ACADEMIA BIOLOGY

Research Article
Published: 2024-08-30
https://doi.org/10.20935/AcadBiol7325

LASSO–MOGAT: a multi-omics graph attention framework for cancer classification

Fadi Alharbi[1], Aleksandar Vakanski[1,*], Murtada K. Elbashir[2], Mohanad Mohammed[3]

Academic Editor(s): Ravindra Kumar, Andre J. van Wijnen

Abstract

The application of machine learning (ML) methods to analyze changes in gene expression patterns has recently emerged as a powerful approach in cancer research, enhancing our understanding of the molecular mechanisms underpinning cancer development and progression. Combining gene expression data with other types of omics data has been reported by numerous works to improve cancer classification outcomes. Despite these advances, effectively integrating high-dimensional multi-omics data and capturing the complex relationships across different biological layers remain challenging. This article introduces Least Absolute Shrinkage and Selection Operator–Multi-omics Gated Attention (LASSO–MOGAT), a novel graph-based deep learning framework that integrates messenger RNA, microRNA, and DNA methylation data to classify 31 cancer types. By utilizing differential expression analysis (DEG) with Linear Models for Microarray (LIMMA) and LASSO regression for feature selection and leveraging graph attention networks (GATs) to incorporate protein–protein interaction (PPI) networks, LASSO–MOGAT effectively captures intricate relationships within multi-omics data. Experimental validation using fivefold cross-validation demonstrates the method's precision, reliability, and capacity to provide comprehensive insights into cancer molecular mechanisms. The computation of attention coefficients for the edges in the graph, facilitated by the proposed graph attention architecture based on PPIs, proved beneficial for identifying synergies in multi-omics data for cancer classification.

Keywords: *gene expression analysis*, *machine learning*, *multi-omics data*, *graph attention networks*

Citation: Alharbi F, Vakanski A, Elbashir MK, Mohammed M. LASSO–MOGAT: a multi-omics graph attention framework for cancer classification. *Academia Biology* 2024;2. https://doi.org/10.20935/AcadBiol7325

1. Introduction

Recent advances in machine learning (ML) and deep learning (DL) approaches have advanced our capacity to categorize various forms of cancer [1]. Graph-based DL architectures, in particular, offer the potential to leverage complex biological networks, such as protein–protein interaction (PPI) networks and gene regulatory networks, to extract meaningful data representations. By treating biological networks as graphs, where nodes represent biological entities (e.g., genes and proteins) and edges represent interactions between them, graph-based architectures effectively capture the topological and functional characteristics of these networks [2]. Graph-based architectures, including graph neural networks (GNNs), graph convolutional neural networks (GCNNs), graph attention networks (GATs), and graph transformer networks (GTNs), have been successfully utilized for integrating multi-omics data and biological networks [3, 4] and have shown promise in learning complex patterns from biological networks and omics data [5]. For instance, GATs employ attention mechanisms to focus on important nodes and edges in a graph, providing insights into the molecular mechanisms underlying cancer development and progression [6]. The utilization of graph-based networks in multi-omics data analysis is likely to continue to grow, with further advancements in model architectures and applications [7].

Gene expression data is a valuable resource in cancer research, offering insights into the activity levels of genes within specific tissues or cell types and enabling comparisons between cancerous and normal cells [8]. By measuring the amount of messenger RNA (mRNA) produced by each gene, researchers can determine which genes are being actively transcribed and expressed in these specific contexts [9]. This information is crucial for understanding the molecular mechanisms underlying cancer development and progression, where certain genes may be upregulated (increased expression) or downregulated (decreased expression) compared to normal cells [10]. Changes in gene expression can be indicative of the dysregulation of cellular processes, such as cell growth, division, and apoptosis, which are distinguishing

[1]Department of Computer Science, College of Engineering, University of Idaho, Moscow, ID 83844, USA.
[2]Department of Information Systems, College of Computer and Information Sciences, Jouf University, Sakaka, Aljouf 72441, Saudi Arabia.
[3]School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Pietermaritzburg, Scottsville 3209, South Africa.
*email: vakanski@uidaho.edu; vakanski@uidaho.com

features of cancer. By analyzing gene expression patterns, researchers can identify genes that are specifically altered in different cancer types, providing valuable biomarkers for early detection and diagnosis.

To gain a more comprehensive understanding of cancer biology, researchers have integrated gene expression data with other types of omics data, such as DNA methylation and microRNA (miRNA) expression data [11]. This multi-omics approach allows for the exploration of complex relationships between different molecular layers and the identification of key molecular alterations driving cancer development and progression [12]. For example, DNA methylation data can reveal epigenetic changes that silence tumor suppressor genes or activate oncogenes, contributing to cancer initiation and advancement [13]. MiRNA expression data, on the other hand, can offer insights into post-transcriptional gene regulation and its role in cancer pathogenesis [14]. Overall, the integration of gene expression with multi-omics data offers a more inclusive view of cancer biology and holds great promise for advancing precision oncology.

This article introduces Least Absolute Shrinkage and Selection Operator–Multi-omics Gated Attention (LASSO–MOGAT), a graph-based DL approach that integrates multi-omics data—comprising mRNA, miRNA, and DNA methylation data—to classify 31 cancer types. To select the most informative multi-omics features, we implemented data preprocessing using differential expression analysis (DEG) with Linear Models for Microarray (LIMMA) [15]. This approach involves fitting a linear model to each gene to estimate the difference in expression between conditions while accounting for variability. LIMMA uses an empirical Bayes approach to moderate the standard errors of the estimated log-fold changes, improving the accuracy of differential expression detection, particularly for genes with low expression levels. This method provides robust statistical inference for identifying genes that are differentially expressed across conditions, with high sensitivity and specificity. We applied LASSO regression to further refine the feature selection process by penalizing the absolute size of the regression coefficients, encouraging sparsity, and selecting the most relevant features for the classification task. We used the extracted lower-dimensional data representation as inputs to a GAT, leveraging the topological information from PPI networks for classification. Experimental validation using fivefold cross-validation demonstrates the efficacy of the proposed integration of multi-omics data with the graph-based GAT model for precise and reliable cancer classification.

Unlike other studies on cancer classification that focus on single types of omics data or simpler feature selection methods [16, 17], the proposed LASSO–MOGAT approach employs DEG with LIMMA and LASSO regression for robust feature selection. The employed GAT model effectively captures complex relationships within multi-omics data by incorporating PPI networks. Additionally, our approach differs from other studies based on GATs, such as Multi-omics Data Integration Graph (MODIG) [18], which aims on identifying cancer driver genes by integrating multi-omics pan-cancer data, including mutations and copy number variants, with multidimensional gene networks. MODIG leverages GATs to generate gene representations across different dimensions, including PPI, gene sequence similarity, and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway co-occurrence, to enhance the identification of driver genes. While both approaches utilize GATs for processing multi-omics data, their objectives differ: our work focuses on cancer type classification, whereas MODIG [18] is designed for cancer driver gene identification.

Our approach also stands apart from methods, such as Graph-Based Optimization for Analysis of Tumors (GOAT) [19], Multi-omics Regression with Graph Attention Networks (MORGAT) [20], Deep Graph Partitioning with Attention Mechanism for Integrating Omics Data (DGP–AMIO) [21], and MOGAT [7], which either target different diseases (e.g., asthma) or employ distinct graph-based learning strategies. Furthermore, LASSO–MOGAT is distinct from other works that utilize GNNs for processing multi-omics data but focus on different aspects and methodologies. For instance, Li and Nabavi [6] introduced a framework that combines both inter-omics and intra-omics connections using a heterogeneous multilayer graph, emphasizing the integration of different types of GNNs (GCN and GAT) with a focus on the classification of cancer subtypes.

The key contributions of this work to the field of cancer classification are as follows:

- We propose a method to integrate RNA-Seq, miRNA, and DNA methylation data for 31 cancer types in addition to normal samples, providing a comprehensive multi-omics perspective on cancer.
- We develop a GAT model that leverages PPI networks to classify cancer types, including normal samples, and effectively captures the complex relationships within the multi-omics data.
- We empirically demonstrate that the use of DEGs with LIMMA and LASSO regression is effective in selecting the most relevant features from multi-omics data, significantly improving the performance of classification models.

2. Related works

In recent years, the integration of multi-omics data has emerged as a promising approach in cancer research, offering deeper insights into the complex biological mechanisms underlying the disease. Several studies have proposed novel algorithms and models to effectively integrate and analyze multi-omics data for cancer subtype classification, leveraging deep learning (DL) techniques and attention mechanisms to enhance both model interpretability and performance.

Mostavi et al. [22] introduced several convolutional neural network (CNN) models for cancer-type prediction based on gene expression data. Their study presented three different CNN architectures—1D-CNN, 2D-Vanilla-CNN, and 2D-Hybrid-CNN—which demonstrated the importance of addressing tissue origin effects in cancer marker identification. This work highlighted that CNN models could effectively classify cancer types and identify cancer marker genes through model interpretation techniques, making significant strides in cancer-type prediction using gene expression profiles.

Ramirez et al. [23] applied graph convolutional networks (GCNs) for classifying cancer types using gene expression data. They utilized different graph structures, including co-expression and PPI graphs, to capture molecular interactions and identify cancer-specific marker genes. Their work demonstrated the power of GCNs in leveraging multi-omics data for cancer classification and marker identification, achieving high prediction accuracies and providing insights into the effects of gene perturbations.

Kaczmarek et al. [24] developed a GTN for cancer classification, leveraging miRNA and mRNA expression data to model biological interactions and target pathways. The GTN provided high interpretability through self-attention mechanisms, identifying important pathways and biomarkers. Although their GTN did not outperform all baseline models in terms of accuracy, its interpretability offers valuable insights into miRNA–mRNA interactions, making it a crucial tool for understanding complex biological relationships in cancer.

Other significant contributions include Moon and Lee [25], who proposed the Multi-omics Module Analysis (MOMA) model for cancer subtype classification using a geometrical approach and attention mechanisms to integrate multi-biomedical data. Zhang et al. [26] presented a DL method that integrates multi-omics data using similarity network fusion and a graph autoencoder, enhancing performance on the Cancer Genome Atlas (TCGA) datasets. Sun et al. [27] introduced Self-Attention-based Deep Learning Network (SADLN) for cancer subtype recognition, integrating multi-omics data and modeling sample relationships based on latent low-dimensional representations.

Shanthamallu et al. [28] presented Graph Attention Models for Multilayered Embeddings (GrAMME) for semi-supervised learning with multilayered graphs. They developed two architectures to leverage inter-layer dependencies. Qiu et al. [16] introduced Gated Graph Attention Network (GGAT) for cancer prediction, combining gating and attention mechanisms to improve correlation capture and enhance prediction accuracy. Zhao et al. [18] proposed MODIG, which integrates multi-omics data with multidimensional gene networks using GATs for effective cancer research. Baul et al. [17] developed omicsGAT for RNA-Seq data in cancer subtyping, employing graph-based learning and attention mechanisms to enhance subtype prediction. Ouyang et al. [29] proposed Multi-omics Graph Learning and Attention Mechanism (MOGLAM), which integrates multi-omics data for disease classification and biomarker identification using dynamic graph convolution and attention mechanisms. Gong et al. [30] introduced Multi-omics Attention-Based Deep Learning Network (MOADLN), which employs self-attention to capture patient correlations within omics types and cross-omics correlations. Song et al. [31] presented GraphSAGE, which integrates multi-omics data with GNNs to predict cancer driver genes, demonstrating effectiveness across multiple tumor types.

Autoencoders have been widely used for dimensionality reduction and feature extraction in multi-omics data analysis. Zhang et al. [32] utilized a variational autoencoder for classifying different cancer types using RNA-Seq gene expression and DNA methylation data. Chai et al. [33] designed the Denoising Autoencoder for Accurate Cancer Prognosis Prediction (DCAP) to integrate multi-omics data for cancer prognosis prediction. Li et al. [34] proposed the Multi-omics Graph Convolutional Network (MoGCN), which uses an autoencoder for feature extraction and similarity network fusion to construct patient similarity networks. Zhou et al. [35] and Khadirnaikar et al. [36] used autoencoders to analyze multi-omics data, reducing dimensionality and identifying novel cancer subtypes through clustering and ML models.

GNNs have emerged as powerful models for analyzing multi-omics data, effectively capturing the relationships between different biological entities. Zhu et al. [37] proposed the Geometric Graph Neural Network (GGNN), incorporating geometric features for improved predictive power and interpretability. Xiao et al. [38] introduced the Multi-Prior Knowledge Graph Neural Network (MPKGNN) for multi-omics data analysis, leveraging multiple prior knowledge graphs for cancer molecular subtype classification. Chatzianastasis et al. [39] developed the Explainable Multilayer Graph Neural Network (EMGNN) to identify cancer genes by leveraging multiple gene–gene interaction networks and multi-omics pan-cancer data.

In addition to GNNs, various models integrating transformer and GCNs have been developed to enhance disease classification accuracy using multi-omics data. Wang et al. [40] introduced Multi-omics Self-Encoder Graph Convolutional Network (MOSEGCN), which combines transformer and similarity network fusion (SNF) for precise disease subtype classification. Yao et al. [41] developed Graph Convolutional Network Transformer (GCNFORMER), a model that integrates GCN and transformer to predict lncRNA-disease associations, demonstrating the effectiveness of combining these approaches for improved performance and interpretability.

CNNs are primarily utilized for grid-like data, such as images, where the spatial relationships between pixels are crucial for feature extraction. The interpretability of CNNs often involves visualizing filters and understanding which regions of an image activate specific neurons. In contrast, GATs are designed for graph-structured data, where nodes represent entities and edges represent relationships, which can vary in connectivity and significance. GATs enhance interpretability through their attention mechanisms, which dynamically weigh the importance of neighboring nodes during message passing. This attention mechanism provides insights into which nodes influence predictions and how relationships are weighted within the graph. Consequently, GATs are particularly adept at interpreting relational data and complex dependencies within graphs. While CNNs emphasize spatial hierarchies in images, GATs focus on relational structures in graphs, thereby offering a different perspective on interpretability.

In summary, the integration of multi-omics data in cancer research has been greatly enhanced by advanced techniques, such as autoencoders, GNNs, GCNs, and GATs. These methods have shown remarkable performance in cancer subtype classification, feature extraction, and understanding complex biological mechanisms underlying the disease. They offer improved interpretability, predictive power, and performance compared to traditional methods, making them valuable tools for researchers. Despite these advancements, several challenges remain. Multi-omics data are generated from various platforms, each with unique characteristics and measurement techniques, making their integration complex. The vast number of features compared to the number of samples in these datasets (large p small n issues) can lead to overfitting and reduced model performance, necessitating effective dimensionality reduction techniques. Additionally, multi-omics data often contain noise and sparse measurements, which can obscure true biological signals and require robust preprocessing methods. Batch effects or variations between different batches of data must be corrected to avoid misleading results. While advanced models like GNNs and GATs offer improved performance, their complexity can make interpretation difficult, posing a challenge in understanding the underlying biological mechanisms. Finally, the computational complexity of integrating and analyzing multi-omics data requires substantial resources and efficient algorithms to handle large-scale datasets. Therefore, there is a pressing need to develop new methods that can effectively address these challenges, improve data integration, and enhance our understanding of the molecular mechanisms driving cancer progression.

3. Materials and methods

3.1. Data collection

We retrieved the omics data for the various cancer types from the Pan-Cancer Atlas [42] using the Genomic Data Commons (GDC) query function from the TCGAbiolinks library [43]. GDC, a project by the National Cancer Institute (NCI), provides unified data storage for cancer genomic studies and facilitates data sharing within the research community.

The GDCquery function requires specifying parameters related to “project”, “legacy”, “data.category”, “data.type”, and “sample.type”. The “project” parameter refers to the name of the cancer research project within TCGA from which to retrieve data. TCGA consists of multiple cancer projects, each focused on a specific cancer type or cancer-related research area. In our case, we set the “project” parameter to "TCGA-*" to specify all the 33 TCGA projects, including those with normal tissues. The “legacy” argument was configured as “True”, signifying that the query was directed to the legacy repository to retrieve the original, unaltered data stored in the TCGA Data Portal. The “data.category” parameter specifies the relevant data category for the project. In our case, we set the “data category” to “Transcriptome Profiling” for the mRNA or RNA-Seq and miRNA datasets, while for methylation data, we specified the “data category” as “DNA Methylation”. The “data.type” parameter defines the specific type of data used for filtering the files to be downloaded. We specified the “data.type” parameter as “Gene Expression Quantification”, “miRNA Expression Quantification”, and “Methylation Beta Value” for mRNA or RNA-Seq, miRNA, and DNA methylation data, respectively. The “sample.type” parameter defines the type of sample used for filtering the data to be downloaded. In our case, we set the sample type to “Primary Solid Tumor” and “Solid Tissue Normal”, indicating that we aimed to download gene expression data specifically from both normal and tumor cases. The data was saved in a format where columns represent the samples or cases, and rows represent the features of interest for each data type. A summary of the collected data is presented in **Table 1**.

3.2. Data preprocessing

3.2.1. Differential gene expression analysis

In genomics, differential gene expression analysis is a technique used to identify genes that exhibit different expression levels under two or more biological conditions, such as treatment versus control samples or tumor versus healthy tissue [44]. The goal is to understand how gene expression patterns change in response to various circumstances, providing insights into the underlying biological processes. We conducted differential gene expression analysis for mRNA or RNA-Seq data using the DESeq2 package in R. This package fits a generalized linear model to the count data for each gene, assuming a negative binomial distribution to account for biological variability and overdispersion. The Wald test is used to evaluate the significance of the estimated log-fold changes, with p values derived from the Wald statistic are calculated to determine differential expression. By specifying a p value threshold of 0.001, we focused on identifying statistically significant genes that are likely to play a role in the biological processes under study.

3.2.2. The LIMMA model

We employed LIMMA to fit a linear model to the DNA methylation data, where the methylation levels of CpG sites are modeled as a function of the experimental sample groups [45]. The DNA methylation is from the Human Methylation 450K (HM450) platform [46] and includes 485,577 features and 9,171 samples. We applied LIMMA on the DNA methylation data to identify differentially methylated CpG sites between tumor and normal samples. The LIMMA method estimates the effect size (difference in methylation levels between groups) and calculates a moderated t-statistic for each CpG site. The P value associated with the t-statistics indicates the statistical significance of the differences in methylation levels between groups. In our case, we set the p value for LIMMA to 0.05, which reduced the number of features in the DNA methylation data to 139,321 features.

3.2.3. The LASSO regression model

LASSO regression is a type of linear regression that uses L1 regularization [47], defined as:

$$\text{minimize}\left(\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}\left|\beta_j\right|\right) \tag{1}$$

where n is the number of samples, p is the number of features, y_i is the target variable for the i^{th} sample, x_{ij} is the j^{th} feature of the i^{th} sample, β_0 is the intercept, β_j is the coefficient for the j^{th} feature, and λ is the regularization parameter that controls the strength of the penalty term. The term $\lambda\sum_{j=1}^{p}\left|\beta_j\right|$ is the L1 penalty term that encourages sparsity in the coefficient vector, effectively selecting only a subset of the most important features while setting the coefficients of less important features to zero.

We used LASSO regression for feature selection and regularization for both mRNA or RNA-Seq and DNA methylation data. The mRNA or RNA-Seq data contain a large number of genes as features after applying differential gene expression, and LASSO regression identified the most relevant genes for cancer-type prediction, reducing the features obtained using differential gene expression from 26,768 to 520. For DNA methylation data, the number of features (methylation sites) was reduced from 139,321 to 393. The pipeline for data processing is summarized in **Table 2**.

3.3. Multi-omics data integration

We integrated mRNA or RNA-Seq, miRNA, and DNA methylation data based on the sample ID. The goal was to combine the datasets so that the omics data pertaining to each sample are merged into a single record. To achieve the integration, we utilized an inner join operation on the sample ID column across the three datasets. An inner join retains only the samples that have data in all three datasets, ensuring that the integrated dataset contains only samples with complete omics data and avoids missing data issues. By merging the datasets using an inner join, we created a unified dataset that incorporates mRNA or RNA-Seq, miRNA, and DNA methylation data for each sample. During the integration process, we encountered instances where certain cancer types did not have data available for specific omics layers. For example, the RNA-Seq data was null for the cancer type TCGA_LAML, and the miRNA data was null for the cancer type TCGA_GBM. These cancer types were excluded from the integrated dataset to ensure that only samples with complete omics data were retained. Despite these exclusions, the integrated dataset still encompasses a diverse range of cancer types, totaling 31 cancer types along with normal samples. In total, the integrated dataset comprises 8,464 samples, each with 2,794 omics features. The preprocessing steps and data integration are depicted in **Figure 1**.

Table 1 • Tumor types and number of samples of TCGA multi-omics data (mRNA, miRNA, and DNA methylation) used in the analysis

Available cancer types		Number of samples		
		mRNA	**miRNA**	**DNA methylation**
Uterine corpus endometrial carcinoma	UCEC	588	578	484
Adrenocortical carcinoma	ACC	79	80	80
Brain lower-grade glioma	LGG	516	512	516
Sarcoma	SARC	261	259	265
Pancreatic adenocarcinoma	PAAD	182	182	194
Esophageal carcinoma	ESCA	197	199	201
Prostate adenocarcinoma	PRAD	553	550	552
Acute myeloid leukemia	LAML	–	–	–
Kidney renal clear cell carcinoma	KIRC	613	615	484
Pheochromocytoma and paraganglioma	PCPG	182	182	182
Head and neck squamous cell carcinoma	HNSC	564	567	578
Ovarian serous cystadenocarcinoma	OV	421	490	10
Glioblastoma multiforme	GBM	163	5	142
Uterine carcinosarcoma	UCS	57	57	57
Mesothelioma	MESO	87	87	87
Testicular germ cell tumors	TGCT	150	150	150
Kidney chromophobe	KICH	91	91	66
Rectum adenocarcinoma	READ	176	164	105
Uveal melanoma	UVM	80	80	80
Thyroid carcinoma	THCA	564	565	563
Liver hepatocellular carcinoma	LIHC	421	422	427
Thymoma	THYM	122	126	126
Cholangiocarcinoma	CHOL	44	45	45
Lymphoid neoplasm diffuse large B-cell lymphoma	DLBC	48	47	48
Kidney renal papillary cell carcinoma	KIRP	322	325	320
Bladder urothelial carcinoma	BLCA	431	436	439
Skin cutaneous melanoma	SKCM	104	99	106
Lung squamous cell carcinoma	LUSC	553	523	412
Stomach adenocarcinoma	STAD	448	491	397
Lung adenocarcinoma	LUAD	598	565	505
Colon adenocarcinoma	COAD	522	463	350
Cervical squamous cell carcinoma and endocervical adenocarcinoma	CESC	307	310	310
Breast invasive carcinoma	BRCA	1,224	1,200	890
Total		10,668	10,465	9,171

Table 2 • Pipeline for data processing

Data type	mRNA	miRNA	DNA methylation
Original features	60,660	1,881	485,577
Differentially expressed analysis	26,768	–	–
LIMMA model (selected features)	–	–	139,321
LASSO regression model (selected features)	520	–	393
All tumor samples and normal	10,668	10,465	9,171
Unique tumor samples and normal	10,667	10,465	8,674
Common samples and features	8,464 samples and 2,794 features		
Network nodes and edges	504 nodes and 343 edges		

Row one (original features): number of original features; row two (differentially expressed genes): number of features after applying differentially expressed analysis; row three (LIMMA model): number of features after applying LIMMA model; row four (LASSO regression model): number of features after applying LASSO regression model; row five (all tumor samples and normal): number of tumor samples and normal samples; row six (unique tumor samples and normal): number of tumor and normal samples after removing duplicates; row seven (common samples and features): number of tumor and normal samples and features common across all data types; and row eight (network nodes and edges): number of nodes and edges for the PPI network for each data type.

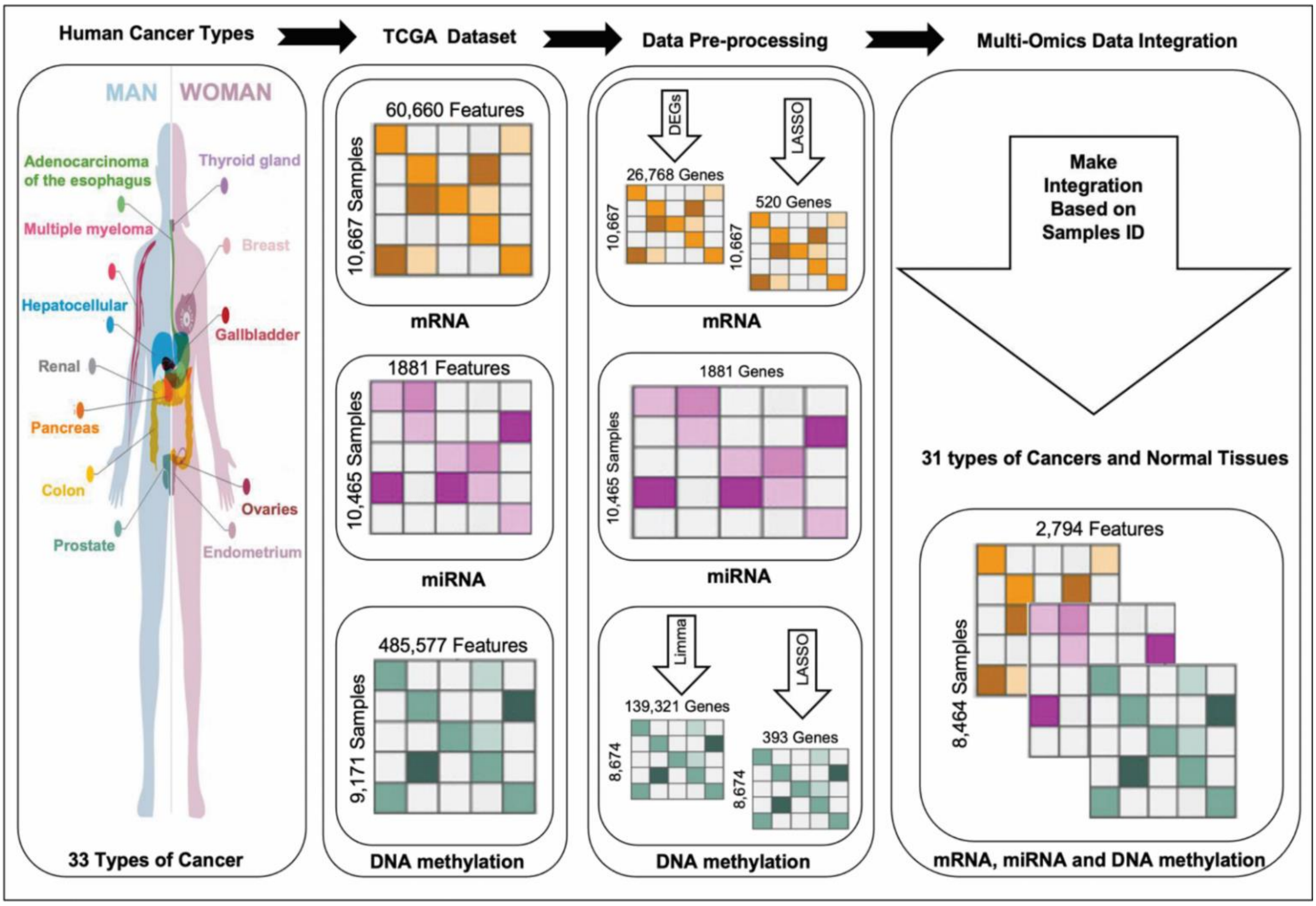

Figure 1 • Preprocessing steps and data integration. First, omics data (mRNA, miRNA, and DNA methylation) were obtained from the Pan-Cancer Atlas using the TCGAbiolinks library. Next, DEG and LASSO regression were applied to mRNA data, while LIMMA and LASSO regression were applied to DNA methylation data. Subsequently, mRNA or RNA-Seq, miRNA, and DNA methylation data were integrated based on the sample ID using an inner join operation.

3.4. The graph attention network

GATs employ attention mechanisms to enhance the functionality of standard GNNs [48]. GATs are particularly well-suited for handling data with complex relationships, such as multi-omics data, where features can be represented as nodes in a graph [49]. GATs function by computing attention coefficients for every edge in the graph, enabling the network to determine which nodes are more important to the representation of each node. By using the attention mechanism, GATs focus on pertinent nodes while simultaneously gathering information from nearby nodes. The propagation in a GAT layer is described by:

$$h_i^{(l+1)} = \sigma\left(\sum_{j\in\mathcal{N}(i)} \alpha_{ij}^{(l)} W^{(l)} h_j^{(l)}\right) \qquad (2)$$

where $h_i^{(l)}$ is the representation of node i in layer l, $N(i)$ is the neighborhood of node i, $\alpha_{ij}^{(l)}$ is the attention coefficient for edge $i \rightarrow j$ in layer l, $W^{(l)}$ is the weight matrix for layer l, and σ is the activation function.

The attention coefficients $\alpha_{ij}^{(l)}$ are computed using a softmax function over all neighbors of node i as follows:

$$\alpha_{ij}^{(l)} = \text{softmax}_j\left(\text{LeakyReLU}\left(\vec{a}^{(l)^T}\left[W^{(l)}h_i^{(l)} \,||\, W^{(l)}h_j^{(l)}\right]\right)\right) \quad (3)$$

where || denotes concatenation, $\vec{a}^{(l)}$ is a weight vector to be learned, and Leaky ReLU is the activation function.

In the context of multi-omics data, GATs are particularly important due to their capacity to efficiently extract and combine data from several omics layers represented graphically. More accurate and transparent models in multi-omics data analysis are made possible by the ability to learn which characteristics or nodes are more pertinent for each sample's representation through the use of attention mechanisms.

We designed a GAT model to operate on PPI networks represented as graphs. The model takes as input the PPI network structure and node features, where each node represents a protein and the edges represent interactions between proteins. The node features include multi-omics data, such as gene expression, miRNA expression, and DNA methylation levels. The GAT architecture, depicted in **Figure 2**, consists of four GAT convolutional layers, each followed by batch normalization and Leaky ReLU activation functions (with a default negative slope parameter of 0.01). These steps are applied after each GAT convolutional layer to normalize and activate the outputs. A dropout layer with a rate of 0.5 is added after each GAT convolutional layer to reduce overfitting by randomly dropping a fraction of the units during training. In the forward pass, the input data is passed through each GAT convolutional layer, followed by batch normalization, activation, and dropout. The final output is passed through a softmax function to obtain class probabilities. Our model was trained and evaluated using fivefold cross-validation to ensure robust performance metrics and avoid overfitting. For each fold, the model was initialized and trained from scratch, with parameters optimized using the training set and performance evaluated on the corresponding test set. The dataset was divided into five subsets, with each subset used once as a test set while the remaining four subsets were used for training. This process was repeated five times, and the average performance metrics across all folds were reported. The loss function used in our method, as described in our model implementation, is the negative log-likelihood loss. This loss function is suitable for multiclass classification tasks and is applied to the output of the final fully connected layer of our model, which utilizes the graph attention network (GAT) layers for feature aggregation and classification.

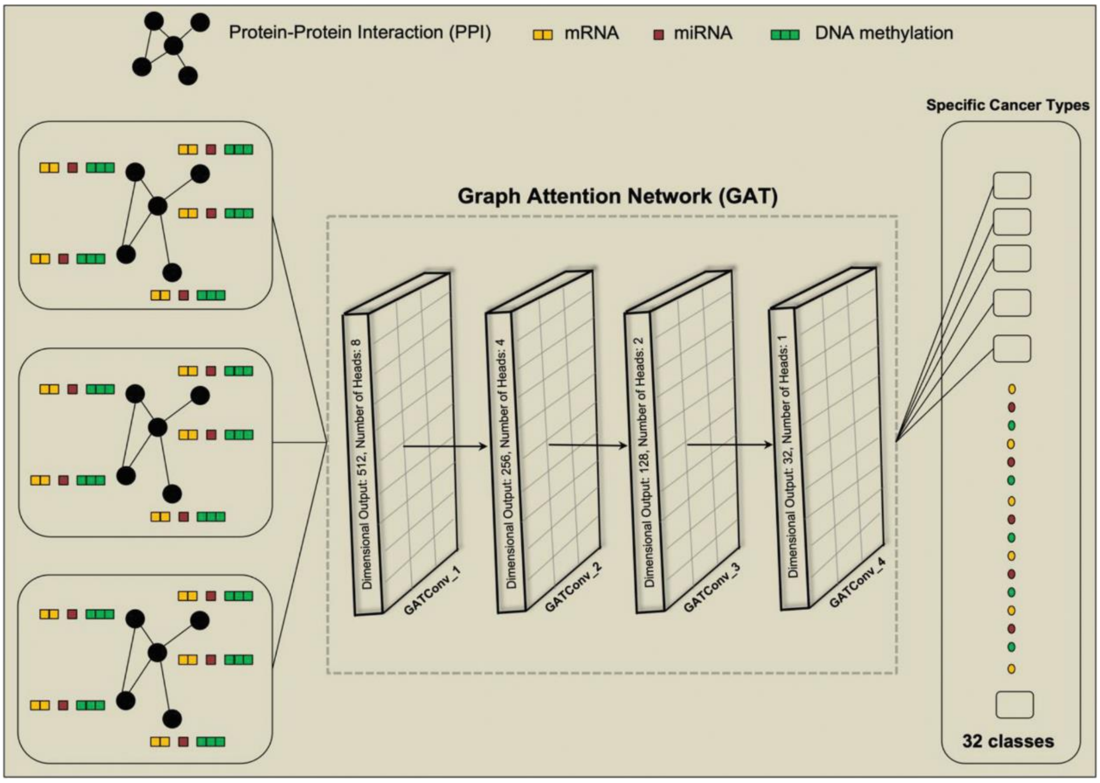
Protein-Protein Interaction (PPI)
mRNA
miRNA
DNA methylation
Specific Cancer Types
Graph Attention Network (GAT)
Dimensional Output: 512, Number of Heads: 8
GATConv_1
Dimensional Output: 256, Number of Heads: 4
GATConv_2
Dimensional Output: 128, Number of Heads: 2
GATConv_3
Dimensional Output: 32, Number of Heads: 1
GATConv_4
32 classes

Figure 2 • The architecture of the GAT model for multiclass cancer classification.

We designed a GAT model to operate on PPI networks represented as graphs. The model takes as input the PPI network structure and node features, where each node represents a protein and the edges represent interactions between proteins. The node features include the multi-omics data, such as gene expression, miRNA expression, and DNA methylation levels. The GAT architecture, depicted in **Figure 2**, consists of four GAT convolutional layers. Batch normalization and Leaky ReLU activation functions are applied after each GAT convolutional layer to normalize and activate the outputs. A dropout layer is added to reduce overfitting by randomly dropping a fraction of the units during training (i.e., in the forward pass, the input data is passed through each GAT convolutional layer, followed by batch normalization, activation, and dropout). The final output is then passed through a softmax function to obtain class probabilities.

3.5. The PPI network

The PPI network is a graphical representation of the connections between proteins in a cell [50]. Proteins work together to perform a variety of biological tasks, including metabolism, gene control, and signaling [51]. Understanding the molecular mechanisms behind cellular functions and illnesses, such as cancer, is made possible with the use of PPI networks [52]. A valuable tool for obtaining anticipated and experimentally verified PPI is the STRING database [53]. For the used TCGA cancer types, the PPI network is shown in **Figure 3**. To create the PPI network, we retrieved the protein interaction data for the genes of the mRNA or RNA-Seq data from the STRING database. The data include information regarding the interacting proteins and the confidence scores of the interactions. Next, the data are processed to build the PPI network as a graph, where each protein is represented as a node and the interactions between proteins are represented as edges. Subsequently, the PPI network is inputted as a graph structure for GAT by coding the edges and nodes as tensors. Each patient is represented by the topology of a PPI network in the graph structure. Multi-omics features from mRNA, miRNA, and DNA methylation data enrich the nodes. Although each patient's network has the same topology, the node feature values differ, representing their unique cancer-specific molecular patterns.

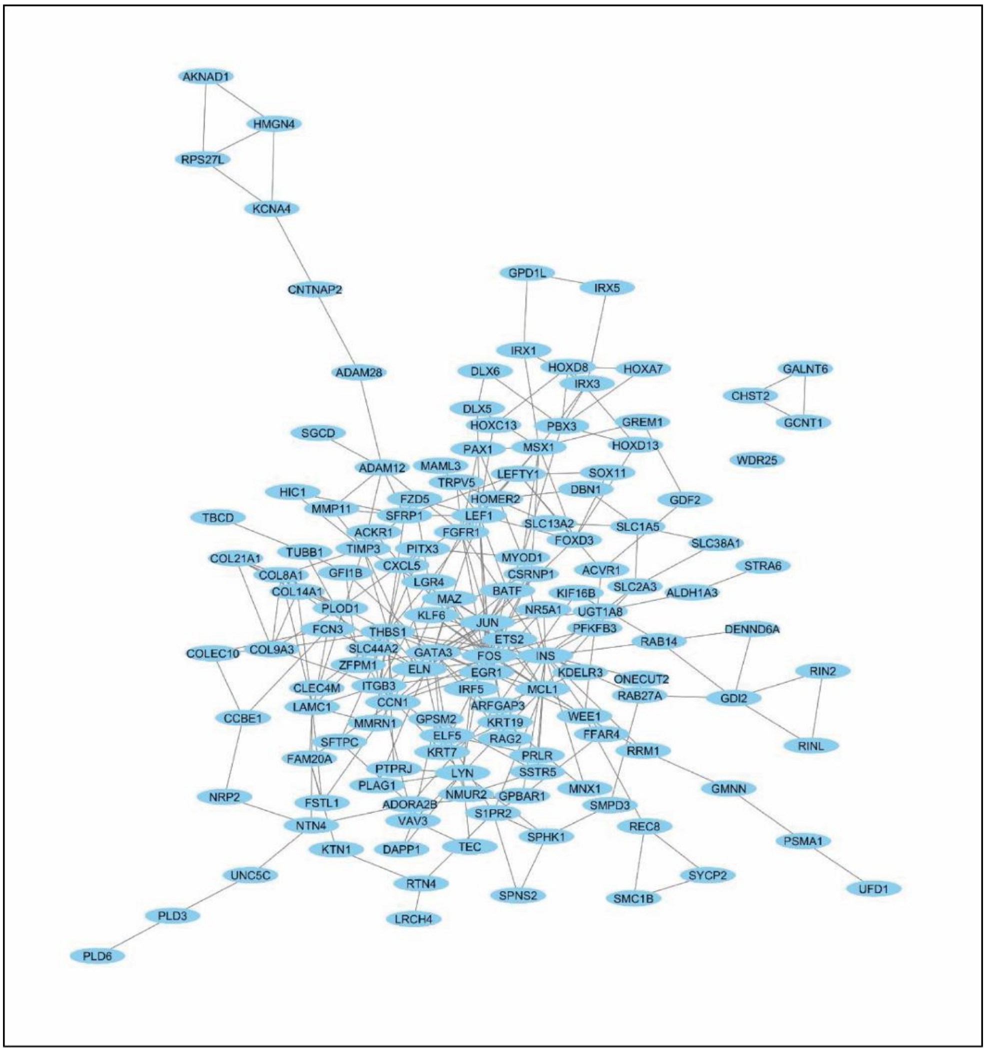

Figure 3 • Protein-Protein Interaction (PPI) network.

3.6. Evaluation metrics

For model evaluation, we used accuracy, precision, recall, and F1 score performance metrics. Accuracy is computed using the following equation and pertains to the corrected classified cancer type:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \times 100\% \quad (4)$$

True positives (TP) denote the proportion of samples correctly classified as positive, true negatives (TN) denote the proportion of samples correctly predicted as negative, false positives (FP) denote the proportion of negative samples incorrectly predicted as positive, and false negatives (FN) represent the proportion of positive samples incorrectly predicted as negative.

Precision measures the ratio of true positives to the total correct classification, where high precision indicates that the model performs well in classifying true positives. Recall is the ratio of true positives to all positives, where high recall indicates that the model can effectively discriminate between cancer types. Precision and recall have trade-offs, therefore improving only one metric does not necessarily result in a more efficient model. The

F1 score is the harmonic mean of recall and precision, representing both recall and precision in one metric. The metrics are frequently expanded to handle multitude of classes in multiclass classification scenarios through the use of macro-averaging (MA) or micro-averaging techniques, which offer a means of summarizing performance across numerous classes. While micro-averaging takes into account the overall counts of true positives, false positives, and false negatives across all classes, MA treats each class equally. We selected MA, which computes metrics independently for each class and then averages them because it gives each class's performance equal weight. Since MA is insensitive to class imbalance and fairly represents overall model performance, every class contributes equally to the final average. MA for recall, precision, and F1 score are calculated as follows:

$$\text{Recall}(\text{MA}) = \frac{1}{K}\sum_{i=1}^{k}\frac{\text{TP}_i}{\text{TP}_i + \text{FN}_i} \times 100\% \tag{5}$$

$$\text{Precision}(\text{MA}) = \frac{1}{K}\sum_{i=1}^{k}\frac{\text{TP}_i}{\text{TP}_i + \text{FP}_i} \times 100\% \tag{6}$$

$$\text{F1}(\text{MA}) = \frac{1}{K}\sum_{i=1}^{k} 2 \times \frac{\text{Recall}_i \times \text{Precision}_i}{\text{Recall}_i + \text{Precision}_i} \times 100\% \tag{7}$$

where k denotes the total number of classes. In our case, $k = 32$, including the normal samples.

4. Results

The values of the performance metrics obtained with the proposed LASSO–MOGAT approach based on PPI networks for cancer classification are presented in **Table 3**. The integration of multi-omics data, including mRNA or RNA-Seq, miRNA, and DNA methylation, yields the highest accuracy of 94.68%, with precision of 90.38%, recall of 89.87%, and an F1 score of 89.87%. The results imply that a representation of cancer biology through multi-omics integration enhances classification accuracy and maintains a well-balanced model with high precision and recall, which is crucial for medical applications. This aligns with the systems biology concept, emphasizing the importance of considering interactions among different biological components.

Table 3 • Performance metrics of the proposed LASSO–MOGAT approach based on PPI network

Data types	Multi-omics data type	Cancer types	Accuracy mean ± std	Precision	Recall	F1 score
Single omics data	mRNA or RNA-Seq	31 types of cancer plus normal tissues	92.08% ± 0.0043	0.8769	0.8716	0.8674
Single omics data	miRNA	31 types of cancer plus normal tissues	87.61% ± 0.0083	0.8117	0.8188	0.8126
Single omics data	DNA methylation	31 types of cancer plus normal tissues	92.23% ± 0.0094	0.8770	0.8599	0.8580
Multi-omics data	mRNA or RNA-Seq and miRNA	31 types of cancer plus normal tissues	92.78% ± 0.0065	0.8902	0.8902	0.8882
Multi-omics data	mRNA or RNA-Seq and DNA methylation	31 types of cancer plus normal tissues	94.05% ± 0.0039	0.8966	0.8954	0.8935
Multi-omics data	miRNA and DNA methylation	31 types of cancer plus normal tissues	93.82% ± 0.0046	0.9035	0.8926	0.8926
Multi-omics data	mRNA or RNA-Seq, miRNA, and DNA methylation	31 types of cancer plus normal tissues	94.68% ± 0.0060	0.9038	0.8987	0.8987

The results in **Table 3** indicate that models based on single omics data types, while still effective, exhibit lower accuracies compared to multi-omics models. For example, the model with miRNA data performs the poorest, with an accuracy of 87.61%, precision of 81.17%, recall of 81.88%, and F1 score of 81.26%. This underlines the advantage of multi-omics integration in improving classification models, as seen in the higher precision, recall, and F1 score values in the multi-omics models.

Interestingly, certain combinations of omics data types, such as mRNA or RNA-Seq with miRNA or DNA methylation, do not provide significant performance improvements, highlighting the complexity of omics data interactions. Conversely, the combination of mRNA and DNA methylation in multi-omics data achieves a high accuracy of 94.05%, with a precision of 89.66%, recall of 89.54%, and F1 score of 89.35%. This indicates potential synergistic effects between these data types.

Overall, the findings suggest that while single omics data types can be effective, integrating multiple omics data types, particularly with GATs based on PPI networks, offers a more comprehensive and accurate approach to cancer classification.

Table 4 shows the comparison between our proposed LASSO–MOGAT model and related models for classifying different types of cancer. The models by Mostavi et al. [22], based on 1D-CNN, 2D-Vanilla-CNN, and 2D-Hybrid-CNN, exhibit high performance using mRNA data for 33 cancer types, with accuracies of 95.50%, 94.87%, and 95.70%, respectively, demonstrating the capability of CNNs in processing omics data. The graph-based model with transformer architecture GTN by Kaczmarek et al. [24] combines mRNA and miRNA data for 12 cancer types and achieves an accuracy of 93.56%. Similarly, the GCNN model by Ramirez et al. [23], based on PPI networks, achieves accuracies of 89.99% and 94.71%, respectively. These models highlight the significance of adding graph structures and multi-omics data to improve classification accuracy. By leveraging a large dataset comprising DNA methylation, miRNA, and mRNA levels for 31 different forms of cancer as well as normal tissues, the proposed LASSO–MOGAT model achieves an accuracy of 94.68% in cancer classification.

Table 4 • Performance metrics for related multi-omics graph methods based on PPI networks

Authors	Models	Pan-cancer	Multi-omics data type						
			mRNA	miRNA	DNA methylation	Accuracy mean ± std	Precision	Recall	F1 score
Mostavi et al. [22]	1D-CNN	33 types of cancer and normal tissues	√			95.50% ± 0.100	–	–	–
	2D-Vanilla-CNN					94.87 % ± 0.040	–	–	–
	2D-Hybrid-CNN					95.70% ± 0.100	–	–	–
Ramirez et al. [23]	GCNN–PPI graph	33 types of cancer and normal tissues	√			89.99% ± 0.883	87.75%	83.79%	–
	GCNN–PPI + singleton graph					94.71% ± 0.107	92.76%	92.19%	–
Kaczmarek et al. [24]	GTN	12 types of cancer	√	√		93.56% ± 0.910	–	–	–
Proposed LASSO–MOGAT		31 types of cancer and normal tissues	√	√	√	94.68% ± 0.006	90.38%	89.87%	89.87%

5. Conclusions

This study demonstrates the potential of integrating multi-omics data with advanced graph-based deep learning models for cancer classification. The proposed LASSO–MOGAT framework effectively combines RNA-Seq, miRNA, and DNA methylation data with GATs and PPI networks, capturing the intricate relationships within cancer biology. The experimental results, validated through fivefold cross-validation, highlight the performance of LASSO–MOGAT in terms of accuracy, precision, recall, and F1 score. By employing DEG with LIMMA and LASSO regression for feature selection, this method offers a robust and comprehensive understanding of cancer molecular mechanisms. These findings underscore the critical role of multi-omics integration and graph-based models in enhancing cancer classification accuracy and reliability, paving the way for improved diagnostic and therapeutic strategies in precision oncology.

Funding

The authors declare no financial support for the research, authorship, or publication of this article.

Author contributions

Conceptualization, F.A. and A.V.; methodology, F.A., A.V., M.K.E., and M.M.; software, F.A.; validation, M.K.E. and M.M.; data curation, F.A.; writing—original draft preparation, F.A.; writing—review and editing, A.V. All authors have read and agreed to the published version of the manuscript.

Conflict of interest

The authors declare no conflict of interest.

Data availability statement

Data supporting these findings are available within the article, at https://doi.org/10.20935/AcadBiol7325, or upon request.

Institutional review board statement

Not applicable.

Informed consent statement

Not applicable.

Sample availability

The authors declare no physical samples were used in the study.

Additional information

Received: 2024-06-01

Accepted: 2024-08-16

Published: 2024-08-30

Academia Biology papers should be cited as *Academia Biology 2024*, ISSN 2837-4010, https://doi.org/10.20935/AcadBiol7325. The journal's official abbreviation is *Acad. Biol.*

Publisher's note

Academia.edu Journals stays neutral with regard to jurisdictional claims in published maps and institutional affiliations. All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright

References

1. Alharbi F, Vakanski A. Machine learning methods for cancer classification using gene expression data: a review. Bioengineering. 2023;10(2):173. doi: 10.3390/bioengineering10020173
2. Pfeifer B, Saranti A, Holzinger A. GNN-SubNet: disease subnetwork detection with explainable graph neural networks. Bioinformatics. 2022; 38(Suppl 2):ii120–6. doi: 10.1093/bioinformatics/btac478
3. Wekesa JS, Kimwele M. A review of multi-omics data integration through deep learning approaches for disease diagnosis, prognosis, and treatment. Front Genet. 2023;14: 1199087. doi: 10.3389/fgene.2023.1199087
4. Leng D, Zheng L, Wen Y, Zhang Y, Wu L, Wang J, et al. A benchmark study of deep learning-based multi-omics data fusion methods for cancer. Genome Biol. 2022;23(1):171. doi: 10.1186/s13059-022-02739-2
5. Gogoshin G, Rodin AS. Graph neural networks in cancer and oncology research: emerging and future trends. Cancers. 2023;15(24):5858. doi: 10.3390/cancers15245858
6. Li B, Nabavi S. A multimodal graph neural network framework for cancer molecular subtype classification. BMC Bioinform. 2024;25(1):27. doi: 10.1186/s12859-023-05622-4
7. Tanvir RB, Islam MM, Sobhan M, Luo D, Mondal AM. MOGAT: a multi-omics integration framework using graph attention networks for cancer subtype prediction. Int J Mol Sci. 2024;25(5):2788. doi: 10.3390/ijms25052788
8. Narrandes S, Xu W. Gene expression detection assay for cancer clinical use. J Cancer. 2018;9(13):2249–65. doi: 10.7150/jca.24744
9. Singh KP, Miaskowski C, Dhruva AA, Flowers E, Kober KM. Mechanisms and measurement of changes in gene expression. Biol Res Nurs. 20(4):369–82. doi: 10.1177/1099800418772161
10. Li M, Sun Q, Wang X. Transcriptional landscape of human cancers. Oncotarget. 2017;8(21):34534. doi: 10.18632/oncotarget.15837
11. Heo YJ, Hwa C, Lee GH, Park JM, An JY. Integrative multi-omics approaches in cancer research: from biological networks to clinical subtypes. Mol Cells. 2021;44(7):433–43. doi: 10.14348/molcells.2021.0042.
12. Menyhárt O, Győrffy B. Multi-omics approaches in cancer research with applications in tumor subtyping, prognosis, and diagnosis. Comput Struct Biotechnol J. 2021;19: 949–60. doi: 10.1016/j.csbj.2021.01.009
13. Geissler F, Nesic K, Kondrashova O, Dobrovic A, Swisher, EM, Scott CL, et al. The role of aberrant DNA methylation in cancer initiation and clinical impacts. Ther Adv Med Oncol. 2024;16:17588359231220511. doi: 10.1177/17588359231220511
14. Ankasha SJ, Shafiee MN, Wahab NA, Ali RAR, Mokhtar NM. Post-transcriptional regulation of microRNAs in cancer: from prediction to validation. Oncol Rev. 2018;12(1). doi: 10.4081/oncol.2018.344
15. Mohamed TI, Ezugwu AE. Enhancing lung cancer classification and prediction with deep learning and multi-omics data. IEEE Access. 2024;12:59880–92. doi: 10.1109/ACCESS.2024.3394030
16. Qiu L, Li H, Wang M, Wang X. Gated graph attention network for cancer prediction. Sensors. 2021;21(6): 1938. doi: 10.3390/s21061938
17. Baul S, Ahmed KT, Filipek J, Zhang W. omicsGAT: Graph attention network for cancer subtype analyses. Int J Mol Sci. 2022;23(18):10220. doi: 10.3390/ijms231810220
18. Zhao W, Gu X, Chen S, Wu, J, Zhou Z. MODIG: integrating multi-omics and multi-dimensional gene network for cancer driver gene identification based on graph attention network model. Bioinformatics. 2022;38(21):4901–7. doi: 10.1093/bioinformatics/btac622
19. Jeong D, Koo B, Oh M, Kim TB, Kim S. GOAT: gene-level biomarker discovery from multi-omics data using graph attention neural network for eosinophilic asthma subtype. Bioinformatics. 2023;39(10):btad582. doi: 10.1093/bioinformatics/btad582
20. Shi H, Gu Y, Zhang H, Li X, Cao Y. MORGAT: a model based knowledge-informed multi-omics integration and robust graph attention network for molecular subtyping of cancer. In: International Conference on Intelligent Computing; 2023 Jul; Singapore: Springer Nature Singapore; 2023.
21. Yang K, Cheng J, Cao S, Pan X, Shen, HB, Jin C, et al. Integration of multi-source gene interaction networks and omics data with graph attention networks to identify novel disease genes. bioRxiv. 2023;12. doi: 10.1101/2023.12.03.569371
22. Mostavi M, Chiu YC, Huang Y, Chen Y. Convolutional neural network models for cancer type prediction based on gene expression. BMC Med Genomics. 2020;13:1–13. doi: 10.1186/s12920-020-0677-2
23. Ramirez R, Chiu YC, Hererra A, Mostavi M, Ramirez J, Chen Y, et al. Classification of cancer types using graph convolutional neural networks. Front Phys. 2020;8:203. doi: 10.3389/fphy.2020.00203
24. Kaczmarek E, Jamzad A, Imtiaz T, Nanayakkara J, Renwick N, Mousavi P. Multi-omic graph transformers for cancer classification and interpretation. Pac Symp Biocomput. 2022;27:373–84. doi: 10.1142/9789811250477_0034

25. Moon S, Lee H. MOMA: a multi-task attention learning algorithm for multi-omics data interpretation and classification. Bioinformatics. 2022; 38(8):2287–96. doi: 10.1093/bioinformatics/btac080

26. Zhang G, Peng Z, Yan C, Wang J, Luo J, Luo, H. MultiGATAE: a novel cancer subtype identification method based on multi-omics and attention mechanism. Front Genet. 2022;13:855629. doi: 10.3389/fgene.2022.855629

27. Sun Q, Cheng L, Zhang L. SADLN: self-attention based deep learning network of integrating multi-omics data for cancer subtype recognition. Front Genet. 2023;13:1032768. doi: 10.3389/fgene.2022.1032768

28. Shanthamallu US, Thiagarajan JJ, Song H, Spanias A. Gramme: semisupervised learning using multilayered graph attention models. IEEE Trans Neural Netw Learn Syst. 2019;31(10):3977–88. doi: 10.1109/TNNLS.2019.2948797

29. Ouyang D, Liang Y, Li L, Ai N, Lu S, Yu M, et al. Integration of multi-omics data using adaptive graph learning and attention mechanism for patient classification and biomarker identification. Comput Biol Med. 2023;164:107303. doi: 10.1016/j.compbiomed.2023.107303

30. Gong P, Cheng L, Zhang Z, Meng A, Li E, Chen J, et al. Multi-omics integration method based on attention deep learning network for biomedical data classification. Comput Methods Programs Biomed. 2023;231:107377. doi: 10.1016/j.cmpb.2023.107377

31. Song H, Yin C, Li Z, Feng K, Cao Y, Gu Y, et al. Identification of cancer driver genes by integrating multiomics data with graph neural networks. Metabolites. 2023;13(3):339. doi: 10.3390/metabo13030339

32. Zhang X, Zhang J, Sun K, Yang X, Dai C, Guo Y. Integrated multi-omics analysis using variational autoencoders: application to pan-cancer classification. In: IEEE International Conference on Bioinformatics and Biomedicine (BIBM); 2019 Nov. Bengaluru: IEEE; 2019. p. 765–9.

33. Chai H, Zhou X, Zhang Z, Rao J, Zhao H, Yang Y. Integrating multi-omics data through deep learning for accurate cancer prognosis prediction. Comput Biol Med. 2021;134:104481. doi: 10.1016/j.compbiomed.2021.104481

34. Li X, Ma J, Leng L. MoGCN: a multi-omics integration method based on graph convolutional network for cancer subtype analysis. Front Genet. 2022;13:806842. doi: 10.3389/fgene.2022.806842

35. Zhou N, Wang S, Tan Z. AEMVC: anchor enhanced multi-omics cancer subtype identification. Proceedings of the 3rd International Symposium on Artificial Intelligence for Medicine Sciences; 2022 Oct. New York: Association for Computing Machinery; 2022. p. 57–63.

36. Khadirnaikar S, Shukla S, Prasanna SRM. Integration of pan-cancer multi-omics data for novel mixed subgroup identification using machine learning methods. PLoS One. 2023;18(10):e0287176. doi: 10.1371/journal.pone.0287176

37. Zhu J, Oh JH, Simhal AK, Elkin R, Norton L, Deasy JO, et al. Geometric graph neural networks on multi-omics data to predict cancer survival outcomes. Comput Biol Med. 2023;163:107117. doi: 10.1016/j.compbiomed.2023.107117

38. Xiao S, Lin H, Wang C, Wang S, Rajapakse JC. Graph neural networks with multiple prior knowledge for multi-omics data analysis. IEEE J Biomed Health Inform. 2023;27(9):4591–600. doi: 10.1109/JBHI.2023.3284794

39. Chatzianastasis M, Vazirgiannis M, Zhang Z. Explainable multilayer graph neural network for cancer gene prediction. Bioinformatics. 2023;39(11):btad643. doi: 10.1093/bioinformatics/btad643

40. Wang J, Liao N, Du X, Chen Q, Wei B. A semi-supervised approach for the integration of multi-omics data based on transformer multi-head self-attention mechanism and graph convolutional networks. BMC Genomics. 2024;25(1):86. doi: 10.1186/s12864-024-09985-7

41. Yao D, Li B, Zhan X, Zhan X, Yu L. GCNFORMER: graph convolutional network and transformer for predicting lnc RNA-disease associations. BMC Bioinformatics. 2024;25(1):5. doi: 10.1186/s12859-023-05625-1

42. Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, Ellrott K, et al. The cancer genome atlas pan-cancer analysis project. Nat Genet. 2013;45(10):1113–20. doi: 10.1038/ng.2764

43. Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Garolini D, et al. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. Nucleic Acids Res. 2016;44(8):e71. doi: 10.1093/nar/gkv1507

44. Rapaport F, Khanin R, Liang Y, Pirun M, Krek A, Zumbo P, et al. Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. Genome Biol. 2013;14:1–13. doi: 10.1186/gb-2013-14-9-r95

45. Chen J, Long MD, Sribenja S, Ma SJ, Yan L, Hu Q, et al. An epigenome-wide analysis of socioeconomic position and tumor DNA methylation in breast cancer patients. Clin Epigenetics. 2023;15(1):68. doi: 10.1186/s13148-023-01470-4

46. Pidsley R, Wong CCY, Volta M, Lunnon K, Mill J, Schalkwyk LC. A data-driven approach to preprocessing Illumina 450K methylation array data. BMC Genomics. 2013;14:1–10. doi: 10.1186/1471-2164-14-293

47. Wang Z, Fu G, Ma G, Wang C, Wang Q, Lu C, et al. The association between DNA methylation and human height and a prospective model of DNA methylation-based height prediction. Human Genet. 2024;143(3):401–21. doi: 10.1007/s00439-024-02659-0

48. Sheng J, Zhang Y, Wang B, Chang Y. MGATs: motif-based graph attention networks. Mathematics. 2024;(12):293. doi: 10.3390/math12020293

49. Lazaros K, Koumadorakis DE, Vlamos P, Vrahatis A. Graph neural network approaches for single-cell data: a recent overview. Neural Comput Appl. 2024;36(17):1–25. doi: 10.1007/s00521-024-09662-6

50. Zainal-Abidin RA, Afiqah-Aleng N, Abdullah-Zawawi MR, Harun S, Mohamed-Hussein ZA. Protein-protein interaction (PPI) network of Zebrafish Oestrogen receptors: a bioinformatics workflow. Life. 2022;12(5):650. doi: 10.3390/life12050650

51. Morris R, Black KA, Stollar EJ. Uncovering protein function: from classification to complexes. Essays Biochem. 2022; 6(3):255–85. doi: 10.1042/EBC20200108

52. Hu H, Wang H, Yang X, Li X, Zhan W, Zhu H, et al. Network pharmacology analysis reveals potential targets and mechanisms of proton pump inhibitors in breast cancer with diabetes. Sci Rep. 2023;13:7623. doi: 10.1038/s41598-023-34524-x

53. Szklarczyk D, Gable AL, Nastou KC, Lyon D, Kirsch R, Pyysalo S, et al. The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. Nucleic Acid Res. 2021;49(D1):D605–12. doi: 10.1093/nar/gkaa1074