or computers. 3- Building federated learning systems may be complicated, and dealing with failure circumstances can be difficult. 4- Image from the tiny Dataset 250

Author(s)/ Year	Techniques used	The summary	Dataset used	The advantages	The limitaions
11- B. Ma et al. (2020)[15]	extreme gradient boosting (XGBoost) "XGBoost classification"	To differentiate between early and late-stage cancers, the authors of this study proposed a classification model that uses XGBoost's capabilities in conjunction with more complex multi-omics data.	TCGA https://github.com/lab319/Cancer_prognosis_classification-/tree/master/Data	1-The experimental findings demonstrated that their strategy produced statistically significant superior or equivalent predicted performance. 2-AUC, XGBoost outperforms SVM, RFDNN, KNN, NB, and Elastic Net.	1- Findings from a comparative analysis of the predictive accuracy of many classification algorithms on the KIRC dataset ACC: SVM outperforms XGBoost AUPR: RF outperforms XGBoost MCC: SVM outperforms XGBoost Remember that Elastic Net is superior to XGBoost. When the data type is DNA methylation
12-Jacob J. Chabon et al. (2020)[16]	(Lung-CLiP) classification framework	The Alizadeh and Diehn labs developed Lung-CLiP, an innovative method for diagnosing noninvasive early non-small cell of lung cancer (NSCLC). Utilising ML, this method predicts the presence of tumor-derived cfDNA in a blood sample by incorporating advanced sequencing library preparation techniques.	WBC+ and WBC−	Adenocarcinoma and non-adenocarcinoma histology patients were detected with a sensitivity and specificity of 98% using lung-CLiP. It used a future cohort of students from a separate school for independent validation. This lessens the likelihood of model overfitting, which might lead to results that could be more optimistic.	1- More patients must be studied before the performance features of Lung-CLiP can be clearly established. 2- Because developed Lung-CLiP in a population mostly consisted of smokers, it is probable that performance among nonsmokers will be worse.

Author(s)/ Year	Techniques used	The summary	Dataset used	The advantages	The limitaions
13-Akram Mohammed et al. (2017) [17]	The SVM, and RF	By using RF classification and implementing nine-fold cross-validation plus one, a total of 10 classifiers were created. Out of these, nine were designed for individual tissues, two were specifically for distinguishing cancerous tissues from normal tissues across numerous tissue types, and one was built for identifying normal tissues across diverse tissue types. Distinguishing between healthy and cancerous tissues	The NCBI (GEO) repository was used to acquire microarray gene expression data.	1- With an impressive testing accuracy of 97.89%, the sample was correctly identified as normal or cancerous by the heterogeneous tissue bio-class classifier. 2- The single-tissue models exhibited an accuracy ranging from 85.29% to 100% in correctly classifying an instance of a specific tissue-type as either normal or malignant.	1-The classification characteristics (variables) employed by random forests substantially influence their performance. If the feature set is inadequate or missing critical information, classification accuracy may suffer. 2- Random forests are susceptible to class imbalance.
14- V. Crippa et al. (2023)[18]	CIMLR	CIMLR is a machine learning approach based on kernels that leverages multi-omics data to stratify patients and categorize cancer subtypes. Several omics data types are transformed into kernel matrices, which represent the degree to which samples are comparable according to the feature values, as the first step in the program.	TCGA, they obtained seven omics data sets from cBioPortal for each cancer kind. https://www.cbioportal.org/	1-Complex structures and nonlinear interactions across data types can be captured using CIMLR's kernel-based technique. 2-They concentrated on tumours with no strong consensus on multi-omics subgroups.	1- By using TCGA multi-omics data, which consists of seven different omics characteristics for each patient along with chosen clinical outcomes, the researchers focused specifically on four forms of cancer for the purpose of this study.

Author(s)/ Year	Techniques used	The summary	Dataset used	The advantages	The limitaions
15-Moloud Abdar, Vladimir Makarenkov (2019) [19]	CWV-BANNSVM (the confidence-weighted, voting method and the boosting ensemble methodology) are used to boost an ANN (BANN). The CWV-BANNSVM model combines boosting, ANNs (BANN) and two SVMs).	The researchers offer a new data mining approach for predicting breast cancer (BC). Both the SVM and, the ANN were used to analysis BC data. According to the first experiment, adjusting these regularisation parameters may greatly improve the performance of the traditional, (SVM) algorithm used for breast cancer diagnosis.	(WBCD), available in the UCI repository	1-Achieving 100% accuracy. 2-To avoid overfitting, they identified and applied some optimal polynomial SVM parameter values. 3-very adaptable and requires less recordings for the training step 4- The CWV-BANN-SVM approach is very effective for early breast cancer detection.	1-overfitting issue
16-Dingming Wu et al. (2015) [20]	LRAcluster is a method that is unsupervised used to identify the primary low-dimensional subspace of multi-omics data with large-scale and high-dimensional characteristics, specifically for molecular classification purposes.	Their novel LRA-based integrated probability model has shown exceptional computational efficiency and stability, enabling it to effectively process various data kinds.	TCGA multi-omics dataset	1-The results reveal that the LRAcluster approach is substantially quicker than iCluster+ for reliable model fitting. 2- LRAcluster can rapidly converge in a few rounds. 3- LRAcluster is five times quicker than iCluster.	1-(From 11), Based on existing omics data, LRAcluster did not uncover any robust molecular subgroups for the remaining 7 cancer types.

Author(s)/ Year	Techniques used	The summary	Dataset used	The advantages	The limitaions
17-Naorem, Muthaiyan and Venkatesan. (2019)[21]	Using the training data set, several classification models, triple-negative breast cancer. 1-Naïve Bayes 2- Sequential minimal optimization (SMO) 3- Random Forest [RF]	The RRA approach was used in the study to incorporate TNBC miRNA expression profiling datasets.	Gene Expression Omnibus (GEO)	1-Th accuracy of Naïve Bayes =96.8447 % 2- Th accuracy of SMO = 96.966% 3- Th accuracy of RF = 96.4806%	
18- S. Liu et al (2021)[22]	Crystall	A revolutionary algorithm With Crystall, they were able to identify the methylomic features of this regression model. Since they assumed that features with small positive numbers for the model parameters didn't add much to the class labels, they excluded them from further analysis.	TCGA	1-Crystall achieves better results than the four most popular feature selection algorithms that rely on regression (ISVR, Lasso, E-net, and Ridge). 2-The experiment was carried out, and the Crystall worked admirably, identifying 40 markers with 72% accuracy. In the provided study, just one type of omics data (methylation) is taken into account.	1- Out of the 221 tests, 71 were deemed negative due to death within five years, while the remaining 150 were classified as positive. Thirty-two patients survived for over five years, although 103 of 221 samples had known expiration dates. 2-A single omics type is employed to identify breast cancer-related genes.

Author(s)/ Year	Techniques used	The summary	Dataset used	The advantages	The limitaions
19-P. Mohapatra et al. (2016)[23]	kernel ridge regression (KRR)	Two versions of (KRR), (RKRR) ,and (WKRR) are introduced in this study for microarray medical dataset categorization within the MCSO method to classify the traits it produces.	Microarray medical Datasets http://www.gems-system.org/ http://datam.i2r.a-star.edu.sg/datasets/krbd/	1-No matter the dataset, KRR performs better than other models in the experiments, and WKRR performs better than RKRR. 2-Computationally expensive CSO beats PSO, according to the literature in this field. 3-The testing accuracy reached by KRR for BC, PC, colon tumor, leukemia tow tupe, leukemia1, leukemia2, SRBCT, and the Brain Tumor1 is 0.97, 0.97, 0.95, 0.92, 0.96, 0.86, and 0.94, sequentially. The results show that KRR outperforms the other approaches, including RR, OSRR, SVMRBF, SVMPoly, and RF.	1-Except for the prostate cancer dataset, WKRR has superior accuracy compared to RKRR, one of two variants of KRR. 2-The curse of dimensionality is an eternal problem for 2-microarray datasets used in medical research.
20-Meshoul, S. et al. (2022)[25]	multi-stage feature selection (FS) framework	They describe a (FS) system with many phases and two techniques to data integration. At each level, four ML models, namely SVM, RF, additional trees, and XGBoost, were evaluated using multi-omics data. The SHAP framework was used to illustrate the impact of various attributes on the categorization process.	TCGA multi-omics data	Step 1: Conduct preliminary trials with early integration. XG-Boost was attained. The accuracy: 79.885% The precision is 80.101%. Recall rate: 79.885% The F1-score is 79.1940%. The ROC-AUC is 92.4380%. Step 2: Selecting 1st-level features with late integration. Extra trees were planted. The accuracy is 80.8580%. Precision is 79.9400%. Recall rate: 80.8580% F1	1- While the objective is to eliminate noise and irrelevant characteristics, there is a danger that features that help to model performance in certain scenarios or under certain conditions may be removed.

Author(s)/ Year	Techniques used	The summary	Dataset used	The advantages	The limitaions
				rating: 79.7980% The ROC-AUC is 93.4590%. The use of an extra-trees filter with second-tier feature selection resulted in the attainment of optimal results. Step 3: Choose a second level Extra trees were planted. The accuracy is 84.500%. The precision is 84.756%. Recall rate: 84.500% F1-score: 84.0210% The ROC-AUC is 95.313%.	
21- Bo Wang et al. (2014)[26]	SNF, to integrate data	In order to get a comprehensive understanding of a certain illness or biological process, it is important to use computational techniques to gather and consolidate the relevant data. SNF addresses this challenge by creating networks of samples (such as patients) for each accessible form of data and then merging them into a unified network that encompasses the whole spectrum of underlying data.	TCGA	1- In terms of identifying tumour kinds and predicting survival, SNF outperforms both current integrative approaches and single data type analysis. 2- The SNF algorithm has the potential to provide valuable insights from a little number of samples. It is also robust against noise and variations in the data. Furthermore, SNF is scalable to analyse a high number of genes due to its construction based on sample networks. 3- The NFS technique has the advantage of going beyond standard subtyping procedures to find continuous phenotypes.	1- Data pretreatment is a vital stage, and preparing various datasets for SNF can be hard and time-consuming. It might be difficult to ensure that data is in an adequate format and that appropriate similarity measures are used. 2- SNF necessitates the selection of several parameters, such as the number of nearest neighbours, the similarity measure, and the fusion

Author(s)/ Year	Techniques used	The summary	Dataset used	The advantages	The limitaions
					parameter. The selection of these factors can have an influence on the outcomes and may necessitate optimisation. 3- when the SFN dependent P values are compared Survival rates for KRCCC and LSCC are worse. The reason for this is that both KRCCC and LSCC have a minimum one subtype that is specific to a small group of patients.

Author(s)/ Year	Techniques used	The summary	Dataset used	The advantages	The limitaions
22-Q. Liu et al. (2022)[27]	BTF-CNMF	The reason to this study was to formulate a new unsupervised learning BTF-CNMF technique for classifying breast cancer tumours into prognostically different subgroups.	TCGA dataset used https://www.cancer.gov/tcga	1. It effectively identified six intrinsic subtypes of breast cancer. 2-The proposed strategy demonstrated unique survival patterns (p 0.05). 3-An important enhancement in the proposed technique is the use of an approach based on models to integrate three complex omics-data sets for the purpose of tumor stratification . 4-The model-based approach is capable of directly handling high-dimensional data.	1- One limitation of this study is that the researchers only employed 762 TCGA breast cancer patients, hence the results cannot be generalized to other instances. 2- BTF-CNMF misclassified the third subtype of the disease. 3-When it came time for subtyping, the researchers only used the patient-directional factorized matrix, ignoring the other two factorized matrices' latent information. The data type-directional matrix and the gene-directional matrix are two examples of such matrices; the former links data types to latent rankings while the latter links genes to latent ranks.

Author(s)/ Year	Techniques used	The summary	Dataset used	The advantages	The limitaions
23-Marin Urol et al. (2019) [28]	LIME algorithm	They employed the (ANN) automatic genomic classification approach, which is based on the LIME algorithm and allows the network to pick the traits with the greatest discriminative potential.	499 prostate cancer (PCa) genomes	1-The resultant ANN can accurately predict the timing of reoccurring within a period of three months, with an accuracy level of 96.90% and a false-positive rate of less than 5.0%, using genomic expression data. 2-The findings obtained from the LIME algorithm demonstrate that this particular group of genes have predictive capabilities for recurrence and holds significant importance in the prediction process.	1- ANN are commonly referred to as "black box" models since the reasoning behind their predictions is difficult to explain. This might be a disadvantage when professionals need to defend or explain medical choices to patients.
24-Zhenlei Wang et al. (2021)[29]	ES-JSS	With this approach, spectral clustering and structural sparsity are combined into one framework, resulting in a space-and test-case-resistant ensemble selection solution.	Twenty-five KEEL & UCI datasets are used.	1-ES-JSS performed best on 64.0% (16/25) of the datasets, whereas none of the other techniques performed better than 20.0% (5/25). ES-JSS achieved superior performance compared to KNORA-E on 84.0% (21/25) of the datasets. Additionally, ES-JSS beat the other methods by at least 60.0% (15/25) on an absolute minimum of 60.0% (15/25) of the datasets, such as METADES (15/25), and KNOP (15/25). Based on t-tests conducted at a significance threshold of 5.0%, which is ES-	1-The effectiveness of this method depends on the characteristics of the data, and the specific issue being addressed. It may not always outperform simpler approaches, notably for limited datasets or when the techniques' underlying assumptions are not satisfied.

Author(s)/ Year	Techniques used	The summary	Dataset used	The advantages	The limitaions
				JSS was not shown to be statistically less effective than equivalent techniques on 80.0% of the datasets. The findings indicate that ES-JSS performs better.	
25-B. Pfeifer et al. (2023)[30]	$Parea_{hc}$	A strategy for identifying distinct subcategories of illnesses. Method using real-world multi-view patient data from seven types of cancer.	ACGT	1- In six of the seven cancer types tested, Parea outperforms the modern method. 2- This feature enables the easy and adaptable development of collective procedures, using a wide range of fusion and clustering techniques. 3-Adaptable approach for creating clustering ensembles of any complexity. 4- $Parea^{1}_{hc}$ is restricted to using just two hierarchical clustering techniques, whereas $Parea^{2}_{hc}$ facilitates hierarchical clustering.	1- Given Cox log-rank test was employed to evaluate the data, using a significance level of α=0.05. HCfused shown superior outcomes only for skin cutaneous melanoma (SKCM) across all types of malignancy. 2- There is a total of 606 patients. Additional research is required to get clinically significant results due to the limited number of individuals in the remaining clusters.
26- I. Khan et al. (2021) [31]	a new ensemble clustering algorithm proposed algorithm (PA)	The popular fuzzy k-means approach was updated by adding two more phases to its target function. One approach is to include a penalty term into the algorithm to ensure that it is not affected by the initial selection of cluster centroids. The second phase involves automating a clustering technique that iteratively modifies feature weights.	Cancer gene expression profiles that are authentic. Armstrong in 2002, Risinger in 2003, Yeoh in 2002, Bhattacharjee in 2001, Chowdary in 2006, Gordon in 2002, and Laiho in 2007.	1- The noise levels in the dataset are handled by the feature weighting portion. These approaches are very useful in determining the optimum number of subsets of features	1- The approach does not automatically execute cluster analysis on category, financial, or

Author(s)/ Year	Techniques used	The summary	Dataset used	The advantages	The limitaions
				from a given dataset. 2-The overall quality of clustering is higher. 3- On these data sets, the suggested algorithm (PA) outperforms the bulk of cutting-edge clustering approaches.	data that is streaming. 2-The proposed algorithm (PA) requires validation on a current real-world dataset.
27- Majumdar, A. et al. (2021)[32]	kESVR was developed to forecast each pharmaceutical response value. for a particular patient.	The kESVR is a hybrid model that combines both supervised as well as unsupervised learning techniques, using data as its main driver. The system employs embedded clustering, PCA, k-means, local regression, and SVM to predict drug response and capture a global trend, while effectively handling missing data and filtering out noise from outliers.	(CCLE) http://www. broadinstitute.org/ccl (CTRP). https://portals.broadinstitute.org/ctrpCurated data used in technique comparison may be available at the GitHub URL provided by Chen and Zhang's study. [34]. https://github.com/abhishekmaj08/kESVR.gi	1- This model outperforms earlier forecasting approaches in terms of speed and precision, and it avoids overfitting. 2- It is a strong, driven by data model that incorporates both supervised as well as unsupervised features into its operation.	1-kESVR relies only on gene expression data. 2- At the moment, kESVR does not provide feature selection.

3. DEEP LEARNING FOR MULTI-OMICS DATA

The importance of DL in omics data analysis has drawn a lot of interest from researchers in the last years, rendering DL and omics data among the main areas of study. No clear limits on the advancement of this field have been identified in this setting. Consequently, further study is needed to follow this research path and provide a complete picture. We aim to provide a thorough understanding and in-depth analysis by analysing and classifying the relevant literature. Consequently, the final selection of relevant publications is categorized in this area of the research into three primary classifications: evaluations, clinical implementations, and further investigations that use deep learning and omics data. This comprehensive review assists researchers in comprehending the essential aspects of deep learning in omics by emphasizing and elaborating on new areas of study[35]. Deep learning methods are perfect for analyzing intricate, varied, and high-dimensional data sets, including omics datasets. An overview of some of the main applications in precision medicine, including the discovery of biomarkers for the categorization of illnesses, is given in the section that follows. Key applications are summarized in Figure 1.

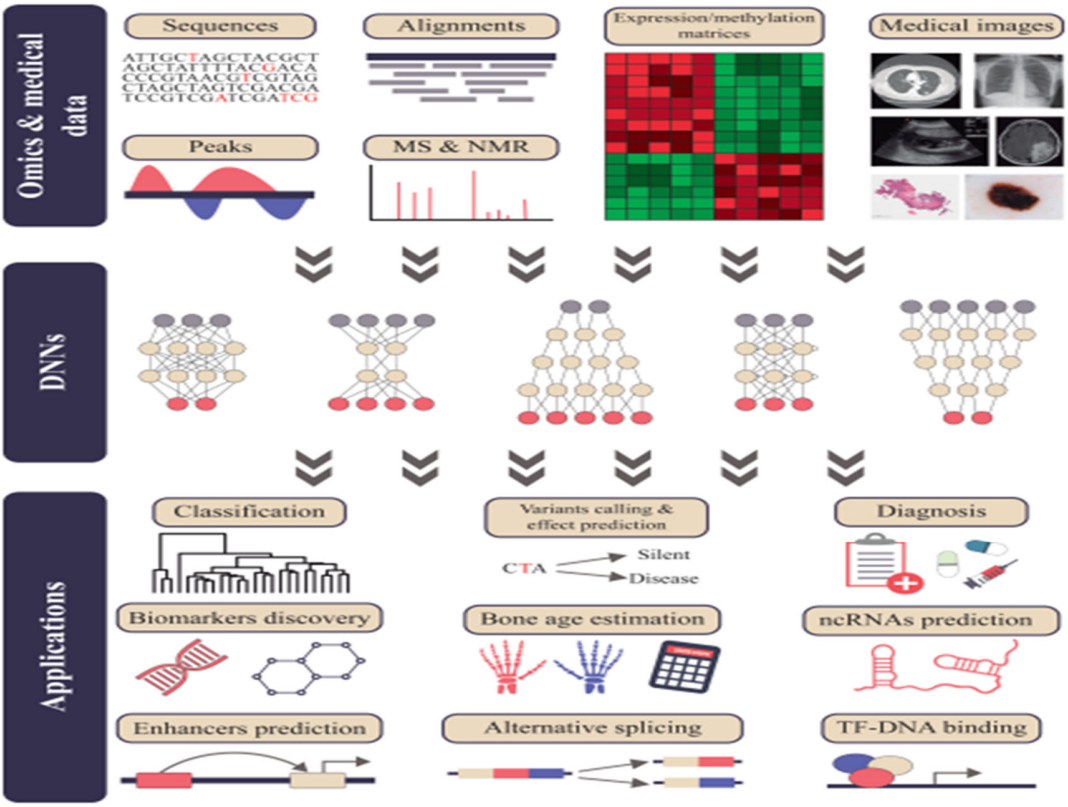

Figure 1. shows how DNNs have been used for several biological data categories [36]

The many forms of data are listed at the top. In the center, there are not many examples of DNN structures. At the bottom is a list of some of the most important applications that these methods have created. Pictures from medical imaging sources include TCIA [37] for MRI and CT scans; Chest X-Ray database [38] for X-rays; MedPix® (https://medpix.nlm.nih.gov) for the US; TCGA [39] for histopathological images; and ISIC (https://www.isic-archive.com) for skin lesions. Several graphical components were obtained using Stockio (https://www.stockio.com/) and Freepik (https://www.freepik.com/). Choosing Features: Multi-omics datasets often include a large number of characteristics. Deep learning algorithms work well for feature selection or determining which biomarkers or variables are most significant. Feature selection is a crucial data preparation technique that has been experimentally shown to lower the dimensionality of features and enhance the performance of learning algorithms in practical applications. A feature selection model is constructed using a deep model approach, which consists of a fully connected network, a pairwise linked layer, and a decision network. The fully connected layer is utilised for converting the input into scores that the decision network can recognize. It subsequently provides the errors for training the complete network.

On the other hand, the pairwise connected layer establishes connections between two networks [40]. DL models general, connect many problems that influence the outcomes of classification or regression operations. The deep learning models used for omics research identified several challenges connected to deep learning. The subsequent subsections (refer to Figure 3) delineate the process of preparation, datasets, model validation, and testbed applications [35].

3.1. Omics and DL Model Challenges.

Several challenges exist to overcome when integrating deep learning models with omics data. Due to the high dimensionality and missing values of omics data, complex methods are required for preprocessing to preserve biological integrity. Computational resources and scalable DL architectures are necessary for handling large-scale datasets. In healthcare applications, model interpretability is of the utmost importance, and to overcome overfitting and capture complicated patterns, advanced regularization is required. Robust validation procedures are necessary because there aren't enough labeled datasets, and people's bodies differ. Healthcare applications raise ethical concerns, focusing on protecting patient privacy and reducing prejudice. The absence of defined standards makes model evaluation more complex, and integrating data from many omics sources adds another layer of complexity. Its need to work together to find new solutions to overcome these obstacles and ensure deep learning is used ethically and effectively in omics research and analysis. We can summarize them such as:

1- Low Noise Reduction Efficacy. 2- Traditional Feature Selection Methods. 3- High Computational Cost. 4- Selection of Best Prediction Model. 5- Ignoring Prior Knowledge. 6- DL Model Explain ability. 7- Omics Interaction Identification. 8-Time-Consuming. 9-Single-Omics Dimension. 10- Reproducibility and Generalization. 11-Selection of Suitable Evaluation Metric. 12- Classification Type.

There are also many challenges in dealing with datasets provided in Figure 2.

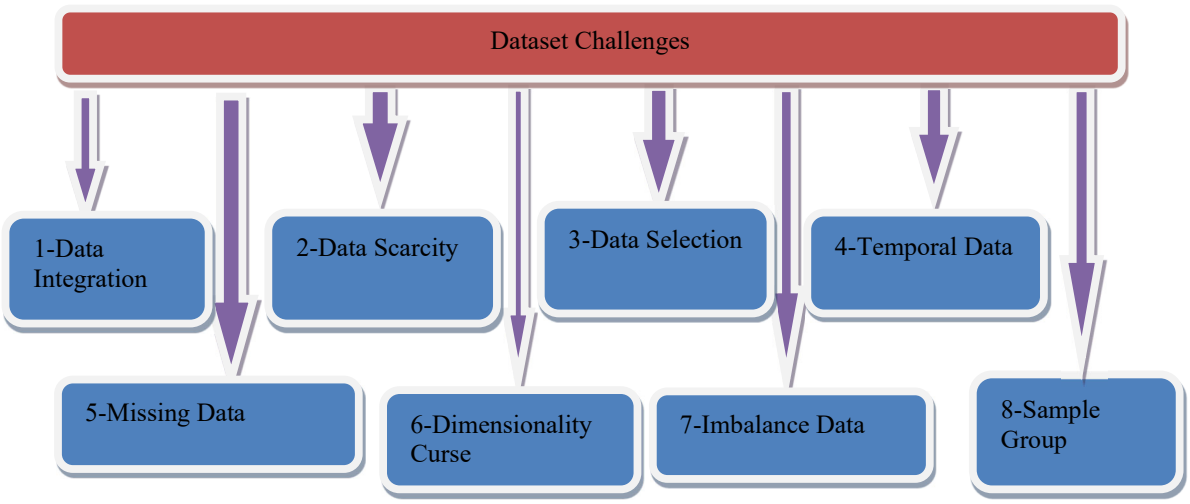

Figure 2. challenges in dealing with datasets

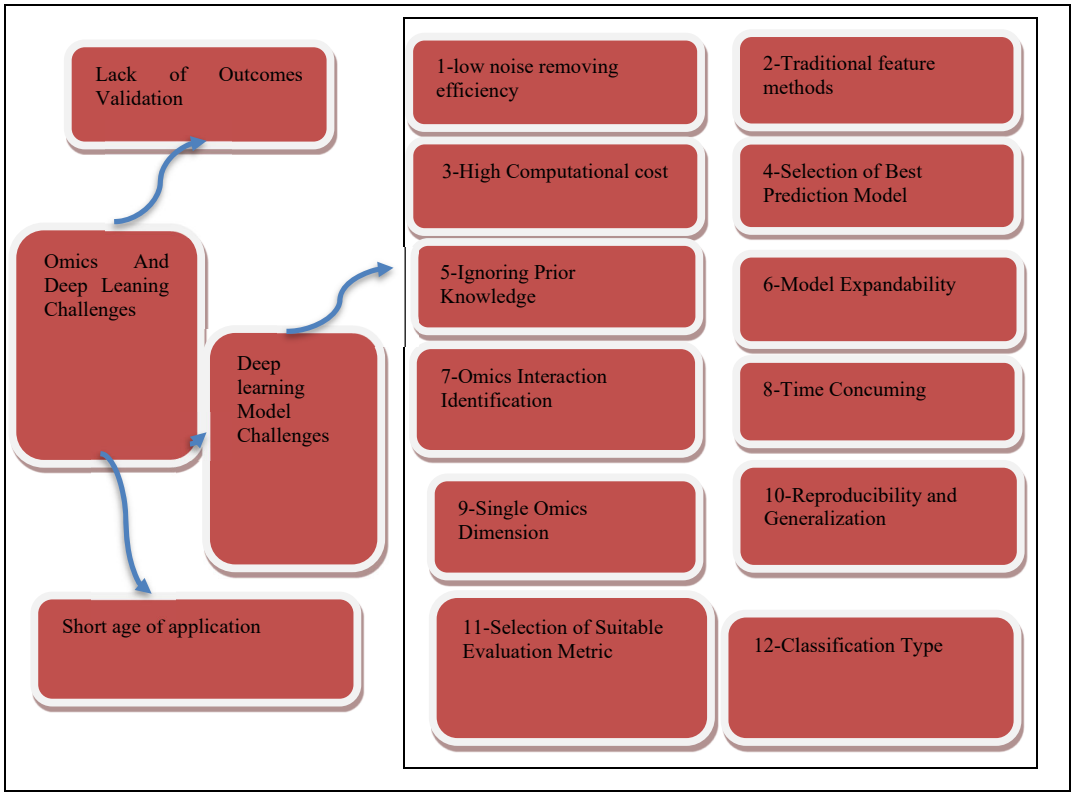

Figure 3. Omics and DL Model Challenges

To overcome all these challenges, researchers have developed many techniques for deep learning that they will review in this section. Considering the high incidence of breast cancer, distinguishing the inherent characteristics of the different subtypes is essential to identifying the underlying causes. By effectively integrating the multi-omics data available, it is possible to enhance the precision in identifying different subtypes of BC. In research by [41], the DNN model DeepMO based on multi-omics data—was offered to identify breast cancer subtypes. TCGA collected three types of omics data: mRNA, DNA methylation, and CNV data. after feature selection and data preparation, including an encoding and a classification subnetwork. The accuracy and AUC of the binary classification results obtained via DeepMO, which utilises multi-omics data, outperform other approaches.

Moreover, DeepMO demonstrated higher prediction accuracy on multi-classification when compared to other methods using single and multi-omics data. Furthermore, they verified how feature selection affects DeepMO. Any sickness detected early enough can be cured with minimal human effort. The computer-aided diagnostic (CAD) is an automated help for practitioners that produces accurate data to assess the severity of diseases according to the study done by [42], they proposed an approach that used (DNN) as a classifier model and recursive feature elimination (RFE) to pick features. DNN, with numerous layers of processing, outperformed SVM in classification rate. As a result, the researchers employed deep learning to classify hyperspectral data. A DNN is used to categorize the breast cancer data, using many processing layers. The UCI repository's (WBCD) was used to test the system. The dataset was partitioned into many train-test split subsets. The system's performance is evaluated via the assessment of accuracy, sensitivity, specificity, precision, and recall. The accuracy achieved was 98.620% according to the findings.

The researchers in study [43] present BioSurv, a system that uses DL and, ML techniques to predict cancer survival, and detect biomarkers. The researchers analyzed the multi-omics data, including DNA methylation, miRNA, CNV, and mRNA, using BRCA and LUAD data. The collected dataset is statistically tested for feature selection using the random spatial local best CSO (RSLBCSO) approach. To identify prognostic indicators, KEGG and survival studies are then conducted. Fifteen LUAD and thirteen BRCA poor prognostic markers were identified. A Bayesian optimised DNN is used to predict cancer survival, and it has a high accuracy of 90.0% for BRCA and 91.0% for LUAD. BioSurv's poor prognostic predictor performance in triple-negative breast cancer patients is one of its limitations (TNBC).

They conducted the first study[44] that used deep autoencoding and four-omics data to build a robust survival prediction model, and the results demonstrate that the method is beneficial for predicting LUAD prognosis. The most

frequent kind of cancer and the one with the highest death rate is lung cancer, of which LUAD accounts for about forty percent. There is an urgent need for a lung cancer prognostic prediction model. Prior research on LUAD prognosis only used single-omics information, such mRNA or miRNA. In order to achieve this, they suggested using a deep learning-based autoencoding method to combine four-omics data, including DNA methylation, mRNA, miRNA, copy number variations, and survival (P = 4.08e-09). This allowed them to create an autoencoder. The model that learned representative features to differentiate between the two ideal patient subgroups, showing a good (C-index = 0.65) and significant difference in survival.

The use of genomic data in cancer therapy is becoming more and more common. Strong cancer prognostic prediction requires many kinds of omics data since each type provides a singular perspective prone to bias and noise in the data. However, the large number of duplicate variables combined with the limited sample size makes it challenging to integrate multi-omics data adequately. Deep learning techniques have recently made it possible to combine multi-omics data and extract representative features using autoencoder. However, data noise might affect the produced model.

Furthermore, earlier research typically focused on specific cancer types rather than broad pan-cancer examinations. The DCAP was created by [45]. An Autoencoder-based Denoising Method for Combining Multi-omics Data to Provide Precise Cancer Prognosis Prediction the DCAP deep learning system is used in this work to predict cancer risk by using multi-omics data. Applying the DCAP methodology to 15 tumors from TCGA improved C-index values 6.5% over previous approaches. The research[46] employs the DNN technique. A deep learning system identifies primary and metastatic malignancies by analysing passenger mutation patterns. By using the PCAWG data set, which is the most extensive compilation of primary cancer whole genomes that have been handled consistently to date, they developed a supervised machine-learning system capable of accurately differentiating among 24 primary tumour types solely based on characteristics extracted from DNA sequencing. When implemented in a cross-validation setting, the system demonstrated an overall accuracy of 91.0%. Notably, out of 24 distinct tumour types, 20 attained an F1 score of 0.83 or higher. Upon evaluating the likelihood ratings of the tumor-type predictions, the precise forecast rated within the uppermost three positions 98.0% of the time.

A primary objective of modern bioinformatics is to create hybrid models that can analyze gene expression data and provide diagnosis methods for various diseases by research [47] The researchers combine a RF classifier, CNN, inductive spectrum clustering approach, and other voting techniques to solve this issue by determining the patient's condition at the end. For gene expression profiles, they first use the spectral clustering technique. In the first phase, they use the spectral clustering to analyse gene expression patterns. This involves calculating internal and external and equilibrium clustering quality criteria while employing inductive approaches to achieve objective clustering. This leads to clusters of DNA expression patterns that are differently expressed and mutually connected. The examined objects having gene expression levels in the designated clusters as characteristics are identified in the second step using a CNN and RF classifier. When both classifiers were employed, the simulation results demonstrated how successful and accurate the recommended approach was in detecting objects. However, the CNN showed a substantially greater data processing performance than the RF technique because of its significantly shorter processing time.

Cancer is a very intricate illness, and typically, each kind of cancer consists of several subcategories. Multi-omics data may provide a broader range of biological information to identify and uncover different forms of cancer. Current unsupervised cancer subtyping methods, however, are unable to acquire a complete understanding of both common and particular information from multi-omics data. Developed by [48], a unique shared and particular representation learning approach for multi-omics data clustering and cancer subtyping (MOCSS). Consistent and distinct information may be efficiently mined and used by the proposed MOCSS in multi-omics data. MOCSS is not without its restrictions. The TCGA has a variety of malignancies with varying sample sizes. For instance, the LUAD sample size is less than other cancer datasets, which might lead to statistical bias. It may be difficult to distinguish between Luminal A & Luminal B subtypes in BRCA due to ambiguity in the categorization of Luminal A and Luminal B based on gene expression profile criteria.

In addition, [49] validate the significance of the primary prognostic variables identified by MOCSS for LUAD, which offered a deep-learning method. Under a COX hazards model, three methods were used for simultaneous data integration and estimation: regular autoencoders, penalized principal component analysis (PPCA), and hyper-parameter optimised autoencoders (HPOAE). HPOAE showed a higher capacity to identify genes linked with survival subgroups in colon cancer patients. The fact that this research only included methylation of DNA, RNA-seq, and miRNA-seq results from 368 TCGA cases and excluded imaging data on cancer tissue and clinical features of the patients from the optimized autoencoder was a major drawback. Bigger samples could provide better outcomes.

Furthermore, identifying cancer driver genes is crucial in precision oncology research since it aids in understanding cancer genesis and progression. Currently, most computational approaches used to discover cancer

driver genes consider directed gene regulatory networks (GRNs) as undirected gene-gene relationship networks or rely on protein-protein interaction (PPI) networks as their main approach. This led to the loss of unique structure regulatory information in the directed GRNs, which affected the identification of cancer driver genes. Using a combination of directed graph convolutional network (DGCN) and multilayer perceptron (MLP). [50] introduced a novel approach (called DGMP) for finding cancer driver genes based on pan-cancer multi-omics data (i.e., gene expression, mutation, CNV, and DNA methylation). DGMP is a semi-supervised training method to separate cancer-driver genes from non-cancer genes. The model is built upon the DGCN and MLP architectures. Information was extracted from the TCGA database. DGMP searches for significantly altered cancer driver genes and identifies driver genes that are engaged in gene regulatory networks (GRN) with other cancer genes. Furthermore, it predicts driver genes based on other changes, such as differential expression and abnormal DNA methylation.

The results obtained from the analysis of three networks, namely DawnNet, KEGG, and RegNetwork, indicate that DGMP outperforms other state-of-the-art techniques in identifying cancer-driver genes. It is best to avoid overfitting while creating this model. Cancer subtype identification can help researchers comprehend hidden genetic pathways, increase diagnosis accuracy, and improve clinical therapy. Created MODEC according to a study [51], an unsupervised clustering algorithm that does not rely on prior information (MODEC). MODEC is a DL-powered clustering technique that iteratively generates cluster centroids and assigns cluster labels to each sample by minimizing the Kullback-Leibler divergence loss. Six available cancer datasets are available in the Using TCGA database. MODEC beat eight competing methods in terms of subtype correctness and reliability. MODEC was highly competitive in identifying survival patterns and major clinical traits. MODEC finds it difficult to expedite algorithm execution and manage incomplete datasets. The goal of the DL-TODA programmed, as proposed by [52]involves using a deep learning model trained on a dataset consisting of over 3000 distinct bacterial species to categorise metagenomic data. The CNN architecture built explicitly for computer vision applications, was used to model features that are unique to each species. DL-TODA demonstrated a high level of accuracy in classifying around 75.0% of the reads using synthetic testing data derived from 2454 chromosomes from 639 species. DL-TODA achieves a classification accuracy above 98% at taxonomic ranks higher than the genus level, making it comparable to two advanced taxonomic classification tools: Centrifuge and Kraken2. DL-TODA achieved a species-level accuracy of 97.0% on the same test set, surpassing the accuracies of 93.0% recorded by Kraken2 and 85.0% acquired by Centrifuge. The CRC CMSs are classified using an attention-based multi-instance learning (MIL) CNN model, which utilizes the whole-slide image and TCGA data.

To detect breast cancer, a hybrid DL system with decision-level fusion was described. By study [53] The framework used GRU and LSTM as classifiers. A CNN architecture is used to extract features. Proposing and using two distinct classifiers, a hybrid deep learning model is implemented to facilitate decision-making based on information from several sources. LSTM and GRU are classifiers for multi-omics data, including medical, gene expression, and CNV data. The choice of fusion model, which combines (LSTM, and GRU) achieves the greatest accuracy rate = 98.0%., while LSTM has an accuracy of 97.0% and GRU has an accuracy of 97.5%. Using data from several sources to inform decisions is another feature of this paradigm. DL classifiers (GRU and LSTM) were applied to predict BC survival based on the stacked features created by concatenating the gathered information. This model has a few shortcomings. First, the results of the experiments were limited in scope, which may have affected the generalizability of the conclusions. The second constraint is that only one ensemble technique can be used, as suggested by[54]. two predictive models for personalised therapy based-on DL techniques and multi-omics data: a model to identify a patient's BC subtype, and a framework for drug response prediction, DCNN-DR. They developed the first model via omics-data and clinical data from patients with BC. The second model was developed using omics data from BC cell types. Regarding prediction performance, the proposed models use late integration techniques and outperform previous approaches.

However, learning from a continuously changing function helps NN models; this may mimic a more refined relationship between changes to input and output. The DCNN-DR deficiency was found, and some patients' luminal A & B subtypes were misclassified. Furthermore, more clinical trials are required before clinicians can confidently embrace the models. The study by [55] proposed the cascading Deep Forest ensemble model. The cascading Deep Forest model combines the best features of deep NN with ensemble models. The cascading Deep Forest learns class distribution features by creating decision tree-based forests while monitoring the input. The METABRIC dataset was used in the present study. The results' accuracy was 83.45% for five subtypes and 77.55% for ten subtypes. Deep Forest models are well-known for resisting noise and outliers in data and working well with imbalanced training sets. Building and fine-tuning a Cascade Deep Forest model can be difficult and time-consuming, particularly if you have a large amount of multi-omics data. The OmiEmbed. a unified multi-task deep learning framework for multi-omics data, was proposed by [56] Deep embedding and downstream task modules are used in the system to extract biomedical information from high-dimensional omics data. The BTM dataset used in this investigation is available

from GEO. OmiEmbed can help you reduce dimensionality, integrate multi-omics, classify tumors, reconstitute phenotypic features, and predict survival. OmiEmbed defeated the most advanced techniques in all three kinds of downstream tasks: regression, classification, and survival predicting. To improve interpretability, the OmiEmbed framework is required. The study by [57] suggested research analyses of the effectiveness of several DL autoencoders for cancer subtype identification. Despite the fact that the performance of various autoencoders differed among datasets, they achieved superior performance compared to traditional data fusion methods involving PCA, kernel PCA, and sparse PCA. The denoising autoencoder is designed to restore the original input by reconstructing it from a corrupted version. Autoencoders performed the best at detecting subtypes. The decoder architecture remains a dark box, with no way to see how the various input components contribute. Autoencoder vanilla: The vanilla autoencoder has regularised variants such as denoising, sparse, and variational autoencoders. The vanilla autoencoder is a basic sort of autoencoder that typically consists of just one layer of the encoder and decoder. The researchers examined data from the database maintained by the TCGA in their investigation. Although the vanilla autoencoder is simple, it is vulnerable to over-fitting.

GDP (Group lass regularised) was created by Researchers according to the study [58] Cancer Prognosis Using DL, GDP is the first technique to merge group lasso regularisation, DL framework, and the Cox model for survival prediction. GDP is a computational survival prediction tool that uses clinical and multi-omics data from (TCGA) project, both simulated and actual data. In simulated data, their findings validated the group prior information's importance in model regularisation. They demonstrated that when group prior knowledge was provided, group lasso obtained higher prediction accuracy than ordinary lasso regularisation. Although it was previously shown that clinical data may have a bigger effect on cancer survival predictions than molecular information, they didn't have any previous experience to employ in the clinical characteristics obtained from group lasso.

The researchers in [59] created a novel category of autoencoders known as Sparsely Connected Autoencoders (SCA), because they provide regulated connections between the data input layer and the decoder's component, which is a benefit. An autoencoder is a highly effective technique for data compression and noise reduction. The benefit of this new architecture is that the decoder model is no longer a black box and can be utilized to display new biologically interesting aspects from single-cell data. One limitation of this study is the absence of specific results. DeepProg is an innovative and generic computational framework proposed according to [60]. It handles different kinds of omics data sets by combining DL (autoencoder) and ML algorithms for survival prediction. It correctly predicts patient survival subgroups using multi-omics data. DeepProg is strongly predictive, as indicated by C-indices of 0.73-0.80 in two liver cancer datasets and 0.68-0.73 in five breast cancer datasets. DeepProg requires thorough testing in real-world patient populations. The most recent studies using DL on multi-omics data are summarized in Table 2.

Table 2. The summary on utilization of DL techniques on Omics data

Author(s)/ Year	Techniques used	The summary	Dataset used	The advantages	The limitaions
1- Yuqi Lin et al. (2020) [41]	Following preprocessing of data and selection of features, an encoding and a classification subnetwork comprise the (DNN), which was fed various types of omics data. It's DeepMO.	In this study, the researchers used DeepMO, a model that combines DNN with multi-omics data, to classify different subtypes of BC.	There are three forms of omics data: mRNA data, DNA methylation data (TCGA), and genomic data.	1- DeepMO's multi-omics-based binary classification findings beat other algorithms in terms of accuracy and (AUC). 2-DeepMO also performed better in multi-classification prediction.	1-DeepMO is a closed-loop model. This makes understanding how the model produces predictions challenging.
2- S. Karthik et al. (2018)[42]	DNN	DNN use a complicated network model to follow the structure of a standard (ANN). The BC data is classified using DNN with a lot of processing.	The system was tested using the (WBCD) from the UCI library.	1- The achieved accuracy was 98.62%, surpassing that of other cutting-edge techniques. 2-It supports all algorithms, including, unsupervised, supervised, semi-supervised, and reinforcement learning. 3-As predicted, this model excels and achieves a promising 98.62% for an 80-20% partition split. Furthermore, for 70-30% and 60-40% splits, this approach achieves 97.66 and 96.88%, respectively.	1-Because this network includes numerous layers and a large number of inner nodes, it is computationally costly yet produces promising results after training the model. 2-The training time of the algorithm is a restriction of this system because it has thoroughly trained the neural network.
3- Arwinder Dhillon et al. (2023)[43]	BioSurv	The BioSurv framework, which is based on random spatial local best CSO and Bayesian optimised DNN, is presented in this article for biomarker identification and cancer survival prediction using miRNA, mRNA, CNV, and DM.	Patients with BRCA and LUAD were studied in a clinical setting. The dataset may be obtained via the Linked Omics Portal. LinkedOmics :: Login from TCGA	1- On integrated multi-omics data, the BioSurv worked well, achieving the desired levels of accuracy, sensitivity, specificity, and precision, as measured by the AUC and c-index values for BRCA samples (91.60 percent, 88.20 percent, 89.12 percent, 90.01%, 90%, and 0.69), and LUAD samples (90.1%, 87.5%, 88.3%, 86.4%, and 0.67 percent). 2-DNN, XGBoost, RF, SVM, GBM, KNN, and DT	1-Unfavorable predictive indicator in patients with TNBC 2- When comparing BioSurv against DNN in LUAD in a single instance, the Wilcoxon signed-rank test does not find a significant difference. Its p-value of 0.06 indicates that it does not meet the test requirement of having a p-value of 0.05 or below. 3- The addition of new feature extraction approaches, as well as the

Author(s)/ Year	Techniques used	The summary	Dataset used	The advantages	The limitaions
				results were compared against a variety of cutting-edge models. BioSurv outperforms all other models, with an improvement in accuracy of roughly 0.08 and 3% for BRCA and LUAD cancers, respectively. 3-A comparison of single and multi-omics is presented. BioSurv performed well, with AUC values of 90.0% for BRCA and 87.0% for LUAD.	examination of other DL models, show promise for improving BioSurv's effectiveness in predicting cancer survival.
4- Lee T-Yiet et al. (2020)[44]	deep autoencoding and four-omics data	They used four-omics data, mRNA, miRNA, methylation, and CNV, to significantly differentiate the survival differences among the two survival groups of lung adenocarcinoma patients	TCGA 4-omics data	1-Following the combination of four-omics data, mRNA, miRNA, methylation, and CNV, the log-rank P value was calculated to be 4.08e-09, and a C-index of 0.65 was obtained between the two survival categories. 2- Autoencoders identify key characteristics from data automatically, facilitating the detection of hidden patterns and biomarkers that would not be apparent using standard research approaches.	1-focused on particular cancer types rather than doing pan-cancer testing 2-multi-omics data has a large dimensionality, especially when all four data types are considered at the same time. This high dimensionality can cause computational difficulties, longer training durations, and an increased risk of overfitting.
5- H. Chai et al. (2021) [45]	DCAP, which stands for "A framework to integrate multi-omics data by Denoising Autoencoder for Accurate cancer prognosis prediction"	They used a DL framework called DCAP in this study to incorporate multi-omics data for cancer risk estimate.	TCGA https://tcga-data.nci.nih.gov/tcga/	1-improve C-index values by 6.5% on average over earlier techniques. 2- The models demonstrated an average C-index of 0.627. As an example, the BC prognosis prediction system underwent separate testing on three datasets from the GEO. The findings showed that it could differentiate between dangerous and low-risk patients, with p-	1- C-index values of 15 TCGA carcinomas were used to compare methods. The AE-Cox model ("The Cox model using compressed features by Autoencoder") outperformed the suggested model in diagnosing colon cancer. DCAPC-index=0.622, AE-CoxC index=0.628 2- The abundance of censored samples in the dataset diminished the precision of predicting

Author(s)/ Year	Techniques used	The summary	Dataset used	The advantages	The limitaions
				values less than 0.05 and a C-index more than 0.6. 3- A reliable and accurate foundation for combining data from many omics studies to predict cancer outcomes.	cancer outcomes. Estimated censored rates for STAD (The stomach) and LUAD (The lungs), respectively 60.70% and 64.50%, correspondingly. The method's effectiveness was hindered by the elevated rates of censorship.
6- Wei Jiao et al. (2020) [46]	DNN	The detection of both primary and metastatic tumors is accomplished using a deep learning system via the use of passenger mutation trends.	PCAWG data set dcc.icgc.org/releases/PCAWG/	1- In the setting of cross-validation, the system achieved an overall accuracy of 91.0%. Out of the 24 tumor types, 20 of them achieved an F1 score equal 0.83 or higher. 2- The classifier achieved prediction accuracies of 88% & 83% for both primary and metastatic tumors, respectively, when evaluated using external validation datasets. 3- Out of the twenty-four kinds of tumors, twenty-one got F1 scores more than 0.80. This includes eight out of the nine types for RF models that were constructed using single-feature categories.	1- The CNS-PiloAstro tumor type had a mean F1 score of 0.79 across 10 independents trained DNN approach. The Lung-AdenoCA tumor type achieved an F1 score= 0.77, while the Stomach-AdenoCA tumor type achieved an F1 score equal 0.67, which were the lowest among the tested tumor types. 2- Tumor types with fewer than one hundred occurrences in the dataset became more prone to erroneous predictions, while tumor types with a larger number of samples performed better.
7-Sergii Babichev et al. (2023)[47]	hybrid model	In this work, hybrid models were created to interpret gene expression data using an inductive spectral clustering technique, RF classifier, CNN, and an alternative voting approach to make a final conclusion regarding the patient's status.	The gene expression datasets https://www.cancer.gov/about-nci/organization/ccg/	1- When both classifiers were employed, the simulation results revealed that the suggested approach was extremely successful, reaching high accuracy in object detection. 2- Due to its vastly lower processing time, the CNN algorithm exhibited significantly greater data processing efficiency than the RF method. 3- enhancing the recognition of objects accuracy and objectivity by	1-Adrenocortical carcinoma—ACC 79 sample 2-the hybrid model may be complex.

Author(s)/ Year	Techniques used	The summary	Dataset used	The advantages	The limitaions
				parallelizing the data analysis method.	
8- Yuxin Chen et al. (2023)[48]	MOCSS	MOCSS (shared and specific representation learning) is a novel approach for multi-omics data clustering and, cancer subtyping.	TCGA https://github.com/ChenyuxinXMU/MOCSS/	1- MOCSS beats current modern multi-omics clustering techniques in the field of clustering performance, according to experimental data. 2- MOCSS can distinguish physiologically relevant clusters, it can adapt to a broad variety of omics data, and it is able to be used to effectively subtype tumours.	1- The TCGA exhibits variability in sample sizes across various cancer types. Notably, the sample size for LUAD is comparatively less compared to other cancer datasets, potentially leading to analytical bias. 2- The classification of the luminal A & Luminal B subtypes in BRCA might be confusing due to gene expression profiling criteria, making it difficult to differentiate between the two subtypes.
9- S. Salimy et al. (2023)[49]	a deep-learning method	Three methods were used for the simultaneous integration and estimation of data within a COX hazards framework, (HPOAE), conventional autoencoders, and penalised principal component analysis (PPCA).	TCGA	1- The HPOAE model considerably improved the model's mean squared error (MSE), obtaining a low MSE of 0.015. 2- In patients with colon cancer, HPOAE demonstrated a superior ability to identify survival subgroups and the genes linked with them. 3- One of the challenges is the high processing complexity caused by high data dimensionality. The researchers used a deep (HPOAE) model to decrease the dimensionality	1- This study was flawed in its omission of clinical information on individuals and imaging information on cancer tissue in the optimized autoencoder. Additionally, it only used DNA methylation, RNA-seq, and miRNA-seq information collected from 368 samples.

Author(s)/ Year	Techniques used	The summary	Dataset used	The advantages	The limitaions
				of the input data. This was then followed by a based-on rules C5 model to identify genes and situations that are linked with survival.	
10-Shao-Wu Zhang et al. (2022)[50]	a novel method of DGMP	DGMP, a novel approach for detecting tumours driver genes that combines (DGCN) and (MLP), is proposed by the researchers. DGMP is a semi-supervised training system that uses DGCN and MLP to identify cancer-driver genes from non-cancer driver genes.	(TCGA) database	1- The primary objective of DGMP is to identify cancer driver genes that have undergone significant modifications Additionally, DGMP aims to uncover driver genes that are engaged in gene regulatory networks (GRN) and are also associated with additional cancer-related genes. 2- The findings from three networks, namely DawnNet, KEGG, and RegNetwork, demonstrate that DGMP surpasses other cutting-edge methods in identifying cancer driver genes. 3- By using the DGCN model, DGMP fully utilises gene regulation information.	1-Overfitting is a possibility. 2-time-consuming.

Author(s)/ Year	Techniques used	The summary	Dataset used	The advantages	The limitaions
11- Yanting Zhang and Hisanori Kiryu. (2022)[51]	MODEC	MODEC is a totally unsupervised clustering algorithm that does not rely on prior information. A DL-based clustering module is used by MODEC in order to build cluster centroids and assign cluster labels to each sample on a recurring basis. This is accomplished by minimizing the Kullback-Leibler divergence loss wherever possible.	The Cancer Genome Atlas database contains six public cancer datasets. https://www.cancer.gov/tcga	1- outperformed eight other approaches in terms of subtyping accuracy and dependability. 2- There is no need to choose complicated parameters with MODEC, so it is easy to use and useful for troubleshooting in the real world. 3- After merging omics data into a common low-dimensional subspace with a customized rank, MODEC repeatedly determines cluster centers and labels each sample using manifold optimization and deep learning.	1- Handle incomplete datasets and accelerate algorithm execution. 2- The MODEC method has its limitations, such as its inability to process datasets with different views that include the same samples. 3- In terms of execution time, MODEC does not have a significant benefit. 4- A typical issue is that manifold optimisation and deep-learning models are difficult to comprehend.
12- Cres, C.M. et al. (2023)[52]	DL-TODA	The goal of the DL-TODA software is to classify metagenomic data using a DL-based model that has been trained on more than three thousand different bacterial species. DL-TODA is a CNN-based DL model that can identify brief metagenomic data from more than three thousand different bacterial species. The (DNN) AlexNet, an effective CNN in computer vision, is used to train DL-TODA.	Bacterial genomes are accessible via the NCBI Reference Sequence database. https://github.com/zhanglab/dl-toda (Accessed on 6, October, 2023).	1- The capacity of DL-TODA to confidently categorise around 75.0 Percentage of the reads was proved using simulated test results obtained from 2454 genomes comprising 639 species. 2 -Comparing DL-TODA against two state-of-the-art taxonomic classification tools, Centrifuge and Kraken2, it showed a classification accuracy of over 98.0 percent at taxonomic levels higher than the genus level. DL-TODA outperformed Kraken2 (93.0% accuracy) and Centrifuge (85.0%) on the same test set, reaching an accuracy of 97.0 percent at the species level. 3- DL-TODA allows for seamless training with additional data without the need to re-analyze the current training sets. This	1- In DL-TODA, macro averages were lower than micro averages. A paucity of training data might be one cause for DL-TODA's low performance on certain species. 2- To speed up training and testing processes, a solution for lowering the size and number of parameters in DL-TODA is required.

Author(s)/ Year	Techniques used	The summary	Dataset used	The advantages	The limitaions
				allows for easy updating of the model using recently identified genomes.	
13- Othman, N.A. et al. (2023)[53]	For breast cancer diagnosis, a hybrid DL approach with decision-level fusion was described. The framework used (LSTM) and (GRU) as classifiers. A CNN architecture is used for feature extraction.	To facilitate decision making based on data from many sources, a hybrid DL model is described and implemented using two independent classifiers. By merging several omics data sets (clinical, gene expression, and CNV),	from the METABRIC dataset https://github.com/USTC-HIlab/MDNNMD (Accessed on 6 October 2023).	1- Both GRU and LSTM are implemented as classifiers. The reliability of LSTM is 97.0 percentage, that of GRU is 97.50 percentage, and that of decision fusion (LSTM + GRU) is 98.0 percentage. 2- Decisions may be made based on data from several sources. 3- The ACC of the LSTM model was 97.0 percentage, the Pre was 98.0%, the Sen was 98.60 percentage, the MCC was 92.0 percentage, and the AUC was 95.3, while the GRU model had an ACC of 97.50 percentage, the Pre was 98.0 percentage, the Sen was 99.20 percentage, the MCC was 93.0 percentage, and the AUC was 96.0. Using the voting classifier, the proposed model achieved an ACC of 98.0 percentage, Pre of 99.0 percentage, Sen of 99.2%, MCC of 93.60%, and AUC of 98.20, which may reduce pathologist errors and efforts throughout the clinical process.	1- The first limitation is that our findings are not generalizable because our tests were done on a limited dataset. 2- The second constraint is that only one ensemble technique can be used.

Author(s)/ Year	Techniques used	The summary	Dataset used	The advantages	The limitaions
14- Khan, D. et al. (2023)[54]	DCNN-DR, a deep convolution neural network-drug response method	They put out a technique to determine a patient's BC subtype and a model for predicting pharmaceutical response, DCNN-DR, both using DL approaches and multi-omics data. These models would be used in personalised medicine. Both models were trained using omics data from BC cell lines; the first one was built using clinical and omics data from BC patients.	(CCLE) and (TCGA) https://github.com/Deebamajeed/BC-Subtype-DR(Accessed on 6 March 2022)	1-The suggested models employ late integration approaches and have somewhat higher prediction performance than existing methods. 2- The model successfully detects Her2+, basal-like, and unknown subtypes are all possible.	1- Certain luminal A & luminal B subtypes were incorrectly categorized. 2- More clinical studies are required before physicians may confidently adopt the models.
15- El-Nabawy, A. et al. (2021)[55]	cascade Deep Forest ensemble model	A combination of DNN and ensemble model strengths, the Deep Forest model uses a cascade effect. The cascading Deep Forest learns properties of class distribution by creating forests based on decision trees while it monitors the input.	The METABRIC dataset	1. Deep Forest models provide improved interpretability and are useful for training sets that are unbalanced. 2-The acquired accuracy was 83.45% for 5 subtypes and 77.55% for 10 subtypes. 3- Gene expression data alone, when combined with the cascade Deep Forest classifier, achieves accuracy equivalent to previous systems utilising bigger processing capabilities in around just five seconds for Ten subtypes and 7 seconds for Five subtypes.	1-Creating and fine-tuning a Cascade Deep Forest model may be difficult and time-consuming, especially if you have a big amount of multi-omics data.
16- Zhang, X. et al. (2021)[56]	OmiEmbed zhangxiaoyu11/OmiEmbed: Multi-task deep learning framework for multi-omics data analysis (github.com)	a unified multi-task deep learning framework for multi-omics datasets the technology employs deep embedding and downstream task modules to glean biological insights from omics data with a high dimensionality.	GEO provides access to the BTM dataset. https://www.ncbi.nlm/ https://xenabrowser.net/datapages/	1- Helping with reducing dimensionality, integrating multi-omics, classifying cancer types, reconstructing phenotypic features, and predicting survival. 2- OmiEmbed is open source, well-organized, and readily adaptable to different customized input data, network architecture, and downstream functions, with the ability to assist more accurate and	1- Improve the interpretability of the OmiEmbed framework.

Author(s)/ Year	Techniques used	The summary	Dataset used	The advantages	The limitaions
				specialized healthcare decision making.	
17- Franco, E.F. et al. (2021) [57]	DL autoencoders.	The purpose of this study was to compare the efficacy of several deep learning autoencoders in identifying tumour subtypes.	TCGA database edianfranklin/autoencoder_for_cancer_subtype (github.com) (Accessed on 7 October 2023)	1-even though the performance of various autoencoders varied between datasets, they beat conventional data fusion methods including PCA, kernel PCA, plus sparse PCA. 2- Using a defective copy as a starting point, the autoencoder denoising method restores the original input. 3. Autoencoders performed the best in detecting subtypes.	1- The outputs of various autoencoders varied between datasets, but were generally vanilla and variational. 2- Because it is impossible to see how each input component contributed, the decoder model remains a mystery.
	1- vanilla autoencoder: Regularised variations of the vanilla autoencoder include denoising autoencoder, sparse autoencoder, and variational autoencoder.	The vanilla autoencoder is the most basic sort of autoencoder, containing just one layer of encoder and decoder.	TCGA database	1- A simple type of autoencoder 2- The vanilla encoder learns nonlinear properties from data. This isn't possible with linear feature derivation algorithms like PCA. 3- Even when the autoencoder includes a high number of hidden units (Sparse autoencoder), the sparsity penalty Ω (h) assists in understanding the essential characteristics of data.	1-Although the vanilla autoencoder is straightforward, there is a substantial risk of overfitting.

Author(s)/ Year	Techniques used	The summary	Dataset used	The advantages	The limitaions
18- Gangcai Xie et al. (2019)[58]	A GDP (Group lass regularized) was developed. Cancer Prognosis Using DL GDP/simulation at master · WGLab/GDP (github.com)	GDP is a computational technique for predicting survival based on clinical and multi-omics data.	TCGA project, both simulated and actual. GDAC in its entirety https://gdac.broadinstitute.org/ TCGA data portal https://cancergenome.nih.gov/	1- GDP might have a substantially greater c-index without regularization either lasso and a naïve approach. 2- across a real-world investigation with large-scale omics data sets, GDP demonstrated superior performance compared to other methods across a range of tumour types, including glioblastoma multiforme, renal clear cell carcinoma, and bladder urothelial carcinoma.	1- They didn't have any group-level knowledge to apply via group-lasso to clinical characteristics, and research suggests that clinical data may be more important than molecular data for estimating cancer survival rates.
19- Alessandri, L. et al. (2021)[59]	(SCA)	They developed a new kind of autoencoder called as (SCA), This has the benefit of enabling a regulated interaction between the input layer and the decoder module, which is a significant advantage.	All of the material and dataset used in this study may be found on the website figshare.com: https://figshare.com/projects/Sparsely_Connected_Autoencoders_a_multi-purpose_tool_for_single_cell_OMICs_analysis/123226, accessed on 10 Nove. 2021.	1- Data compression and noise reduction are two areas where autoencoders shine. 2- In this new architecture, the decoder model is not anymore a black box, and it may be used to present new physiologically important elements from single cell data. 3- The autoencoder's exceptional ability to retain just the relevant component of a signal may help discern between actual variations through subpopulations of cells and clustered overfitting.	1- Inadequate Specific Results

Author(s)/ Year	Techniques used	The summary	Dataset used	The advantages	The limitaions
20- Olivier B. Poirion et al. (2021)[60]	DeepProg lanagarmire/DeepProg: Deep-Learning framework for multi-omic and survival data integration (github.com)	DeepProg is a novel and generic computing framework that handles various kinds of omics data sets for survival prediction by combining DL techniques (autoencoder) and ML methods.	Data used from the omics: RNA-Seq, microRNA-Seq, and DNA methylation. add data · lanagarmire/DeepProg@873b698 (github.com)	1- DeepProg is strongly predictive, as indicated by C-indices of 0.73-0.80 in two liver cancer datasets and 0.68-0.73 in five cancers of women datasets. 2- It uses boosting methods to increase the final model's resilience by mixing weaker models from different subsets of the original data.	1- DeepProg need thorough testing in real-world patient populations to. 2- Deep learning models, in particular, can be computationally costly and may necessitate significant computer resources, thus limiting their accessibility to some researchers and healthcare institutions.

4. OTHER METHODS DEALT WITH MULTI-OMICS AND SINGLE-OMICS.

Data analytics is being used extensively to support biomedical research across several fields, particularly concerning the most critical clinical diseases like cancer. In particular, methods based on bioinformatics have been used to characterize illnesses by their molecular features. In recent years, there has been a proliferation of cancer research studies that use single- and multi-omics data. Gene expression, DNA methylation, and microRNAs are just a few examples of data types used in single-omics studies. Despite this, a major fraction of the literature describes gene expression studies employing microarray datasets. Single-omics data include many characteristics but only a small number of samples. It's clear that cell biology in health and illness is being revolutionized by the simultaneous analysis of a single cell's genome, epigenome, transcriptome, proteome, or metabolome. In less than a decade, astonishing technological breakthroughs have revealed critical new insights into the interplay of internal and intercellular molecular processes that control growth, function, and disease. The billions of cells that make up humans and other eukaryotic organisms fall into many different kinds and functional cell states according to traits inherent to the cell or external to it. From the genome and epigenome to the transcriptome, proteome, and metabolome, and back again, there is a maze of interconnected molecular hierarchies within a cell. Direct physical interactions (like receptor-ligand interactions) and signaling molecules secreted by one cell that act through receptors on distant cells (like morphogen signaling pathways) are extrinsic factors that impact a cell's functional state. Other variables in the microenvironment, such as chemical compound gradients, are also considerations. So, to study how multicellular creatures develop, age, and get sick, it needs one-cell and spatial multi-omics methods, also called multimodal omics approaches.

Another study by [61] study of a single cell's genome, epigenome, transcriptome, proteome, and metabolome is transforming our knowledge of cell biology in health and illness. In less than a decade, extraordinary technological developments have provided vital new insights into the interaction of intracellular and intercellular molecular processes that drive development, function, and disease.

4.1. Current Data Integration Challenges.

Several obstacles to computational data integration remain despite substantial research. These methodologies implicitly assume the expected commonality of collected cellular states between studies and modes. When examining minor differences in the physiological state across several experimental conditions, horizontal data integration across phases may result in overcorrection of true biological variation [62]. Perhaps the most challenging difficulty is data integration. a detailed examination of the challenges in single and multi-omics statistical evaluation of cancer data, emphasizing the choice of genes and data integration approaches. According to the researchers [2], a single-omics study typically ignores the intricate nature of the ecosystem of biological events causing the illness. Consequently, they give limited and perhaps untrustworthy information about disease processes. proposed the SARSA study[4], which resulted in new hybrid cancer detection systems. The study looks at several types of cancer by analyzing, classifying, and processing a multi-omics dataset in the fog cloud network. The study proposes policy schemes based on reinforcement learning state action reward state action (SARSA) based on SARSA on-policy workload learning and multi-omics workload learning enabled by reinforcement learning. The article introduces innovative SARSA on-policy reinforcement and multi-omics workload learning-enabled hybrid cancer detection methods. The system has many levels, including the gathering of clinical data via laboratory tests and medical procedures (such as biopsy, colonoscopy, and mammography), as well as the execution of operations inside the network's decentralized clinics that use omics-based approaches. SARSA reduces processing time by identifying clustering features and categorizing the system's many properties into distinct groups. The primary concern with SARSA is the system's substantial computational load.

The proposed TNM system by [63] . the condition of a lymph node The TNM classification, which includes the assessment of the primary tumor (T), regional lymph node status (N), and distant metastases (M), is widely used for cancer staging. TNM staging offers prognostic information on the patient's illness outcome after surgery by categorizing the resected tumor based on the laboratory's assessment. This helps in making decisions about further therapy. It verifies the existence or nonexistence of metastases at the first diagnosis. The TNM system is user-friendly, and the predictive accuracy across phases is commendable. The TNM classification cannot distinguish between 'favorable' and 'unfavorable' cancers within the same stage since approximately 20 percent of stage II patients may still succumb to recurrent illness. The parallel cat swarm optimization (PCSO) method according to the study [64] is a specialized optimization technique designed to tackle numerical optimization problems characterized by a small population size and a restricted number of iterations. By adding the orthogonal array from the Taguchi technique into the tracing mode operation of the PCSO method, the authors suggested an improved parallel cat swarm optimization (EPCSO) strategy according the study. Compared to earlier PSO-based approaches, the EPCSO method can attain a higher level of accuracy while simultaneously taking less time for processing than the PCSO method. Omics datasets (EPCSO) may also be used to diagnose cancer.

By study [65] proposed a new technique to CancerSEEK that coupled protein biomarkers with genetic biomarkers to enhance sensitivity but drastically reducing specificity in the whole process. It is possible that other cancer biomarkers, such as metabolites, mRNA transcripts, miRNAs, or methylation DNA sequences, might be included in the same manner in order to enhance the sensitivity and localization of cancer locations. Eighty-three percent of patients had cancer found in a few anatomic areas, according to CancerSEEK. For the detection of five different forms of cancer (ovary, liver, stomach, pancreatic, and esophagus), the sensitivities varied from 69 to 98%. In the case of healthy individuals, although in the context of a genuine cancer screening, there is a possibility that some individuals may be suffering from inflammatory or other disorders, which would lead to a greater percentage of false-positive findings than what was shown in the research.

A novel framework known as PROFILE (a technique known as "logical modeling") proposed by Researchers[66] Offers a technique for merging the analytical understanding of logical modeling with the integration of multi-omics data to create models that are relevant to patients. Adopting this technique will ultimately facilitate the use of logical modeling in precision medicine, enabling doctors to more readily choose pharmacological therapy tailored to individual patients. Before inclusion in the logical model, this data must undergo binarization or be scaled to a range between 0 and 1. This may be achieved by altering the node's activity, initial conditions, or state transition rates. Proliferation and apoptosis, two significant cancer phenotypes derived from personalized models, are physiologically coherent prognostic factors: individuals exhibiting strong growth and little apoptosis have a poor survival rate, whereas those with low proliferation and high apoptosis have a better prognosis.

InCRiS was proposed According to the research [67]. InCRiS is an artificial intelligence system designed to identify specific risk sub-pathway regions in cancer and then apply them to people with various types of cancer. The offered data and technique will be very beneficial for future studies on cancer causation and precision therapy. Using the TCGA database, InCRiS outcomes may provide insights into specific locations of sub-pathway dysfunction at the individual level. Offer more precise suggestions for cancer pathogenic pathways and targeted treatment. In order to maximize the number of samples in which the gene used for analysis was expressed, genes with over 20% of the missing expression data were removed from the expression profile. Nevertheless, this might lead to the elimination of several vital mutant genes. Consequently, less filtering might be deemed appropriate when applicable to other investigations.

The suggested methods according to based study [68] include particle-swarm optimization algorithms (PSO) and RNA sequencing of tumor-educated blood platelets (TEPs). Employ PSO and RNA-seq of tumor-educated platelets obtained from patients to produce RNA sets that differentiate individuals with non-small-cell lung cancer (NSCLC), including those in the beginning stages, from people in good health with inflammatory conditions. The data was collected from patients diagnosed with NSCLC. The TEP-based detection of both early and late-stage non-small-cell lung cancer was reliable, as demonstrated by the validation cohort accuracy of 518 late-stage cases. Swarm intelligence can also optimize the diagnostic readout of liquid biopsy BioSource's. This is supported because 88% of gene panels used to diagnose cancer from TEPs were selected using PSO. The ability to identify cancer using tumor-educated platelets (TEPs) is hardly affected by inflammatory conditions.

NEMO algorithm: Community-oriented Multi-Omics Clustering is an innovative method for clustering several omics data sets. The study by [69] propose that NEMO may be used to analyze incomplete datasets that include patients with missing data for certain omics variables, without data imputation. Employing multi-omics data to delineate cancer subtypes has the potential to enhance our comprehension of cancer and facilitate more accurate patient therapy. When all the data was included, NEMO yielded results similar to the top-performing multi-omics clustering algorithms out of the nine tested. Additionally, NEMO showed improvement when working with incomplete data. Every pair of samples in the incomplete data must have at least one common omic. This assumption is valid when just one omic was evaluated for all patients, which is often the situation with gene expression. The most recent studies using other methods dealt with multi- and single-omics, summarized in Table 3.

Table 3. The most recent studies using other methods dealt with multi- and single-omics.

Author(s)/ Year	Techniques used	The summary	Dataset used	The advantages	The limitaions
1-Momeni et al (2020) [2]	data integration methods.	This study offers a comprehensive examination of the challenges involved in analyzing cancer data using single and multi-omics approaches. It particularly focuses on the strategies used for gene selection and data integration.	microarray datasets	1- Single-omics data analysis is performed with the intention of determining which genes are the most important. 2- more dependable outcomes from multi-omics data	1- Single-omics data exhibit a high number of characteristics but a comparatively low number of samples. 2- (Single omics) often fail to consider the intricacies of the molecular processes that underlie disease. 3- give limited and possibly untrustworthy information regarding illness pathophysiology.
	1-Filter	Statistical approaches are frequently used in filtering procedures to choose characteristics with a minimal computational cost. These approaches make no use of a classifier or learning system.	Huge datasets	1-Scalable and quick 2-Unaffected by the classifier 3-More efficient than wrapper approaches	1-Lower accuracy as a result of the classifier not being considered 2-Ignores the connection between features/variables. 3-There may be redundancy.
	2- Wrapper	Wrapper methods use a predictor inside an opaque context. The effectiveness of the predictor is used as the objective function to assess the subset of chosen attributes.	microarray datasets	1- Communicate with the classifier 2- Interaction of features 3- Greater precision than filter techniques	1- Overfitting 2-Extreme computational complexity 3- Expensive procedure 4-Tendencies towards local optimisation

Author(s)/ Year	Techniques used	The summary	Dataset used	The advantages	The limitaions
	3- Embedded	The embedded approach is a strategy for selecting genes that occurs simultaneously with the learning process.	microarray datasets	1- Greater precision and efficiency than filter techniques 2- Has a lower computational complexity than wrapper approaches. 3- A greater emphasis on the interaction between characteristics	1-Dependent to classifier
	4-Ensemble approaches	An ensemble approach seeks to identify a collection of optimum subsets of features by using several feature selection processes, and then combines the outcomes obtained from these subsets.	microarray datasets	1- Less prone to overfitting 2- Greater scalability for large data sets 3-Stability	1- Complexity arises while attempting to comprehend the combination of classifiers.
	5- Hybrid Methods combining several methodologies for gene selection	a minimum of two different methods of gene selection Hybrid techniques include the integration of a filter, wrapper, embedding, or ensemble in a certain order. The hybrid approach incorporates the benefits that are often associated with each individual method.	microarray datasets	1-More effective than filtering techniques. 2- Lower proclivity to overfitting Reduced computational costs.	1- Classifier dependent. 2- Depending on how different gene selection techniques are combined.

Author(s)/ Year	Techniques used	The summary	Dataset used	The advantages	The limitaions
2-Mazin Abed Mohammed et al (2023) [4]	The study resulted in the development of innovative hybrid cancer detection methods. SARSA, which stands for reinforcement learning state action reward state action, is the basis for the policy schemes that are developed in this study.	Through the analysis, classification, and processing of a multi-omics dataset inside a fog cloud network, the study conducts an investigation into a number of different types of cancer. On the basis of SARSA's on-policy workload learning and multi-omics workload learning, which is enabled by reinforcement learning	clinical data, multi-omics dataset	1- Through the extraction of clustering characteristics and the enumeration of the numerous features of diverse classes within the system, the processing time may be reduced.	1-The system is undergoing massive processing.
3-Rejali et al (2023)[63]	TNM classification	TNM staging offers insights into the patient's illness prognosis by categorizing it after surgery, using the pathologist's assessment of the removed tumor. It helps in making decisions regarding further therapy and verifies if metastasis is present or absent at the time of diagnosis.	********	1- The evaluation of lymph node status has significant importance in the TNM method, a widely used method for cancer staging. 2- The TNM system is user-friendly and has a high level of prognostic accuracy across different stages.	1- The TNM classification lacks the ability to differentiate between favorable and unfavorable malignancies within the same stage, given that roughly twenty percent of stage II patients may still succumb to recurrent illness. 2- The predictive capacity of the TNM system is flawed and requires improvement. 3- The TNM staging approach lacks precision in predicting patient outcomes, especially in stages II and III.

Author(s)/ Year	Techniques used	The summary	Dataset used	The advantages	The limitaions
4- P.-W. Tsai et al. (2012) [64]	enhanced parallel cat swarm optimization (EPCSO)	The PCSO technique is an optimisation approach developed to address numerical optimisation problems with a small population size and a limited number of iterations. In this study, the authors suggest the integration of the Taguchi technique's orthogonal array into the tracing mode operation of the PCSO method, which they refer to as the EPCSO approach.	Problem dependent	1- The suggested EPCSO approach may find optimal solutions in a relatively short period. 2- The EPCSO approach outperforms previous PSO-based methods in terms of accuracy while requiring less processing time than the PCSO method.	1- EPCSO has no theoretical promises about convergence or optimality. It may not always identify the global optimum, and the solution's quality is determined on the unique issue and parameter choices. 2- which may hinder its capacity to efficiently explore the search space. It may become trapped in local optima and struggle to escape, particularly in complicated and multimodal optimisation situations.
5- Cohen et al (2018) [65]	CancerSEEK	Protein biomarkers and genetic indications were integrated by the researchers to increase sensitivity without sacrificing specificity. In a similar vein, other cancer markers including metabolites, messenger RNA transcripts, microRNAs, or methylation DNA sequences might be included to improve the precision and localization of cancer.	Dataset including clinical information on protein levels in circulating blood and genetic alterations found in cell-free DNA.	1- CancerSEEK identified cancer in a limited number of specific areas in 83% of people. 2The sensitivities for identifying five kinds of cancer (ovary, liver, stomach, pancreatic, and esophageal) ranged from 69% to 98%. 3- Multiple-fold cross-validation is a widely used and successful method for demonstrating the strong sensitivity as well as specificity of this study's cohort on a comparable scale.	1- The study was limited to individuals without any health conditions. However, in a real cancer screening scenario, some individuals might have inflamed or other disorders, leading to a greater likelihood of false-positive results compared to the findings shown in the research. 2- CancerSEEK's sensitivity is estimated to be 55% for all eight kinds of cancer. The use of this weighting would not impact the exceptional sensitivity of CancerSEEK (ranging from 69% to 98%) in detecting five forms of cancer (ovary, liver, stomach, pancreatic, and esophageal) that now lack screening tests for persons at average risk.

Author(s)/ Year	Techniques used	The summary	Dataset used	The advantages	The limitaions
6- Beal et al (2019) [66]	A unique framework, known as PROFILE (Approach called "logical modeling")	Integrating multi-omics data with the mechanical knowledge of logical modelling, this approach seeks to generate patient-relevant models. In the long run, doctors will be able to choose more effective pharmacological treatments for their patients because to logical modeling's application in precision medicine.	METABRIC dataset. PROFILE/Data/METABRIC at master · sysbio-curie/PROFILE · GitHub	1- infer that model simulations strongly correspond to clinical data, including subgrouping based on the Nottingham prognostic index (NPI) and patients' survival time. 2- Take note that both proliferation and apoptosis are physiologically consistent prognostic indicators patients with low proliferation and strong apoptosis have the best chance of survival, whereas patients with high proliferation and inadequate apoptosis have the worst chance of survival.	1- Prior to incorporation into the logical model, it is necessary to either convert these data into binary form or scale them between 0 and 1. This may be achieved by adjusting the activity of the node, the starting circumstances, or the state transition rates.
7- Best et al. (2017) [68]	PSO and RNA sequencing (RNA-seq) were performed on (TEPs).	Employ (PSO) methods and RNA sequencing (RNA-seq) of tumor-educated platelets obtained from patients to construct RNA sets that may effectively differentiate persons with non-small-cell lung cancer, namely those in the early stages, from healthy individuals, including those with inflammatory conditions.	non-small-cell lung cancer (NSCLC) patients.	1- TEP-based detection of early and late-stage non-small-cell lung cancer was achieved as a result of this discovery (n = 518 late-stage validation cohort, precision, 88.0%; AUC, 0.94; 95.0% confidence interval, 0.92 to 0.96; p 0.001; n = 106 early-stage validation cohort, precision, 81.0%; area under the curve, 0.89; 95.0% confidence interval, 0.83-0.95; p 0.001). 2- Swarm intelligence may be useful for optimising the diagnostic readout of various liquid biopsy biosources, as PSO allowed for the selection of gene panels that could identify cancer from TEPs.	1- Inflammatory disorders Tumor-educated platelet (TEP)-based cancer detection is only minimally confounded. 2- It is necessary to explore the dynamic reorganisation of TEP signatures during therapeutic courses and illness development.
8-Rappoport. (2019) [69]	NEMO algorithm	A one-of-a-kind technique for clustering across several omics. GitHub - Shamir-Lab/NEMO	TCGA datasets spanning 3168 patients. NEMO/data at master · Shamir-Lab/NEMO · GitHub	1- NEMO demonstrated superior performance compared to the top nine cutting-edge multi-omics clustering approaches on both entire and partial data. 2- NEMO exhibits notable improvements in speed and simplicity compared to prior multi-omics clustering methods, without the need for repeated optimization.	1 - At least one common omic is required for each pair of samples when dealing with partial data. When just one omic is considered for every patient, as is often the case with gene expression, this assumption holds. 2 Find the best value for k, the number of nearest neighbours; further study is required. Since all cluster sizes are assumed to be equal, NEMO presently uses a consistent k value for all samples.

ISSN: 2704-1077 eISSN 2704-1069, DOI: 10.59543/ijmscs.v2i.8703

5. THE FUTURE DIRECTION OF OMICS DATA.

Several exciting new directions are emerging in omics that promise to further our knowledge of biological systems using data analysis and technological advances. The first is the anticipated rise of customized medicine, which will use omics data to direct individualized treatment plans according to each person's molecular profile. Combining multi-omics data from genomics, transcriptomics, proteomics, and metabolomics makes it possible to gain a deeper understanding of biological processes and the capacity to make holistic conclusions. When finding valuable patterns from massive omics datasets, cutting-edge computational methods like AI and ML will be pivotal. Furthermore, researchers will be able to study cellular heterogeneity with unprecedented resolution with the advent of single-cell omics technology. This will improve our comprehension of complicated tissues and disease causes. The possibility for large-scale population research, which might aid precision public health efforts, is increasing as omics technologies become cheaper and more widely available. To shape an ethical and fair future for omics research, issues of data privacy, standards, and ethics will continue to play a pivotal role. The omics field is full of potential for revolutionary changes in areas such as population health, customized treatment, and scientific discovery in the years to come.

Quantum Machine Learning A fascinating and promising paradigm within quantum technology has emerged in quantum machine learning. On one side, it can potentially enhance the efficiency of machine learning calculations using quantum devices. On the other hand, it can improve the management of quantum systems using machine learning approaches. Quantum reinforcement learning is a field within quantum machine learning that aims to develop quantum agents capable of intelligently interacting with their environment and adapting their behavior to achieve certain objectives. In the future, omics data and quantum machine learning (QML) can potentially revolutionize cancer detection. Omics data includes several different forms of data, such as genomes, transcriptomics, proteomics, metabolomics, and epigenomics. This information can provide a detailed picture of the molecular changes in cancer cells.

QML is a new branch of machine learning that uses the capabilities of quantum computers to address complicated problems that traditional computers cannot handle. QML algorithms are capable of analyzing omics data to reveal cancer-related patterns and biomarkers. Using omics data and QML in cancer detection has a bright future. These technologies can potentially provide cancer patients with earlier detection, more accurate diagnosis, and personalized treatment. Creating novel ML and DL algorithms tailored specifically for cancer diagnosis using omics data. Integration of machine learning and deep learning with other developing technologies, such as quantum computing, to provide even more powerful and accurate diagnostic tools. Clinical trials are being developed to assess the safety and efficacy of ML and DL models for cancer detection and treatment.

6. CONCLUSION

The application of ML and DL for cancer diagnosis from omics data is a rapidly emerging and promising topic. ML and DL algorithms can increase cancer diagnostic accuracy, speed, and cost-effectiveness. As ML and DL algorithms improve, they are expected to play an increasing role in cancer care. Therefore, more research is required to give a comprehensive vision and follow this line of inquiry. This study offers a broad perspective and thorough comprehension by analyzing and categorizing the relevant literature. Clinical uses, reviews, and other research employing DL and omics data are this study's primary areas for organizing the remaining relevant studies. Diseases classification, biomarker discovery, pathway evaluation, data priority setting, research reviews accomplished to single- or multi-omics data, comparison analysis, and guideline studies are some of the subcategories within the three main classes of articles that present a comprehensive literature review. Data and validation process details and the challenges and limits of DL models are also included. According to our evaluation, researchers have been paying much attention to clinical applications that employ DL to solve many medical issues.

Furthermore, there needs to be more focus on DL traits, feature data, and definitions in the current research regarding a simple medical validation strategy. In addition, healthcare companies seldom adopt or employ testbed DL applications in real-world contexts. So, the omics community must significantly bridge the gap between theory and practice by developing relevant applications. Lastly, by extending and defining new research routes, this systematic review supports investigators in tracking the essential issues of DL in omics.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

REFERENCES

[1] M. Vailati-Riboni, V. Palombo, and J. J. Loor, "What are omics sciences?," in Periparturient Diseases of Dairy Cows: A Systems Biology Approach, Springer International Publishing, 2017, pp. 1–7. doi: 10.1007/978-3-319-43033-1_1.

[2] Z. Momeni, E. Hassanzadeh, M. Saniee Abadeh, and R. Bellazzi, "A survey on single and multi omics data mining methods in cancer data classification," Journal of Biomedical Informatics, vol. 107. Academic Press Inc., Jul. 01, 2020. doi: 10.1016/j.jbi.2020.103466.

[3] J. Paananen and V. Fortino, "An omics perspective on drug target discovery platforms," Brief Bioinform, vol. 21, no. 6, pp. 1937–1953, Nov. 2020, doi: 10.1093/bib/bbz122.

[4] M. A. Mohammed, A. Lakhan, K. H. Abdulkareem, and B. Garcia-Zapirain, "A hybrid cancer prediction based on multi-omics data and reinforcement learning state action reward state action (SARSA)," Comput Biol Med, vol. 154, Mar. 2023, doi: 10.1016/j.compbiomed.2023.106617.

[5] A. Desiani, A. A. Lestari, M. Al-Ariq, A. Amran, and Y. Andriani, "Comparison of Support Vector Machine and K-Nearest Neighbors in Breast Cancer Classification," Pattimura International Journal of Mathematics (PIJMath), vol. 1, no. 1, pp. 33–42, May 2022, doi: 10.30598/pijmathvol1iss1pp33-42.

[6] C. Aroef, Y. Rivan, and Z. Rustam, "Comparing random forest and support vector machines for breast cancer classification," Telkomnika (Telecommunication Computing Electronics and Control), vol. 18, no. 2, pp. 815–821, Apr. 2020, doi: 10.12928/TELKOMNIKA.V18I2.14785.

[7] L. Yuan, J. Zhao, T. Sun, and Z. Shen, "A machine learning framework that integrates multi-omics data predicts cancer-related LncRNAs," BMC Bioinformatics, vol. 22, no. 1, Dec. 2021, doi: 10.1186/s12859-021-04256-8.

[8] S. Ferro, D. Bottigliengo, D. Gregori, A. S. C. Fabricio, M. Gion, and I. Baldi, "Phenomapping of patients with primary breast cancer using machine learning-based unsupervised cluster analysis," J Pers Med, vol. 11, no. 4, 2021, doi: 10.3390/jpm11040272.

[9] J. Li et al., "Molecular breast cancer subtype identification using photoacoustic spectral analysis and machine learning at the biomacromolecular level," Photoacoustics, vol. 30, Apr. 2023, doi: 10.1016/j.pacs.2023.100483.

[10] A. Lopez-Rincon et al., "Machine learning-based ensemble recursive feature selection of circulating mirnas for cancer tumor classification," Cancers (Basel), vol. 12, no. 7, pp. 1–27, Jul. 2020, doi: 10.3390/cancers12071785.

[11] J. P. Sarkar, I. Saha, A. Sarkar, and U. Maulik, "Machine learning integrated ensemble of feature selection methods followed by survival analysis for predicting breast cancer subtype specific miRNA biomarkers," Comput Biol Med, vol. 131, Apr. 2021, doi: 10.1016/j.compbiomed.2021.104244.

[12] J. Liu et al., "A novel consensus learning approach to incomplete multi-view clustering," Pattern Recognit, vol. 115, Jul. 2021, doi: 10.1016/j.patcog.2021.107890.

[13] P. Andreini, S. Bonechi, M. Bianchini, and F. Geraci, "MicroRNA signature for interpretable breast cancer classification with subtype clue," Journal of Computational Mathematics and Data Science, vol. 3, p. 100042, Jun. 2022, doi: 10.1016/j.jcmds.2022.100042.

[14] M. A. Mohammed, A. Lakhan, K. H. Abdulkareem, and B. Garcia-Zapirain, "Federated auto-encoder and XGBoost schemes for multi-omics cancer detection in distributed fog computing paradigm," Chemometrics and Intelligent Laboratory Systems, vol. 241, Oct. 2023, doi: 10.1016/j.chemolab.2023.104932.

[15] B. Ma, F. Meng, G. Yan, H. Yan, B. Chai, and F. Song, "Diagnostic classification of cancers using extreme gradient boosting algorithm and multi-omics data," Comput Biol Med, vol. 121, Jun. 2020, doi: 10.1016/j.compbiomed.2020.103761.

[16] J. J. Chabon et al., "Integrating genomic features for non-invasive early lung cancer detection," Nature, vol. 580, no. 7802, pp. 245–251, Apr. 2020, doi: 10.1038/s41586-020-2140-0.

[17] A. Mohammed, G. Biegert, J. Adamec, and T. Helikar, "Identification of potential tissue-specific cancer biomarkers and development of cancer versus normal genomic classifiers," 2017. [Online]. Available: www.impactjournals.com/oncotarget/

[18] V. Crippa et al., "Characterization of cancer subtypes associated with clinical outcomes by multi-omics integrative clustering," Comput Biol Med, vol. 162, Aug. 2023, doi: 10.1016/j.compbiomed.2023.107064.

[19] M. Abdar and V. Makarenkov, "CWV-BANN-SVM ensemble learning classifier for an accurate diagnosis of breast cancer," Measurement (Lond), vol. 146, pp. 557–570, Nov. 2019, doi: 10.1016/j.measurement.2019.05.022.

[20] D. Wu, D. Wang, M. Q. Zhang, and J. Gu, "Fast dimension reduction and integrative clustering of multi-omics data using lowrank approximation: Application to cancer molecular classification," BMC Genomics, vol. 16, no. 1, Dec. 2015, doi: 10.1186/s12864-015-2223-8.

[21] L. D. Naorem, M. Muthaiyan, and A. Venkatesan, "Identification of dysregulated miRNAs in triple negative breast cancer: A meta-analysis approach," J Cell Physiol, vol. 234, no. 7, pp. 11768–11779, Jul. 2019, doi: 10.1002/jcp.27839.

[22] S. Liu et al., "Survival Time Prediction of Breast Cancer Patients Using Feature Selection Algorithm Crystall," IEEE Access, vol. 9, pp. 24433–24445, 2021, doi: 10.1109/ACCESS.2021.3054823.

[23] P. Mohapatra, S. Chakravarty, and P. K. Dash, "Microarray medical data classification using kernel ridge regression and modified cat swarm optimization based gene selection system," Swarm Evol Comput, vol. 28, pp. 144–160, Jun. 2016, doi: 10.1016/j.swevo.2016.02.002.

[24] S. C. Chu, P. W. Tsai, and J. S. Pan, "Cat swarm optimization," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Springer Verlag, 2006, pp. 854–858. doi: 10.1007/11801603_94.

[25] S. Meshoul, A. Batouche, H. Shaiba, and S. AlBinali, "Explainable Multi-Class Classification Based on Integrative Feature Selection for Breast Cancer Subtyping," Mathematics, vol. 10, no. 22, Nov. 2022, doi: 10.3390/math10224271.

[26] B. Wang et al., "Similarity network fusion for aggregating data types on a genomic scale," Nat Methods, vol. 11, no. 3, pp. 333–337, 2014, doi: 10.1038/nmeth.2810.

[27] Q. Liu, B. Cheng, Y. Jin, and P. Hu, "Bayesian tensor factorization-drive breast cancer subtyping by integrating multi-omics data," J Biomed Inform, vol. 125, Jan. 2022, doi: 10.1016/j.jbi.2021.103958.
[28] E. Urol, "Classification of prostate cancer based on clinical and omic data using neural networks techniques to improve prognostic power," 2019.
[29] Z. Wang, S. Zhao, Z. Li, H. Chen, C. Li, and Y. Shen, "Ensemble selection with joint spectral clustering and structural sparsity," Pattern Recognit, vol. 119, Nov. 2021, doi: 10.1016/j.patcog.2021.108061.
[30] B. Pfeifer, M. D. Bloice, and M. G. Schimek, "Parea: Multi-view ensemble clustering for cancer subtype discovery," J Biomed Inform, vol. 143, Jul. 2023, doi: 10.1016/j.jbi.2023.104406.
[31] I. Khan, Z. Luo, A. K. Shaikh, and R. Hedjam, "Ensemble clustering using extended fuzzy k-means for cancer data analysis," Expert Syst Appl, vol. 172, Jun. 2021, doi: 10.1016/j.eswa.2021.114622.
[32] A. Majumdar, Y. Liu, Y. Lu, S. Wu, and L. Cheng, "Kesvr: An ensemble model for drug response prediction in precision medicine using cancer cell lines gene expression," Genes (Basel), vol. 12, no. 6, Jun. 2021, doi: 10.3390/genes12060844.
[33] H. Huang et al., "A new fruit fly optimization algorithm enhanced support vector machine for diagnosis of breast cancer based on high-level features," BMC Bioinformatics, vol. 20, Jun. 2019, doi: 10.1186/s12859-019-2771-z.
[34] J. Chen and L. Zhang, "A survey and systematic assessment of computational methods for drug response prediction," Briefings in Bioinformatics, vol. 22, no. 1. Oxford University Press, pp. 232–246, Jan. 01, 2021. doi: 10.1093/bib/bbz164.
[35] M. A. Mohammed, K. H. Abdulkareem, A. M. Dinar, and B. G. Zapirain, "Rise of Deep Learning Clinical Applications and Challenges in Omics Data: A Systematic Review," Diagnostics, vol. 13, no. 4. MDPI, Feb. 01, 2023. doi: 10.3390/diagnostics13040664.
[36] H. Husi, Ed., Computational Biology. Codon Publications, 2019. doi: 10.15586/computationalbiology.2019.
[37] K. Clark et al., "The cancer imaging archive (TCIA): Maintaining and operating a public information repository," J Digit Imaging, vol. 26, no. 6, pp. 1045–1057, Dec. 2013, doi: 10.1007/s10278-013-9622-7.
[38] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-ray8: Hospital-scale Chest X-ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases," May 2017, doi: 10.1109/CVPR.2017.369.
[39] J. N. Weinstein et al., "The cancer genome atlas pan-cancer analysis project," Nat Genet, vol. 45, no. 10, pp. 1113–1120, Oct. 2013, doi: 10.1038/ng.2764.
[40] Y. Huang, W. Jin, Z. Yu, and B. Li, "Supervised feature selection through Deep Neural Networks with pairwise connected structure," Knowl Based Syst, vol. 204, Sep. 2020, doi: 10.1016/j.knosys.2020.106202.
[41] Y. Lin, W. Zhang, H. Cao, G. Li, and W. Du, "Classifying breast cancer subtypes using deep neural networks based on multi-omics data," Genes (Basel), vol. 11, no. 8, pp. 1–18, Aug. 2020, doi: 10.3390/genes11080888.
[42] S. Karthik, R. Srinivasa Perumal, and P. V. S. S. R. Chandra Mouli, "Breast cancer classification using deep neural networks," in Knowledge Computing and Its Applications: Knowledge Manipulation and Processing Techniques: Volume 1, Springer Singapore, 2018, pp. 227–241. doi: 10.1007/978-981-10-6680-1_12.
[43] A. Dhillon, A. Singh, and V. K. Bhalla, "Biomarker identification and cancer survival prediction using random spatial local best cat swarm and Bayesian optimized DNN," Appl Soft Comput, vol. 146, Oct. 2023, doi: 10.1016/j.asoc.2023.110649.
[44] T. Y. Lee, K. Y. Huang, C. H. Chuang, C. Y. Lee, and T. H. Chang, "Incorporating deep learning and multi-omics autoencoding for analysis of lung adenocarcinoma prognostication," Comput Biol Chem, vol. 87, Aug. 2020, doi: 10.1016/j.compbiolchem.2020.107277.
[45] H. Chai, X. Zhou, Z. Zhang, J. Rao, H. Zhao, and Y. Yang, "Integrating multi-omics data through deep learning for accurate cancer prognosis prediction," Comput Biol Med, vol. 134, Jul. 2021, doi: 10.1016/j.compbiomed.2021.104481.
[46] W. Jiao et al., "A deep learning system accurately classifies primary and metastatic cancers using passenger mutation patterns," Nat Commun, vol. 11, no. 1, Dec. 2020, doi: 10.1038/s41467-019-13825-8.
[47] S. Babichev, L. Yasinska-Damri, and I. Liakh, "A Hybrid Model of Cancer Diseases Diagnosis Based on Gene Expression Data with Joint Use of Data Mining Methods and Machine Learning Techniques," Applied Sciences (Switzerland), vol. 13, no. 10, May 2023, doi: 10.3390/app13106022.
[48] Y. Chen et al., "MOCSS: Multi-omics data clustering and cancer subtyping via shared and specific representation learning," iScience, vol. 26, no. 8, Aug. 2023, doi: 10.1016/j.isci.2023.107378.
[49] S. Salimy et al., "A deep learning-based framework for predicting survival-associated groups in colon cancer by integrating multi-omics and clinical data," Heliyon, vol. 9, no. 7, Jul. 2023, doi: 10.1016/j.heliyon.2023.e17653.
[50] S. W. Zhang, J. Y. Xu, and T. Zhang, "DGMP: Identifying Cancer Driver Genes by Jointing DGCN and MLP from Multi-omics Genomic Data," Genomics Proteomics Bioinformatics, vol. 20, no. 5, pp. 928–938, Oct. 2022, doi: 10.1016/j.gpb.2022.11.004.
[51] Y. Zhang and H. Kiryu, "MODEC: an unsupervised clustering method integrating omics data for identifying cancer subtypes," Brief Bioinform, vol. 23, no. 6, Nov. 2022, doi: 10.1093/bib/bbac372.
[52] C. M. Cres, A. Tritt, K. E. Bouchard, and Y. Zhang, "DL-TODA: A Deep Learning Tool for Omics Data Analysis," Biomolecules, vol. 13, no. 4, Apr. 2023, doi: 10.3390/biom13040585.
[53] N. A. Othman, M. A. Abdel-Fattah, and A. T. Ali, "A Hybrid Deep Learning Framework with Decision-Level Fusion for Breast Cancer Survival Prediction," Big Data and Cognitive Computing, vol. 7, no. 1, Mar. 2023, doi: 10.3390/bdcc7010050.
[54] D. Khan and S. Shedole, "Leveraging Deep Learning Techniques and Integrated Omics Data for Tailored Treatment of Breast Cancer," J Pers Med, vol. 12, no. 5, May 2022, doi: 10.3390/jpm12050674.
[55] A. El-Nabawy, N. A. Belal, and N. El-Bendary, "A cascade deep forest model for breast cancer subtype classification using multi-omics data," Mathematics, vol. 9, no. 13, Jul. 2021, doi: 10.3390/math9131574.

[56] X. Zhang, Y. Xing, K. Sun, and Y. Guo, "Omiembed: A unified multi-task deep learning framework for multi-omics data," Cancers (Basel), vol. 13, no. 12, Jun. 2021, doi: 10.3390/cancers13123047.

[57] E. F. Franco et al., "Performance comparison of deep learning autoencoders for cancer subtype detection using multi-omics data," Cancers (Basel), vol. 13, no. 9, May 2021, doi: 10.3390/cancers13092013.

[58] G. Xie, C. Dong, Y. Kong, J. F. Zhong, M. Li, and K. Wang, "Group lasso regularized deep learning for cancer prognosis from multi-omics and clinical features," Genes (Basel), vol. 10, no. 3, Mar. 2019, doi: 10.3390/genes10030240.

[59] L. Alessandri et al., "Sparsely connected autoencoders: A multi-purpose tool for single cell omics analysis," Int J Mol Sci, vol. 22, no. 23, Dec. 2021, doi: 10.3390/ijms222312755.

[60] O. B. Poirion, Z. Jing, K. Chaudhary, S. Huang, and L. X. Garmire, "DeepProg: an ensemble of deep-learning and machine-learning models for prognosis prediction using multi-omics data," Genome Med, vol. 13, no. 1, Dec. 2021, doi: 10.1186/s13073-021-00930-x.

[61] K. Vandereyken, A. Sifrim, B. Thienpont, and T. Voet, "Methods and applications for single-cell and spatial multi-omics," Nature Reviews Genetics, vol. 24, no. 8. Nature Research, pp. 494–515, Aug. 01, 2023. doi: 10.1038/s41576-023-00580-2.

[62] M. Büttner, Z. Miao, F. A. Wolf, S. A. Teichmann, and F. J. Theis, "A test metric for assessing single-cell RNA-seq batch correction," Nat Methods, vol. 16, no. 1, pp. 43–49, Jan. 2019, doi: 10.1038/s41592-018-0254-1.

[63] L. Rejali et al., "Principles of Molecular Utility for CMS Classification in Colorectal Cancer Management," Cancers, vol. 15, no. 10. MDPI, May 01, 2023. doi: 10.3390/cancers15102746.

[64] P. W. Tsai, J. S. Pan, S. M. Chen, and B. Y. Liao, "Enhanced parallel cat swarm optimization based on the Taguchi method," Expert Syst Appl, vol. 39, no. 7, pp. 6309–6319, Jun. 2012, doi: 10.1016/j.eswa.2011.11.117.

[65] J. D. Cohen et al., "Detection and localization of surgically resectable cancers with a multi-analyte blood test." [Online]. Available: https://www.science.org

[66] J. Beal, A. Montagud, P. Traynard, E. Barillot, and L. Calzone, "Personalization of logical models with multi-omics data allows clinical stratification of patients," Front Physiol, vol. 10, no. JAN, 2019, doi: 10.3389/fphys.2018.01965.

[67] Y. Xu et al., "Identifying individualized risk subpathways reveals pan-cancer molecular classification based on multi-omics data," Comput Struct Biotechnol J, vol. 20, pp. 838–849, Jan. 2022, doi: 10.1016/j.csbj.2022.01.022.

[68] M. G. Best et al., "Swarm Intelligence-Enhanced Detection of Non-Small-Cell Lung Cancer Using Tumor-Educated Platelets," Cancer Cell, vol. 32, no. 2, pp. 238-252.e9, Aug. 2017, doi: 10.1016/j.ccell.2017.07.004.

[69] N. Rappoport and R. Shamir, "NEMO: Cancer subtyping by integration of partial multi-omic data," Bioinformatics, vol. 35, no. 18, pp. 3348–3356, Sep. 2019, doi: 10.1093/bioinformatics/btz058.