RESEARCH

Open Access

Amogel: a multi-omics classification framework using associative graph neural networks with prior knowledge for biomarker identification

Chia Yan Tan[1*†], Huey Fang Ong[1†], Chern Hong Lim[1†], Mei Sze Tan[1†], Ean Hin Ooi[2] and KokSheik Wong[1]

†Chia Yan Tan, Huey Fang Ong, Chern Hong Lim and Mei Sze Tan have contributed equally to this work.

*Correspondence: chia.tan@monash.edu

[1] School of Information Technology, Monash University Malaysia, Jalan Lagoon Selatan, 47500 Petaling Jaya, Selangor, Malaysia
[2] School of Engineering, Monash University Malaysia, Jalan Lagoon Selatan, 47500 Petaling Jaya, Selangor, Malaysia

Abstract

The advent of high-throughput sequencing technologies, such as DNA microarray and DNA sequencing, has enabled effective analysis of cancer subtypes and targeted treatment. Furthermore, numerous studies have highlighted the capability of graph neural networks (GNN) to model complex biological systems and capture non-linear interactions in high-throughput data. GNN has proven to be useful in leveraging multiple types of omics data, including prior biological knowledge from various sources, such as transcriptomics, genomics, proteomics, and metabolomics, to improve cancer classification. However, current works do not fully utilize the non-linear learning potential of GNN and lack of the integration ability to analyse high-throughput multi-omics data simultaneously with prior biological knowledge. Nevertheless, relying on limited prior knowledge in generating gene graphs might lead to less accurate classification due to undiscovered significant gene-gene interactions, which may require expert intervention and can be time-consuming. Hence, this study proposes a graph classification model called associative multi-omics graph embedding learning (AMOGEL) to effectively integrate multi-omics datasets and prior knowledge through GNN coupled with association rule mining (ARM). AMOGEL employs an early fusion technique using ARM to mine intra-omics and inter-omics relationships, forming a multi-omics synthetic information graph before the model training. Moreover, AMOGEL introduces multi-dimensional edges, with multi-omics gene associations or edges as the main contributors and prior knowledge edges as auxiliary contributors. Additionally, it uses a gene ranking technique based on attention scores, considering the relationships between neighbouring genes. Several experiments were performed on BRCA and KIPAN cancer subtypes to demonstrate the integration of multi-omics datasets (miRNA, mRNA, and DNA methylation) with prior biological knowledge of protein-protein interactions, KEGG pathways and Gene Ontology. The experimental results showed that the AMOGEL outperformed the current state-of-the-art models in terms of classification accuracy, F1 score and AUC score. The findings of this study represent a crucial step forward in advancing the effective integration of multi-omics data and prior knowledge to improve cancer subtype classification.

Keywords: Graph neural network, Association rule mining, Graph classification, Multi-omics, Prior knowledge

Introduction

Cancer remains one of the leading causes of death in this modern day, with millions of new cases diagnosed each year. Although significant advancement in early detection and treatment has been made, the heterogeneity of cancer poses a substantial challenge. Cancer researchers have identified numerous subtypes of cancer where their molecular and clinical characteristics are different for the same cancer type [1, 2]. Understanding these differences and accurately identifying these subtypes are paramount for developing an effective treatment. Over the years, high throughput technologies such as microarray, next-generation sequencing and mass spectrometry have enabled multiple omics (multi-omics) data generation and analysis for various molecular processes. For example, it allows the comparison of gene expression patterns between different groups, such as cancer subtypes [3, 4]. With the advent of these technologies, the focus of cancer research has thus shifted from single-omics analysis to multi-omics integrative analysis.

Existing studies [5–11] have demonstrated improved cancer subtype classification with multi-omics data integration. However, a carefully crafted solution is required due to the high dimensionality of omics data, where there is a large number of features compared with a relatively low sample size [12]. The high-dimensional nature of omics data often results in model overfitting, where the learning of the classification model is biased by noise rather than meaningful biological information [13]. Additionally, it also leads to computational inefficiencies and causes memory shortages due to the vast search spaces required for data analysis. Therefore, numerous studies have addressed these challenges by introducing novel dimensionality reduction techniques [14–18], which have proven not only to enhance cancer classification performance but also lead to better biomarker discovery by removing irrelevant features and identifying potential candidate biomarkers.

The recent development of deep learning models has led to techniques such as deep neural networks (DNN) and convolutional neural networks (CNN) being adopted for multi-omics integrative analysis. Compared to conventional machine learning models, deep learning models have a higher capability of learning complex patterns and representation with minimal feature engineering [19]. Often, most of the proposed deep learning models on multi-omics integrative analysis use early fusion methods with concatenation technique, where multi-omics data for each sample are combined and fed into the deep learning model to encode features for classification tasks. In contrast, late fusion methods employ separate deep learning models to encode feature representations for each omics data type, followed by a concatenation of feature representations and a final deep learning layer to learn the final classification. For instance, [10] proposed a DNN-based multi-omics integration framework with an auxiliary classifiers-enhanced autoencoder (MOCAT) for cancer subtype classification. In this framework, multi-omics data were fed into the DNN autoencoder to learn compact representation and extract omics-specific features. These encoded omics-specific features were fused by concatenation and feed into the final multi-head attention autoencoder. The multi-head attention mechanism emphasized the distinct significance of various omics modalities since different types of omics data contribute differently to the aggregated predictive accuracy.

Traditional deep learning models such as DNN are effective at extracting feature embedding in row-column format, while CNN are suitable for unstructured data, such as image or audio data. Both models operate in the Euclidean domain, but due to the complex nature of biological organisms, the relationships among genes are better represented in a graph structure than in the Euclidean domain. To address this gap, graph neural networks (GNN), a sub-field of deep learning model, are increasingly used in biomedical research. GNN provide improved performance and interpretability due to their ability to model graph structures. [5] proposed multi-omics graph convolutional networks (MOGONET) for biomedical classification. The study proposed modality-specific graph convolutional networks (GCN) for multi-omics mRNA, miRNA and DNA methylation data for feature learning with a view correlation discovery network as the late fusion technique for final prediction. On the other hand, [11] proposed a multi-omics integration strategy using adaptive graph learning and attention mechanism (MOGLAM) that introduced adaptive learning on sample similarity network for each omics and fused the omic-specific representation learning using the multi-omics attention mechanism. Similarly, [6] proposed a multi-omics graph attention network (MOGAT) to improve cancer subtype prediction by integrating eight types of omics data and using a graph attention network (GAT). This method addresses limitations in existing multi-omics integration approaches by leveraging the attention mechanism in GAT to enhance the extraction of significant features from multi-omics data. A patient similarity graph is constructed for each omics data, while patient embedding consists of all eight types of omics expression. The final omics-specific embeddings of the patients were trained using GAT and concatenated to form the final embedding, which was then used for subtype prediction, visualization, and survival analysis.

In addition, some studies incorporate prior biological knowledge from related source domains, such as protein-protein interactions (PPI), as the input graph to further improve the model performance of multi-omics cancer classification. For example, [9] proposed a novel end-to-end deep learning model incorporating prior knowledge and multi-omics data to classify molecular subtypes. Prior knowledge data were used to form a single unified network, and the multi-omics data were concatenated as gene features of the graph, leveraging GCN to learn graph embedding and parallel network as a global feature extractor. In a subsequent study, the authors proposed an enhanced version that integrates multi-omics data in the form of heterogeneous multi-layer graphs, combining both inter-omics and intra-omic connections from prior biological knowledge [8]. In another work, [7] proposed a multiple prior knowledge into graph neural network (MPK-GNN) framework with four main modules. Multi-omics data were concatenated, and sample feature was extracted using a DNN-based sample module, while the GNN-based feature-level module was used to extract features from the prior knowledge graph together with multi-omics data. These studies highlighted the potential of GNN-based frameworks to enhance the analysis and interpretation of multi-omics data by effectively integrating prior biological knowledge. However, gene-based graph models that are based on prior knowledge are challenging as the quality of the extracted features by graph neural network depends on the completeness of the prior knowledge [8, 9].

Besides incorporating prior knowledge, increasing the number of omics data types in integrative modeling is believed to enhance model performance [6]. However, most studies rely on typical early fusion methods, such as concatenation, which fail to account for the varying contributions of each omics data type to the final classification outcome. In contrast, separated feature extractors are utilized for late fusion methods, mainly based on deep neural networks. Nonetheless, this approach increases computational demands as the total input features grow parallel with the number of omics data types. This is evident in the MOGAT model [6], where the experiments were conducted with expensive eight NVIDIA A100 GPUs with large 40GB of GPU memory to handle the integration of eight omics types to the model. In terms of biomarker discovery, some existing studies [5, 10] use feature ablation studies and select potential biomarkers based on feature importance score. The feature importance score is calculated by evaluating the accuracy degradation when the feature is eliminated from the proposed model. Similarly, the discovery becomes challenging when the number of features, such as genes, increases due to more omics data types. Therefore, there are studies [11, 20] that use attention mechanisms or integrated gradients to select the biomarkers as it can avoid gene explosion when the number of genes increases. However, these methods do not consider the gene-gene interactions when ranking the genes. Due to complex human biological gene interaction, some genes might not appear important individually but could play a crucial role when interacting with other genes. This limitation poses a significant challenge when using methods that rank genes solely based on their individual importance scores.

To address the aforementioned limitations stated above, this study proposes a new method called **A**ssociative **M**ulti-**O**mics **G**raph **E**mbedding **L**earning (**AMOGEL**). The main contributions of AMOGEL are summarized as follows:

- Adopting the association rule mining (ARM) model as early fusion multi-omics integrative analysis to mine relationship between the inter-omics datasets, forming information-based multi-omics graph before model training, since early fusion with inter-omics analysis can remove noise (irrelevant genes) before model training.
- Proposing multi-dimensional edges graph with ARM information-based content graph as the main contributor, while the prior knowledge graph act as auxiliary contributor.
- Introducing the gene ranking with consideration of gene-gene interaction contribution by using gene-gene edges attention score.

Methods

The proposed AMOGEL method can be divided into seven parts: (i) data preparation, (ii) data preprocessing, (iii) classification association rule mining and ranking, (iv) graph feature learning, (v) global feature learning, (vi) cancer subtype classification and (vii) gene ranking and biomarkers selection. Fig. 1 shows the overview

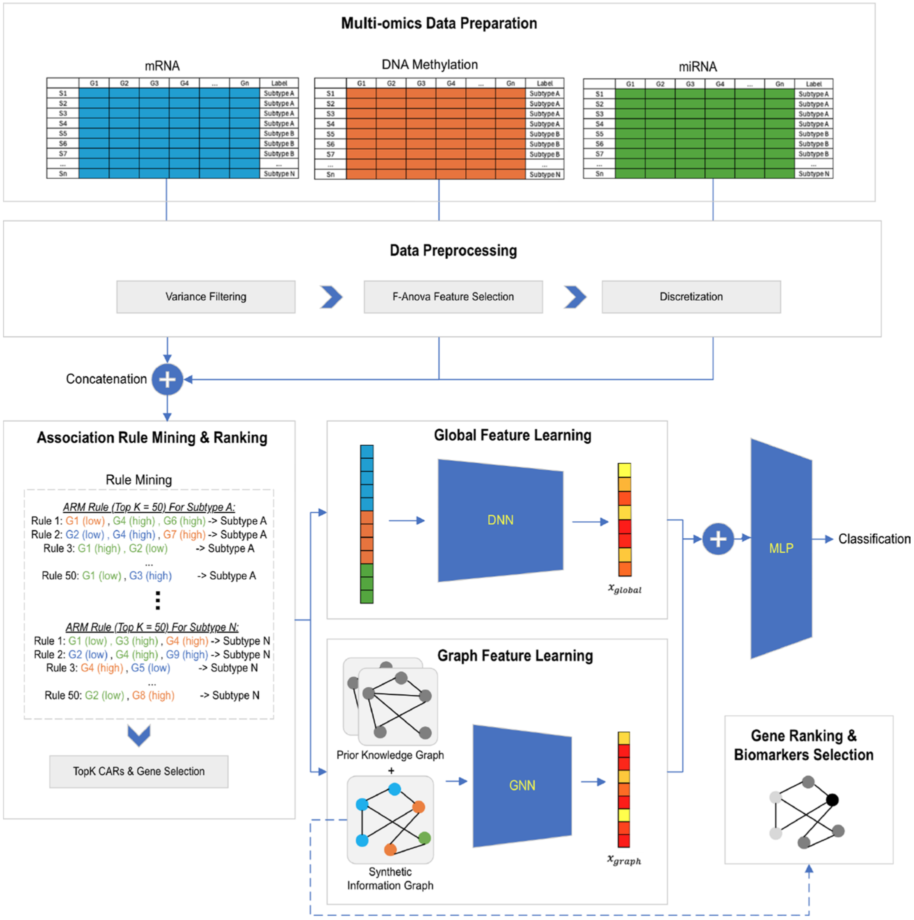

Multi-omics Data Preparation
mRNA
DNA Methylation
miRNA
G1
G2
G3
G4
...
Gn
Label
S1
S2
S3
S4
S5
S6
S7
...
Sn
Subtype A
Subtype B
Subtype N
Data Preprocessing
Variance Filtering
F-Anova Feature Selection
Discretization
Concatenation
Association Rule Mining & Ranking
Rule Mining
ARM Rule (Top K = 50) For Subtype A:
Rule 1: G1 (low) , G4 (high) , G6 (high) -> Subtype A
Rule 2: G2 (low) , G4 (high) , G7 (high) -> Subtype A
Rule 3: G1 (high) , G2 (low) -> Subtype A
...
Rule 50: G1 (low) , G3 (high) -> Subtype A
ARM Rule (Top K = 50) For Subtype N:
Rule 1: G1 (low) , G3 (high) , G4 (high) -> Subtype N
Rule 2: G2 (low) , G4 (high) , G9 (high) -> Subtype N
Rule 3: G4 (high) , G5 (low) -> Subtype N
...
Rule 50: G2 (low) , G8 (high) -> Subtype N
TopK CARs & Gene Selection
Global Feature Learning
DNN
x_{global}
Graph Feature Learning
Prior Knowledge Graph
+
Synthetic Information Graph
GNN
x_{graph}
MLP
Classification
Gene Ranking & Biomarkers Selection

Fig. 1 The overall framework of the proposed associative multi-omics graph embedding learning (**AMOGEL**)

architecture of our proposed method - AMOGEL. The details of each process will be discussed in the subsequent sections.

Data preparation

The same datasets from [11] were used to evaluate the performance of our proposed method - AMOGEL. Cancer subtypes in the breast invasive carcinoma (BRCA) and the pan-kidney (KIPAN) datasets were downloaded from the TCGAbiolink (*https://bioconductor.org/packages/release/bioc/html/TCGAbiolinks.html*) and Broad GDAC Firehouse (*https://gdac.broadinstitute.org/*), respectively. Moreover, three omics data types for each dataset, namely miRNA expression, mRNA expression and DNA Methylation, were chosen due to their highly correlated samples among multiple omics data types. Apart from these multi-omics data, prior biological knowledge, including the KEGG pathway, Gene Ontology (GO) and protein-protein interaction (PPI) networks, are integrated into our study to further improve the model. PPI networks provide the interaction between protein genes with interaction confidence scores based on evidence. Meanwhile, the KEGG Pathway is a collection of biological pathways representing molecular interactions within cells, while the Gene Ontology is the vast knowledge database that unifies the representation of gene and gene product attributes across multiple species, including humans. The PPI dataset was downloaded from the STRING database [21], while the KEGG pathway and GO were retrieved from Database for Annotation, Visualization and Integrated Discovery (DAVID) [22, 23].

Data preprocessing

Retrieved multi-omics datasets are highly noisy and consist of redundant genes. For BRCA, the mRNA dataset consists of 1212 samples with 20531 genes, the DNA methylation dataset has 885 patient samples with 20106 genes, and the miRNA consists of 1189 patient samples with 503 features. While for KIPAN, the mRNA dataset has 1020 samples with 20531 genes, the DNA methylation dataset consists of 867 patient samples with 20116 genes, and the miRNA is made up of 1005 patient samples with 472 features. Considering the large number of features or genes in the datasets, removing irrelevant and redundant ones is essential to ensure better model interpretation and performance. First, genes with missing values were filtered out and duplicated genes and samples were aggregated before the final mean values were calculated. To correlate each sample across different omics datatypes, the datasets were cross-checked to filter out those that do not exist in these three types of omics datasets. Besides, low-variance genes whose gene expressions do not reflect any changes to different classes were removed from the datasets. A variance threshold of 0.001 was used to filter low-variance genes for mRNA, miRNA and DNA Methylation.

In addition to that, the omics data may still have non-significant features that might result in poor classification performance. Thus, the ANOVA F-value statistical test was used to select only statistically significant features for each omic. Based on the works in [5, 11], to avoid selecting only highly correlated features and ignoring complementary information from less relevant features, the first principle component of the data after feature selection should explain < 50% of the variance. Additionally, the best-performing accuracy among different numbers of input features per omics type is evaluated and selected by using the proposed method. Table 1 summarizes the KIPAN and BRCA datasets, the number of samples for each cancer subtype, and the number of genes before and after the data preprocessing.

Table 1 Summary of the multi-omics datasets

Dataset	Subtypes and number of samples after data preprocessing	Number of features in mRNA, DNA Methylation and miRNA	Number of features after data preprocessing[a]
BRCA	Normal-like: 115 Basal-like: 131 HER2-enriched: 46 Luminal A: 546 Luminal B: 147	20531, 20106, 503	1000, 1000, 502
KIPAN	KICH: 66 KIRC: 318 KIRP: 273	20531, 20111, 472	2000, 2000, 471

[a] To ensure the fair comparison, ANOVA-F filtering threshold for BRCA is 1000 and KIPAN is 2000, same threshold reported in MOGONET [5] and MOGLAM [11]

Classification association rule mining and ranking

Due to the complex biological systems, a gene network generated using the prior knowledge (PPI/KEGG pathway/GO) is inadequate to represent the complex human biology system [8, 9]. Therefore, in this research, we proposed to use the association rule mining technique to mine intra-omics and inter-omics gene association interaction as an early fusion technique. The generated class association rules (CARs), which consist of genes and the cancer subtypes, were subsequently used to form the gene network as the main contributor in our proposed classification model. These generated rules are easily interpretable, making this method widely preferred across various fields, such as market basket analysis [24], to identify associations among items that are frequently purchased together.

Association rule mining (ARM) is a technique that is used to discover interesting patterns, relationships, and associations among a set of variables in large datasets. The goal of ARM is to identify rules that describe the co-occurrence of items in the transaction database. The generated rules can be expressed in the form of "if A, then B" where A is called antecedent, and B is called consequent. To discover association rules, ARM finds frequent item sets from large transaction databases by counting the frequency of occurrence of a particular item set. Those items that are more than *minimum support* are defined as frequent itemsets. *Minimum confidence* user-defined threshold was also used to measure the strength of an association rule. The support and confidence of the rules can be calculated using the following conditional probability expressed as Eq. 1 and Eq. 2 below:

$$support(A \Rightarrow B) = P(A \cup B); \tag{1}$$

$$confidence(A \Rightarrow B) = P(B|A). \tag{2}$$

Association classification (AC), on the other hand, combines elements of both classification and association rule mining. In AC, the antecedent of the rules will be the itemsets from the transaction database, and the consequent is the class label.

Instead of the late fusion method, mRNA, DNA Methylation and miRNA features were concatenated into a single view before model learning, as compared to other related work [6, 10, 11], which integrated multi-omics by combining the results generated from different omics-specific models. First, the feature columns of multi-omics datasets, mRNA, DNA methylation and miRNA, were concatenated across the row with the same patient index. The combined feature columns were mapped with unique identifiers to uniquely identify the overlapping gene names that exist in mRNA and DNA methylation. As a result, the BRCA dataset has a total of 985 samples and 2502 multi-omics features, while the KIPAN dataset has a total of 657 samples and 4471 multi-omics features, as shown in Table 1. The datasets were then split into 70% training samples, and 30% testing samples, and the expression value of each feature was discretized into either a low, medium or high value based on the distribution value of each feature from the training dataset. The discretization process is also applied to the test dataset based on the feature distribution from the training dataset.

From existing studies [17, 25, 25–29], ARM typically generates large frequent itemsets, especially in high-dimensional gene expression data. Thus, this study proposed an iterative minimum support search algorithm (see Algorithm 1) to generate a minimum set of closed frequent item sets [30] for each class subset. Transaction databases were generated based on the class output of each sample and cancer subtype, and the initial minimum support value was calculated for each transaction database based on the total number of transactions. By using the initial minimum support value, a list of frequent itemsets was generated using the ARM technique. This process was repeated by decreasing the minimum support value until the generated list of frequent itemsets passed the maximum rules count. By doing so, it helps to prevent class rules imbalance due to imbalance class from the original datasets. The generated frequent itemsets represent the antecedents, while their respective class labels represent the consequences of the generated CARs. Next, weak CARs were filtered out using the *minimum confident threshold*.

Algorithm 1 Iterative minimum support search

```
$min\_support = length(class)$
$frequent\_item\_sets = generate\_close\_itemsets(min\_support)$
while $len(frequent\_item\_sets) < minimum\_rule\_count$ do
    $min\_support = min\_support - 1$
    $frequent\_item\_sets = generate\_close\_itemsets(min\_support)$
end while
$Filter frequent\_item\_set >= min\_confidence\, threshold$
```

The generated CARs that met the user-defined threshold *min_support* and *min_confidence* are considered strong rules. However, strong rules are not necessarily interesting. In order to rank the rules based on their interestingness, modified information-content measurement [14, 31] was used, which is expressed in Eq. 3. In the equation, confidence was added to the interestingness ranking to further enhance the quality of the ranked rules. Top generated CARs (top-*k*) from each class subtype were filtered, and the *k* threshold in this study was set to 1000. The Top-*k* CARs were then used for the graph feature learning and global feature learning, which are detailed in the subsequent subsections.

$$\text{information-content} = log_2(\text{information gain}) + log_2(\text{correlation}) + log_2(\text{confidence}). \tag{3}$$

Graph feature learning

Generally, a network graph can be represented using the notation $G = \{V, E\}$, where V is the set of genes, and E is the edge between genes. For the proposed AMOGEL, a synthetic information graph was generated from the information-content-based top-*k* CARs, and prior knowledge graphs were generated from prior biological knowledge gathered from the PPI, KEGG pathway and GO databases. Both types of graphs were unified to form the final static graph.

Synthetic information graph

Distinct genes that exist from the selected top-k CARs were represented as nodes of the graph. The information-content between genes was represented by an edge between nodes. Given gene i and gene j, the edge between genes can be defined using Eq. 4:

$$e_{ij}^{information} = \frac{infogain_i + infogain_j + correlation_i + correlation_j}{4}, \tag{4}$$

where the $infogain_i$ is defined as mutual information between gene i and the cancer subtypes, while the $correlation_i$ is the correlation between gene i and the class. The gene i and gene j belong to any specific rule R within the set of CARs, $\{i,j\} \subseteq R, \quad R \in CARs$. The adjacency matrix for the synthetic information graph $A_{ij}^{information} \in \mathbb{R}^{m \times m}$ can be constructed using $A_{ij} = e_{ij}$, where m is the number of selected genes from the top-k CARs, as shown in Fig. 2. Less informative edges between nodes were filtered out with a threshold set to 0.3. This helps to reduce the noise contributed to model training and to improve model training time due to a decrease in graphs' sparsity but without compromising the model accuracy as denoted in Eq. 5:

$$A_{ij}^{information} = \begin{cases} e_{ij}, \text{ if } e_{ij} > \text{ threshold}, \{i,j\} \subseteq R, \quad R \in CARs; \\ 0, \quad \text{otherwise}. \end{cases} \tag{5}$$

Prior knowledge graph

Apart from the synthetic information graph, prior knowledge information was also injected into the final static graph to further enhance the model performance. One of the benefits of the graph neural network is its ability to learn from non-Euclidean

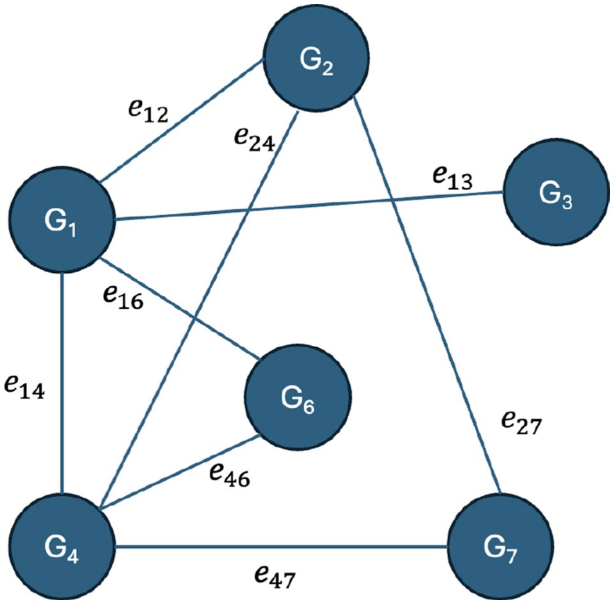
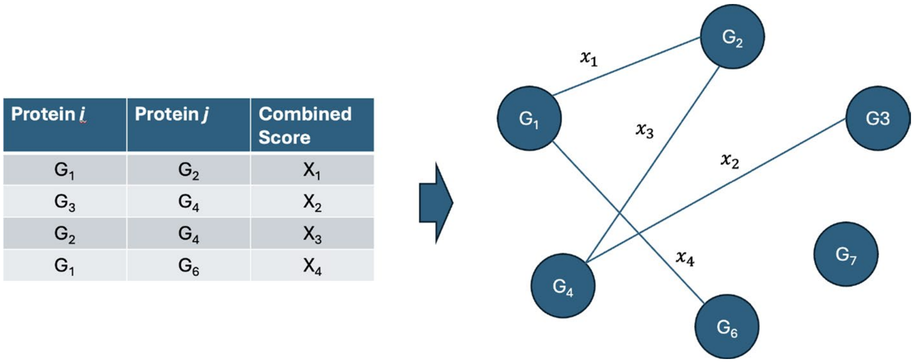


Fig. 2 Information graph. The edge between the node G_i and node G_j is connected if the information-content of both G_i and G_j is more than a certain threshold



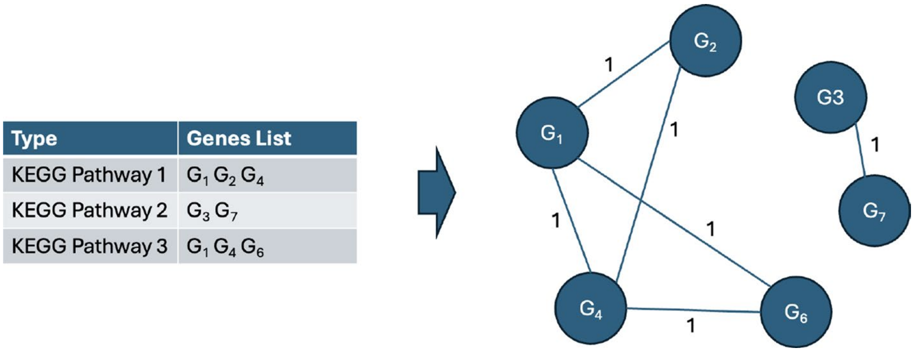
Protein *i*	Protein *j*	Combined Score
G_1	G_2	X_1
G_3	G_4	X_2
G_2	G_4	X_3
G_1	G_6	X_4

Fig. 3 PPI edges construction. The edge between the node G_i and node G_j is connected if there is a PPI pair consisting of both G_i and G_j. The attribute used for the edge is the combined score from the PPI database

relations, such as prior knowledge graph information. The PPI table obtained from the STRING database consists of two gene columns, each representing a gene involved in the interaction, along with a corresponding confidence score indicating the reliability or strength of the interaction between genes. For this study, protein-protein interaction with a confidence score equal to or more than 500 was selected to prevent weak protein gene interaction from being added to the final graph. Each set of genes was used to construct the gene network graph as shown in Fig. 3. By using the same gene set extracted from the synthetic information graph, an edge between gene i and gene j is connected if the genes exist in PPI prior knowledge and the confidence score is used as the attribute value for the edge, which can be denoted as e_{ij}^{ppi}, as shown in Eq. 6. Due to the PPI database being limited to the protein-gene network, there will be no edges between the selected miRNA nodes and miRNA-mRNA nodes from the top-k CARs. In other words, identical genes across inter-omics datasets will have an edge connecting each other if the source and target genes exist in the PPI network. The final normalized adjacency matrix for the PPI graph, $A_{ij}^{ppi} \in \mathbb{R}^{m\times m}$, can be constructed as $A_{ij}^{ppi} = normalize(e_{ij}^{ppi})$, value range from 0 to 1, m is the number of selected genes from the top-k CARs.

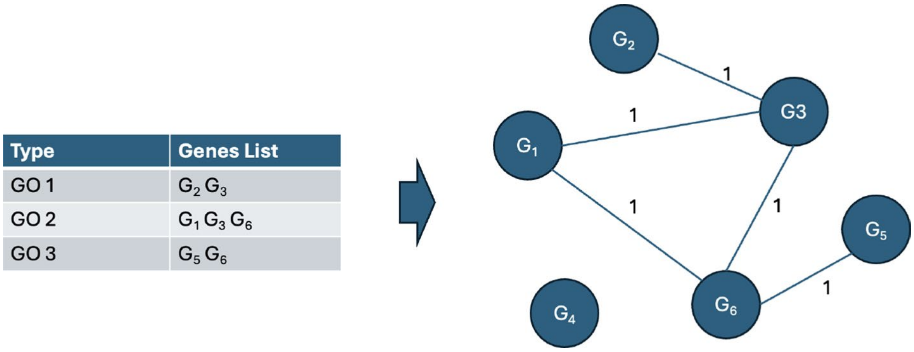
$$e_{ij}^{ppi} = \begin{cases} confidence\_score_{ij}, & \text{if confidence\_score} > 500 \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

Apart from that, two functional enriched graphs were constructed using the KEGG pathway and GO obtained from the DAVID database. It provides valuable information on the involvement of genes in specific biological pathways and their functional roles according to Gene Ontology terms. The functional annotation chart obtained from DAVID provides a comprehensive overview of the association between pathways or Gene Ontology terms and their related genes. In the KEGG pathway annotation chart, each entry includes a specific pathway term along with a list of genes associated with that pathway. For the GO annotation chart, each entry encompasses a Genome Ontology term, and the corresponding genes linked to that specific biological process, cellular component, or molecular function are detailed. Related genes for each KEGG pathway and GO term were used to construct the KEGG pathways



Type	Genes List
KEGG Pathway 1	G_1 G_2 G_4
KEGG Pathway 2	G_3 G_7
KEGG Pathway 3	G_1 G_4 G_6

Fig. 4 KEGG edges construction. The edge between the node G_i and node G_j is connected with an attribute of 1 if there is a KEGG pathway term consisting both G_i and G_j



Type	Genes List
GO 1	G_2 G_3
GO 2	G_1 G_3 G_6
GO 3	G_5 G_6

Fig. 5 GO edges construction. The edge between the node G_i and node G_j is connected with an attribute of 1 if there is a GO term consisting both G_i and G_j

graph and GO genes graph, as shown in Figs. 4 and 5. The edge between gene i and gene j, e_{ij}^{kegg} and e_{ij}^{go} defined in Eq. 7, is constructed with an attribute value of 1 if there is a KEGG pathway or GO term consist of both gene i and gene j. Similar to PPI, the KEGG pathway and GO terms with a significant value of less than 0.05 were filtered out. Adjacency matrix for KEGG pathway graph, $A_{ij}^{kegg} \in \mathbb{R}^{m \times m}$, and GO graph, $A_{ij}^{go} \in \mathbb{R}^{m \times m}$, can be constructed using $A_{ij}^{kegg} = e_{ij}^{kegg}$ and $A_{ij}^{go} = e_{ij}^{go}$.

$$\begin{aligned} e_{ij}^{kegg} &= \begin{cases} 1, \text{ if there is KEGG pathway term consist of both gene } i \text{ and } j. \\ 0, \text{ otherwise.} \end{cases} \\ e_{ij}^{go} &= \begin{cases} 1, \text{ if there is GO term consist of both gene } i \text{ and } j. \\ 0, \text{ otherwise.} \end{cases} \end{aligned} \tag{7}$$

Final static graph

With the synthetic information graph and prior knowledge graphs, a final static graph G^{final} can be constructed. Compared to other existing works that solely used prior knowledge to generate the gene networks, we proposed an information-based graph as the main contributor, while prior knowledge was used as the auxiliary contributor. For the final edge construction for G^{final}, we proposed the use of multi-dimensional edges to retain as much information from both the information graph and prior knowledge graphs. To our knowledge, this is the first time that the gene graph has been constructed using multi-dimensional edges for cancer subtype classification. The final edges adjacency matrix A^{final} can be represented as in Eq. 8:

$$A^{final} = stack[A^{information}, \lambda A^{ppi}, \lambda A^{kegg}, \lambda A^{go}], \tag{8}$$

where the λ is the mean value from $A^{information}$, $A^{final} \in \mathbb{R}^{m \times m \times 4}$. Since deep learning models tend to be influenced by higher values and learn from large weighted edges, the edge attribute of PPI, KEGG pathway and GO were scaled by λ to avoid large contributing factors from prior knowledge in the model training.

Graph neural network

Once the unified gene graphs were constructed, the graphs were parsed into the graph neural network model. Given our final input graph G^{final}, along with a set of node features $X \in \mathbb{R}^{m \times 3}$ and edge information in adjacency matrix $A^{final} \in \mathbb{R}^{m \times m \times 4}$, m is the number of selected genes, the initial node embedding for the input graph is one-hot encoded (low, medium, and high). Information is aggregated from μ's node graph neighbourhood $N(\mu)$ for each node $\mu \in V$ and hidden embedding $h_\mu^{(k+1)}$ of the μ is updated together with the current state of μ's hidden embedding $h_\mu^{(k)}$, which can be expressed using the Eq. 9:

$$h_\mu^{(k+1)} = UPDATE^{(k)}\left(h_\mu^{(k)}, AGGREGATE^{(k)}(\{h_\nu^{(k)}, \forall \nu \in N(\mu)\})\right). \tag{9}$$

To learn the neighbour embedding, graph attention convolution (GATConv) [32] was used as the technique for neural message passing. This is used to capture the varying importance of the gene's neighbor and it is also part of the gene ranking method for biomarker selection. The attention coefficients of neighbour genes were learned through the self-attention mechanism, which enables the model to focus on relevant neighbour genes for each node. The neural passing can be represented as the following Eq. 10 and Eq. 11:

$$h_\mu^{(k+1)} = \alpha_{\mu,\mu}\Theta_s h_\mu^{(k)} + \sum_{\nu}^{N} \alpha_{\mu,\nu}\Theta_t h_\nu^{(k)}, \tag{10}$$

where the attention coefficients $\alpha_{\mu,\nu}$ are computed as

$$\alpha_{\mu,\nu} = \frac{exp\left(LeakyReLU\left(\vec{a}^T[W\vec{h}_\mu \parallel W\vec{h}_\nu]\right)\right)}{\sum_{k \in N} exp\left(LeakyReLU\left(\vec{a}^T[W\vec{h}_\mu \parallel W\vec{h}_k]\right)\right)}, s \tag{11}$$

where $\cdot^T$ represents transposition and $\parallel$ is the concatenation operation.

In AMOGEL, two layers of GATConv message passing are used. For the first layer of GATConv, every node embedding contains information from 1-hop of their neighbourhood. After two layers of GATConv, every node embedding will contain 2-hop of neighbourhood information. Adding more layers of GATConv is not recommended due to the smoothing effect, as the node representations tend to become more similar or "smoothed" across the graph. Apart from that, concatenation and skip connections of each GATConv layer were used to prevent the smoothing effect and improve the model's ability to capture local and global patterns in the graph, which leads to a deeper neural network. After three layers of GATConv message passing, the gene nodes will be encoded with information of neighbouring features from 2-hops away. A final graph embedding is generated by applying global mean pooling on all the trained node embedding and subsequently is passed into a shallow multi-layers perceptron to learn the final graph embedding representation.

Global feature learning

As mentioned before, the embedding learned from graph neural networks is highly dependent on the quality of the graph structures. Instead of relying on the graph structure domain, selected m gene features from the top-k CARs will be used as input to a feedforward deep neural network model. With a deep neural network, global feature representations can be learned independently without considering any gene-gene interactions. To enhance the overall performance of the classification model, the global feature learning module was trained in parallel with the graph feature learning module, as shown in Fig. 1. This approach allows the final classification model to learn both global features and graph features from the selected genes in the top-k CARs.

Cancer subtype classification

The embedding output from graph feature learning and global feature learning denoted as $\boldsymbol{x}_{graph}$ and $\boldsymbol{x}_{global}$ were further concatenated to form lower-dimensional representation and passed through to a multi-layers perceptron to learn the final embedding for classification as shown in Fig. 1.

Gene ranking and biomarker selection

With the trained model, genes were ranked by averaging attention scores generated using the GATConv message passing for all two layers across all the samples. Based on ranking, higher-ranked genes have higher contributing factors and stronger links to neighbour genes, which contribute more to the final classification. The top 100 ranked genes were further evaluated using the DAVID database to evaluate the relevancy of the selected biomarkers. Given the learned attention coefficients $\alpha_{\mu,\nu}$ for gene μ and ν from Eq. 11, gene i ranking can be calculated using the following Eq. 12:

$$Rank(Gene_i) = \sum_{j=1}^{m} \left(\frac{\alpha_{ij}^{GATConv1} + \alpha_{ij}^{GATConv2}}{2} \right), \tag{12}$$

where $\alpha^{GATConv1}$, $\alpha^{GATConv2}$ represented attention coefficients from 2 GATConv layers, m is the number of selected genes from the top 1000 CARs.

Results and discussion

Experiment setup

The KIPAN and BRCA datasets used in this research were split into training and test datasets using a 7:3 ratio. The class distribution for each dataset was equally divided between the training and test datasets. To avoid any bias in data splitting, the experiment was repeated five times with random split sampling. The learning rate was set at 0.00005, and the experiments were run for 500 epochs. The experiments were implemented in Python version 3.9.18, using the PyTorch framework along with the torch-geometric module. The experiments were conducted on a local machine with an Intel Xeon W-2145 processor, 64GB of RAM, and Nvidia Quadro P5000 GPU. For model performance evaluation, the following metrics were used: accuracy, macro F1 score, and macro AUROC. The mean and standard deviation (STD) were recorded for each metric. The accuracy measures the overall correctness of the AMOGEL and be expressed in the following Eq. 13:

$$Acurracy = \frac{\sum_{i=1}^{C} True\,Positives}{Total\,Samples}. \tag{13}$$

where C is the total number of classes. Apart from accuracy, the macro F1 score was used as a metric to measure the performance of the model, which can be interpreted as following Eq. 14:

$$\begin{aligned} Precision_i &= \tfrac{TP_i}{TP_i+FP_i}; \\ Recall_i &= \tfrac{TP_i}{TP_i+FN_i}; \\ F1\,score_i &= \tfrac{2\times Precision_i \times Recall_i}{Precision_i+Recall_i}, \\ Macro\,F1\,Score &= \tfrac{\sum_{i=1}^{C} F1\,score_i}{C}; \end{aligned} \tag{14}$$

where TP_i is true positives for class i, FP_i is false positives for class i, FN_i is false negatives for class i and $Support_i$ is the number of sample for class i. The macro F1 score ranges from 0 to 1, and it measures the balance of precision and recall, with 1 as the perfect classification. The area under the receiver operating characteristic (AUROC) curve measures the model's ability to distinguish the samples' classes. The higher AUROC suggests a better overall ability of the model to distinguish the class of a sample from other classes.

For multi-class classification, one of the common loss functions, the categorical cross-entropy loss, was computed for the AMOGEL for backward propagation optimization, as shown in Eq. 15:

$$Categorical\,Cross\,Entryopy\,Loss = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} y_{ij} log(p_{ij}), \tag{15}$$

where C is the number of classes, y_{ij} is an indicator (0 or 1) of whether sample i belongs to class j, and p_{ij} is the predicted probability that sample i belongs to class j. adaptive moment estimation (Adam) optimizer with L2 regularisation of 0.001 was used to

Table 2 Classification performance on BRCA and KIPAN datasets

Experiment	BRCA			KIPAN		
	Acc (STD)	F1 (STD)	AUROC (STD)	Acc (STD)	F1 (STD)	AUROC (STD)
KNN	72.12(1.6)	45.56(1.1)	79.19(1.5)	93.03(1.6)	92.24(1.5)	96.80(0.9)
SVM	72.47(2.7)	44.39(4.5)	90.56(1.0)	92.89(1.9)	91.86(2.3)	98.45(0.5)
NB	81.39(3.1)	74.87(4.9)	92.50(2.5)	95.86(1.8)	94.58(2.2)	97.21(0.7)
DNN	75.58(2.7)	62.37(5.4)	90.05(1.0)	95.15(1.5)	94.12(2.2)	98.81(0.7)
MOGONET [5]	75.50(2.1)	60.96(5.9)	87.97(2.9)	95.65(1.2)	94.90(1.6)	97.38(1.8)
MOGLAM [11]	76.88(2.3)	64.97(4.1)	92.13(1.0)	94.75(1.5)	93.72(1.4)	98.64(0.6)
AMOGEL	**86.32(1.7)**	**75.67(4.4)**	**94.36(0.6)**	**96.06(1.4)**	**95.08(1.4)**	**99.37(0.4)**

Bold indicates the highest value and the bracket values are the standard deviation

Table 3 Classification performance comparison on BRCA dataset with prior knowledge GGI, PPI & Co-expression Network

Model	Accuracy (STD)
MPK-GNN [7]	66.2 (1.0)
AMOGEL	76.6 (2.4)

The accuracy result of MPK-GNN were retrieved from the experiment result with 7:3 data splitting

update the trainable weight of the model during the backward propagation optimization. This is to prevent any overfitting problem from arising during model training.

Performance comparison and analysis

The proposed method was compared against a few conventional models, including k-nearest neighbour classifier (KNN), support vector machine (SVM), and naive bayes (NB). Apart from that, we compared the proposed method with a feed-forward deep neural network model (DNN), and the state-of-the-art GNN models (MOGONET [5] & MOGLAM [11]), specifically for cancer subtype classification. For KNN, SVM, NB and DNN models, the preprocessed multi-omics datasets mRNA, miRNA and DNA methylation were concatenated horizontally before the model training and testing. For MOGONET [5] and MOGLAM [11] models, each preprocessed multi-omics dataset was trained with omics-specific classifier, followed by the fusion technique for final classification. The results are shown in Table 2.

From the result, AMOGEL outperformed other models in terms of accuracy (Acc), F1-score (F1) and AUROC for both BRCA and KIPAN datasets. The BRCA dataset recorded the highest improvement as compared to other existing methods in terms of accuracy and F1 score performance. For kidney cancer subtype classification, the lower improvements were due to the higher interpretability of its omics dataset, where differences among KICH, KIRC and KIRP can be easily differentiated by observation [5]. This also signifies that our proposed method has the ability to outperform other methods and classify much more complex dataset scenarios. Apart from that, we also performed a comparison study with the current state-of-the-art model MPK-GNN [7], integrated with multi-omics data with multiple prior knowledge. For a fair comparison, AMOGEL model performance was evaluated with the BRCA dataset and prior knowledge from the

Table 4 Summary of mRNA, DNA Methylation and miRNA features count and distribution before DNN model

Experiment	BRCA			KIPAN		
	mRNA (%)	DNA (%)	miNRA (%)	mRNA (%)	DNA (%)	miNRA (%)
DNN	1000(40%)	1000(40%)	502(20%)	2000(45%)	2000(45%)	471(10%)
FS(Dist1)_DNN	150(33%)	150(33%)	150(33%)	333(33%)	333(33%)	333(33%)
FS(Dist2)_DNN[a]	230(51%)	217(48%)	3(1%)	435(44%)	554(55%)	11(1%)
ARM(Top1000)_DNN[ab]	188(42%)	157(35%)	98(22%)	589(64%)	304(33%)	28(3%)

[a] Result was reported from one of the trial of the experiments to compare with other single value reporting experiments.

[b] It was observed that number of selected mRNA features consistently has higher distribution as compared to DNA Methylation and miRNA

Table 5 Number of selected genes and CARs result with different ARM rule pruning for BRCA dataset

ARM rule pruning	Number of selected CARs					Number of selected genes				
	T1	T2	T3	T4	T5	T1	T2	T3	T4	T5
CBA	28	20	18	18	21	56	66	56	45	65
DNN	100	2000	10	1500	100	144	533	73	497	154
Top1000	1000	1000	1000	1000	1000	437	448	402	446	379

5 experiments were carried out with random data splitting, denoted as T1,T2,T3,T4 and T5

same source as mentioned by MPK-GNN [7] research study. Prior knowledge of gene-gene interaction (GGI), PPI and Co-expression network was used, and the comparison result is shown in Table 3. The result showed that the AMOGEL has better accuracy than MPK-GNN model by 10.4%.

Ablation studies

In this study, we conducted ablation studies on the proposed method, which can be separated into three modules, namely: ARM feature selection, graph feature learning, and global feature learning, as shown in Fig. 1. To study the effectiveness of the proposed ARM technique as early fusion, we applied ANOVA-F as a typical feature selection technique to reduce the total number of omics features to be the same as the final total number of omics features using the ARM technique. The total number of omics features after applying the ARM technique is, on average, 450 for the BRCA dataset and 1000 for the KIPAN dataset. For the *FS(Dist1)_DNN* method, each omics data was equally selected using ANOVA-F feature selection, which is 150 mRNA features, 150 DNA Methylation features, and 150 miRNA features for the BRCA dataset, while 333 mRNA features, 333 DNA Methylation features, and 333 miRNA features for the KIPAN dataset. For the *FS(Dist2)_DNN* method, multi-omics data were concatenated, and features were selected using the ANOVA-F feature selection technique. The summary of the total number and distribution of each omics data for the BRCA and KIPAN datasets is shown in Table 4.

For rule pruning in ARM feature selection, we compared our proposed method, fix top-*k* rule pruning based on rule ranking (Top*k*), against the existing ARM rules pruning

Table 6 Number of selected genes and CARs result with different ARM rule pruning for KIPAN dataset

ARM rule pruning	Number of selected CARs					Number of selected genes				
	T1	T2	T3	T4	T5	T1	T2	T3	T4	T5
CBA	71	51	56	72	56	582	275	414	310	268
DNN	200	1000	20	200	200	862	1323	407	550	412
Top1000	1000	1000	1000	1000	1000	1031	920	936	815	718

5 experiments were carried out with random data splitting, denoted as T1,T2,T3,T4 and T5

Table 7 Ablation study

Experiment	BRCA			KIPAN		
	Acc. (STD)	F1 (STD)	AUROC (STD)	Acc. (STD)	F1 (STD)	AUROC (STD)
DNN[a]	75.58(2.7)	62.37(5.4)	90.05(1.0)	94.15(1.5)	94.12(2.2)	98.81(0.7)
FS(Dist1)_DNN[b]	81.56(3.1)	71.32(3.9)	95.86(1.0)	94.34(1.0)	93.02(1.9)	99.04(0.6)
FS(Dist2)_DNN[c]	77.66(2.1)	66.95(3.3)	90.51(1.5)	94.54(0.7)	93.45(2.9)	98.88(0.5)
ARM(Top1000)_DNN[d]	83.20(2.7)	71.78(4.9)	95.80(0.8)	94.95(2.2)	93.74(2.3)	99.03(0.3)
ARM(Top1000)_GNN(noPrior)[e]	83.03(1.2)	70.29(2.2)	94.26(1.3)	90.91(1.9)	90.92(2.0)	97.17(1.7)
ARM(Top1000)_GNN(allPrior)[f]	83.86(1.6)	74.94(1.3)	94.48(1.3)	95.49(1.4)	93.60(1.9)	97.18(1.3)
ARM(CBA)_GNN(allPrior)_DNN[g]	81.73(3.8)	72.86(9.5)	91.80(3.7)	95.35(0.7)	95.04(1.2)	99.05(0.4)
ARM(DNN)_GNN(allPrior)_DNN[h]	85.37(1.5)	74.39(5.5)	94.55(0.9)	95.35(1.2)	94.21(1.4)	99.14(0.4)
ARM(Top1000)_GNN(noPrior)_DNN[i]	83.29(3.5)	72.10(5.3)	95.46(1.6)	95.45(1.7)	94.32(0.9)	98.94(0.6)
ARM(Top1000)_GNN(allPrior)_DNN[j]	**86.32(1.7)**	**75.67(4.4)**	**95.96(0.6)**	**96.06(1.4)**	**95.08(1.4)**	**99.37(0.4)**

Bold indicates the highest value and the bracket values are the standard deviation

[a] Concatenated multi-omics datasets were passed into the DNN model for final classification

[b] Feature selection using ANOVA-F method for each omics data independently

The final distribution of features between omics is equally distributed. Total concatenated omics features are 450 for the BRCA dataset and 999 for the KIPAN dataset

[c] Multi-omics data were concatenated, and subsequently, the features were reduced and selected using the ANOVA-F method. The final number of selected features is 450 for the BRCA dataset and 1000 for the KIPAN dataset

[d] Multi-omics features were selected and concatenated using the ARM technique with Top-1000 ranked CARs. Selected features were passed into the DNN model for final classification

[e] Multi-omics features were selected and concatenated from Top-1000 ranked CARs, generated using ARM technique. The synthetic information graph was constructed for final classification

[f] Multi-omics features were selected and concatenated from Top-1000 ranked CARs, generated using ARM technique. Final static graphs were constructed from the information-based graph and prior knowledge graphs for graph feature learning. Graph feature learning and global feature learning were utilized for the final classification

[g] Multi-omics features were selected and concatenated from pruned CARs by the CBA method. Final static graphs were constructed from the information-based graph and prior knowledge graphs for graph feature learning. Graph feature learning and global feature learning were utilized for the final classification

[h] Multi-omics features were selected and concatenated from the best Top-K CARs based on DNN classifier performance. Final static graphs were constructed from the information-based graph and prior knowledge graph for graph feature learning. Graph feature learning and global feature learning were utilized for final classification

[i] Single dimensional edge graph for AMOGEL model, without prior knowledge information

[j] The proposed AMOGEL method

methods, class-based association (CBA) and dynamic top-*k* rule selection DNN classifier. CBA, which was originally introduced by [33], has been widely used by other researchers due to its effectiveness and efficiency. For DNN, different numbers of top-*k* rules $k = 10, 20, 30, 40, 50, 100, 200, 300, 400, 500, 1000, 1500, 2000$, which were sorted by information-based content ranking, were evaluated using the DNN classifier and the best performing top-k rules DNN classifier was selected. By comparing different classifiers for ARM rule pruning, our proposed method achieved the highest performance compared to the CBA and DNN classifiers, as shown in Table 7. This could be due to the low number of genes selected from the low number of left-over rules after CBA classifier rule pruning. Some of the important genes might be left out of the selected rules. Similar to the CBA classifier, a lower performance from the DNN classifier may be due to the instability of top-*K* selection, which results in varying gene selection in different numbers of trials. Tables 5 and 6 show a summary of the number of selected genes and CARs with different type rule pruning for ARM module for both BRCA and KIPAN datasets.

Apart from that, we conducted an ablation study on the effectiveness of prior knowledge in cancer subtype classification incorporated in graph feature learning module (GNN) and the effectiveness of each integration of DNN module and GNN module in parallel. Table 7 shows the overall ablation study results. Based on the results, the proposed ARM technique, *ARM(Top1000)_DNN*, is proven to be an effective inter-omics fusion strategy by attaining higher accuracy, F1 and AUROC scores for the BRCA and KIPAN datasets when compared to models without ARM, namely *DNN*, *FS(Dist1)_DNN* and *FS(Dist2)_DNN*. By reducing the input features same as *ARM(Top1000)_DNN*, *FS(Dist1)_DNN* and *FS(Dist2)_DNN*, their accuracy and F1 score are still unable to exceed *ARM(Top1000)_DNN* performance. This could be due to mRNA features having more valuable information than DNA methylation and miRNA. The features count, and distribution among omics before input into the DNN model is shown in Table 4, as the ARM feature selection consistently selects more mRNA features for both the BRCA and KIPAN datasets. The result also showed that by using the proposed ARM module as the inter-omics fusion strategy, the method could mine the inter-omics relationship and extract relevant omics features from multi-omics data for subsequent tasks. Besides that, when compared to experiment models with prior knowledge *ARM(Top1000)_GNN(allprior)_DNN* & *ARM(Top1000)_GNN(allprior)* and models without prior knowledge *ARM(Top1000)_GNN(noPrior)_DNN* & *ARM(Top1000)_GNN(noPrior)*, there is an improvement in overall metrics if prior knowledge is integrated in the model for

Table 8 Top 10 ranked genes for BRCA and KIPAN datasets with omics type

Ranking	BRCA		KIPAN	
	Feature	Omics Type	Feature	Omics Type
1	miR-934	miRNA	miR-126	miRNA
2	*RAET1L*	mRNA	*TSPAN5*	mRNA
3	*FOXC1*	mRNA	*VEGFA*	mRNA
4	*ESR1*	mRNA	miR-122	miRNA
5	*MIR563*	DNA Meth.	*UBE2N*	DNA Meth.
6	*FOXA1*	mRNA	*MTUS1*	mRNA
7	*C6orf97*	mRNA	*PLVAP*	mRNA
8	*AGR3*	mRNA	*PRDM16*	mRNA
9	*GATA3*	mRNA	*SLC22A23*	mRNA
10	*TBC1D9*	mRNA	*SNORB30*	DNA Meth.

Table 9 Literature search summary for top 10 selected biomarkers

Cancer type	Biomarker	Omics type	Summary
BRCA	mir934 [40]	miRNA	Promotes breast cancer metastasis by regulating PTEN and epithelial-mesenchymal transition.
BRCA	FOXC1 [34]	mRNA	Highly linked to Basal-like subtype, often overexpressed in Basal-like cancers. Regulates genes involved in cell growth, differentiation, and survival.
BRCA	ESR1 [35, 36] GATA3 [35, 36]	mRNA	Work together in Luminal subtypes to regulate gene expression in response to estrogen and drive the growth of estrogen receptor-positive tumors.
BRCA	FOXA1 [37]	mRNA	Facilitates ESR1 binding to its target genes, aiding transcriptional regulation in Luminal breast cancer cells.
BRCA	C6orf97 [38]	mRNA	Not well-characterized; located near ESR1 and may influence breast cancer susceptibility.
BRCA	AGR3 [39]	mRNA	Associated with estrogen receptor-positive breast cancer.
BRCA	TBC1D9 [41]	mRNA	An important modulator of tumorigenesis in breast cancer.
BRCA	RAET1L	mRNA	Limited publication information; highly ranked biomarkers worth investigating for potential breast cancer connections.
BRCA	MIR563	DNA Meth.	Limited publication information; highly ranked biomarkers worth investigating for potential breast cancer connections.
KIPAN	mir126 [42]	miRNA	Downregulated in kidney cancer, including KIRC. Involved in tumor progression and metastasis
KIPAN	VEGFA [43]	mRNA	Frequently overexpressed in KIRC, targeted by anti-angiogenic therapies.
KIPAN	mir122 [44]	miRNA	Often downregulated in kidney cancer, influences tumor progression by regulating metabolic pathways.
KIPAN	MTUS1 [45] PLVAP [46] PRDM16 [47] SLC22A23 [48]	mRNA	Dysregulated in kidney cancer, contributes to tumor development and progression.
KIPAN	TSPAN5	mRNA	Limited publication information; highly ranked biomarkers worth investigating for potential breast cancer connections.
KIPAN	SNORB30 UBE2N	DNA Meth.	Limited publication information; highly ranked biomarkers worth investigating for potential breast cancer connections.

Table 10 Biomarkers list associated with breast cancer & kidney cancer from DAVID GAD disease database

Dataset	Term	Genes list	*P* Value	Fold enrich.	FDR
BRCA	Breast Cancer	*BCL2, BLM, BRCA1, POLQ,* ***GATA3****, RAD51, RAD54B, ZWINT, CHST3, CENPF,* ***ESR1****, EXO1,* ***FOXA1****, mir125b2*	1.9e−5	4.0	1.3e−2
BRCA	Breast Cancer	*BCL2, BLM, BRCA1, BUB1,* ***GATA3****, NDC80, RAD51, TPX2, ASPM, CLSPN,* ***ESR1****, MCM6, PLK4*	2.9e−4	3.3	5.0e−2
KIPAN	Renal	*CD93, HNF1B, MYCT1, NEK11, FLT1, KDR, LNX1, MGP, NPY, NOTCH4, PTGER3, PRKAG2,* ***VEGFA***	1.9e−2	2.0	6.5e−2

both BRCA and KIPAN datasets. This observation confirms the importance of integrating prior knowledge in the model, as also demonstrated by other studies in the literature [7–9].

Biomarker discovery and interpretation

From the experiment, attention scores on edges were learned by 2 layers of GAT convolution and the best-performing model from five trials was selected and used to generate

Table 11 Summary of external datasets used for biomarker validation across omics types and BRCA and KIPAN cancer subtypes

Dataset	Omic type	Sample count per subtype	Tested biomarkers	*p* value
GSE1992	mRNA	LumA(34), LumB(19), HER(12), Basal(14), Normal(5)	RAET1L	3.9045e-02
GSE19783	miRNA	LumA(41), LumB(12), HER (17), Basal (15), Normal(10)	miR-934	3.1204e-05
GSE70567	DNA Methy.	LumA(41), LumB (30), HER (32), Basal(44), Normal (11)	MIR563	2.9372e-05
GSE20685	mRNA	LumA(71), LumB(52), HER(51), Basal(31), Normal(39)	GATA3	6.2580e-60
			FOXC1	8.6821e-54
			TBC1D9	1.3633e-45
			ESR1	1.0325e-03
			C6orf97	5.2774e-49
			FOXA1	6.3436e-74
			AGR3	2.9687e-59
GSE15641	mRNA	KIRP(11), KIRC(32), KICH(6)	TSPAN5	2.0067e-08
			MTUS1	1.7907e-05
			PLVAP	2.7228e-06
			VEGFA	5.5400e-07
			PRDM16	7.9022e-03
GSE48008	miRNA	KIRP(4), KIRC(5), KICH(27)	miR-122	9.8094e-01
			miR-126	3.9823e-01

the candidate gene biomarkers based on the ranking on Eq. 12. Table 8 show the Top 10 ranking result.

From literature search, gene *FOXC1* is highly linked to Basal-like subtype and often shows overexpression in Basal-like cancers [34]. It is a transcription factor that regulates the expression of genes involved in cell growth, differentiation and survival. genes *ESR1* and *GATA3* work together in breast cancer, particularly in Luminal subtypes, to regulate gene expression in response to estrogen and drive the growth of estrogen receptor-positive tumors [35, 36]. *FOXA1* gene also facilitates the binding of *ESR1* to its target genes [37], and *GATA3* works with *ERS1* to regulate the transcription of genes that drive the growth and survival of Luminal breast cancer cells. Although gene *C6orf97* is not well characterized, it is located near the *ESR1* gene, and it may be involved in breast cancer susceptibility [38]. *ARG3* gene are often associated with estrogen receptor-positive breast cancer [39] and miR-934 promotes breast cancer metastasis by regulation of *PTEN* and epithelial-mesenchymal transition [40]. *TBC1D9* is also an important modulator of tumorigenesis in breast cancer. [41].

For the KIPAN dataset, miR-126 is known to be downregulated in kidney cancer, including KIRC [42]. It is involved in tumour progression and metastasis while *VEGFA* gene is frequently overexpressed in KIRC and is a target for anti-angiogenic therapies in kidney cancer [43]. miR-122 is also often downregulated in kidney cancer and influences tumour progression by regulating metabolic pathways [44]. Genes *MTUS1* [45], *PLVAP* [46], *PRDM16* [47] and *SLC22A23* [48] are also often dysregulated in kidney cancer, contributing to various aspects of tumour development and progression based on the existing studies. From the biomarkers list, *RAET1L* and MIR563 have limited information in publication that relates to breast cancer, but these two genes are highly ranked (second and fifth), and they may be worth studying to establish a clear connection to breast cancer, similar to how genes *TSPAN5*, *UBE2N* and *SNORD30* for KIPAN datasets are related to pan-cancer disease. The summary of the literature search for the top 10 selected biomarkers was summarized in Table 9.

Ranked genes for each dataset were compared with the genetic association database (GAD) disease from the DAVID database [23] as shown in Table 10.

From the associated biomarkers list, the top 10 ranked biomarkers were highlighted in bold. The top 200 ranked genes were compared with the associated disease (breast cancer and kidney cancer) from the DAVID database, and the related biomarkers were listed with fold enrichment and false discovery rate (FDR). Fold enrichment is a measure of how frequently a particular event occurs in a test set compared to a control set, while FDR represents the expected portion of false positives among the declared significant results. Fold enrichment of 4.0 with FDR 1.3% indicates a strong enrichment (four times more frequent in the test set) and a relatively low FDR, suggesting that only 1.3% of the significant findings were expected to be false positive. This also means that this discovered biomarkers list is strong and reliable, indicating a meaningful and trustworthy association. For the second list of detected genes for the BRCA dataset, although it is slightly less than the first biomarkers list, both discovered biomarkers lists were considered significant and reliable, which also demonstrates that the AMOGEL can rank and select the genes accordingly.

To further assess the reliability of identified biomarkers, we conducted external validation using independent datasets obtained from NCBI GEO [49]. The statistical analysis, performed using ANOVA-F tests, confirmed the significance of the top 10 selected biomarkers in distinguishing between subtypes. Table 11 summarize the datasets used, the omic types, subtype sample distributions, and the tested biomarkers.

Figs. 6 and 7 shows the expression distribution of selected top 10 biomarkers across breast cancer subtypes and kidney cancer subtypes, highlighting their statistical significance based on ANOVA-F test p-values. The results indicate that all selected biomarkers exhibit statistically significant differences across subtypes, with p-values < 0.05 for BRCA-selected biomarkers. These findings further validate the robustness of our biomarker selection and provide strong evidence for their discriminative power in the breast cancer subtype. For KIPAN, the plot confirmed the statistical significance of the majority of the tested biomarkers. However, three biomarkers, SNORD30 (DNA Methy.), SLC22A23 (mRNA), and UBE2N (DNA Methy.) could not be tested due to the limited availability of independent datasets. miR-122 and miR-126 were not found to be statistically significant in the given dataset. Despite these limitations, the overall validation results support the relevance of AMOGEL framework for biomarker identification across different cancer types, particularly BRCA-selected biomarkers.

The top 10 genes' attention scores in relation to the top 200 genes for each dataset were plotted using hierarchical edge bundling, as shown in Figs. 8 and 9. Both the results demonstrated that the top 10 genes are not only significant individually but also share strong mutual interactions, which could be crucial for understanding their collective role in the biological process or condition being studied.

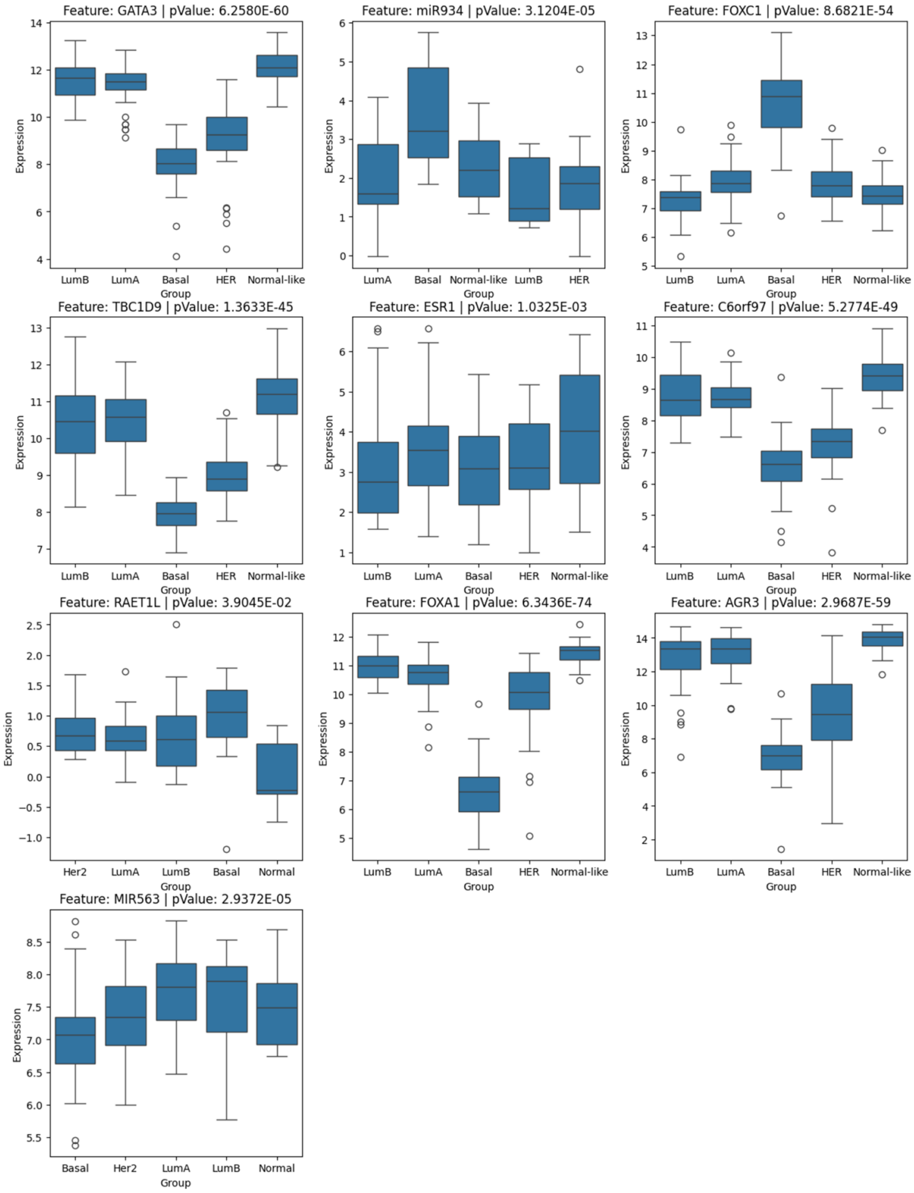


Fig. 6 Boxplots showing the expression levels of selected biomarkers across breast cancer subtypes, with corresponding *p*-values from ANOVA-F tests

Conclusion

This study introduced the Associative Multi-Omics Graph Embedding Learning (AMOGEL) model, which effectively integrates multi-omics data and prior biological knowledge through Graph Neural Networks (GNN) and association rule mining (ARM). The model employed an early fusion technique using ARM to mine inter-omics relationships, forming a comprehensive multi-omics graph before model training. By introducing multi-dimensional edges, information edges as the main contributors and prior knowledge edges as auxiliary contributors, AMOGEL enhances the representation of complex biological interactions. Additionally, the gene ranking technique, which utilizes attention scores and the relationships between neighbouring genes, further improves the model's interpretability and provides useful information on biomarkers discovery.

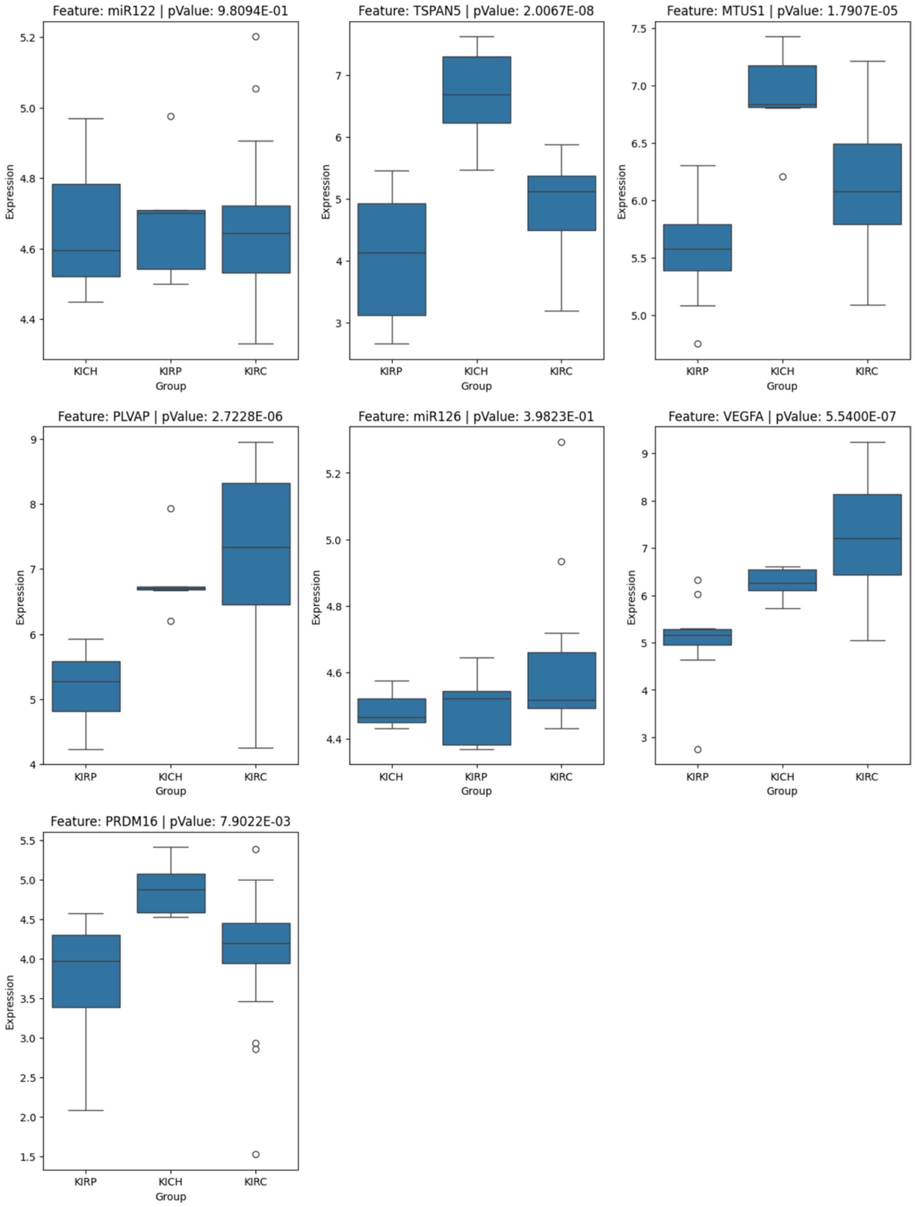


Fig. 7 Boxplots showing the expression levels of selected biomarkers across kidney cancer subtypes, with corresponding *p*-values from ANOVA-F tests

The integration of miRNA, mRNA, and DNA methylation data with prior knowledge from PPI, KEGG pathways, and Gene Ontology (GO) has shown significant improvement in cancer subtype classification accuracy, F1 score, and AUROC. By generating information-based gene-gene interactions and selecting informative genes, AMOGEL addresses the limitations of existing models that rely on limited prior knowledge. To further validate the ranked biomarkers, we compared the identified top 200 genes with the genetic association database (GAD) from DAVID functional annotation tool. The selected biomarkers exhibited strong enrichment in breast cancer and kidney cancer, with fold enrichment values of up to 4.0 and a false discovery rate (FDR) as low as 1.3%. This enrichment analysis confirms that AMOGEL can effectively rank disease-relevant

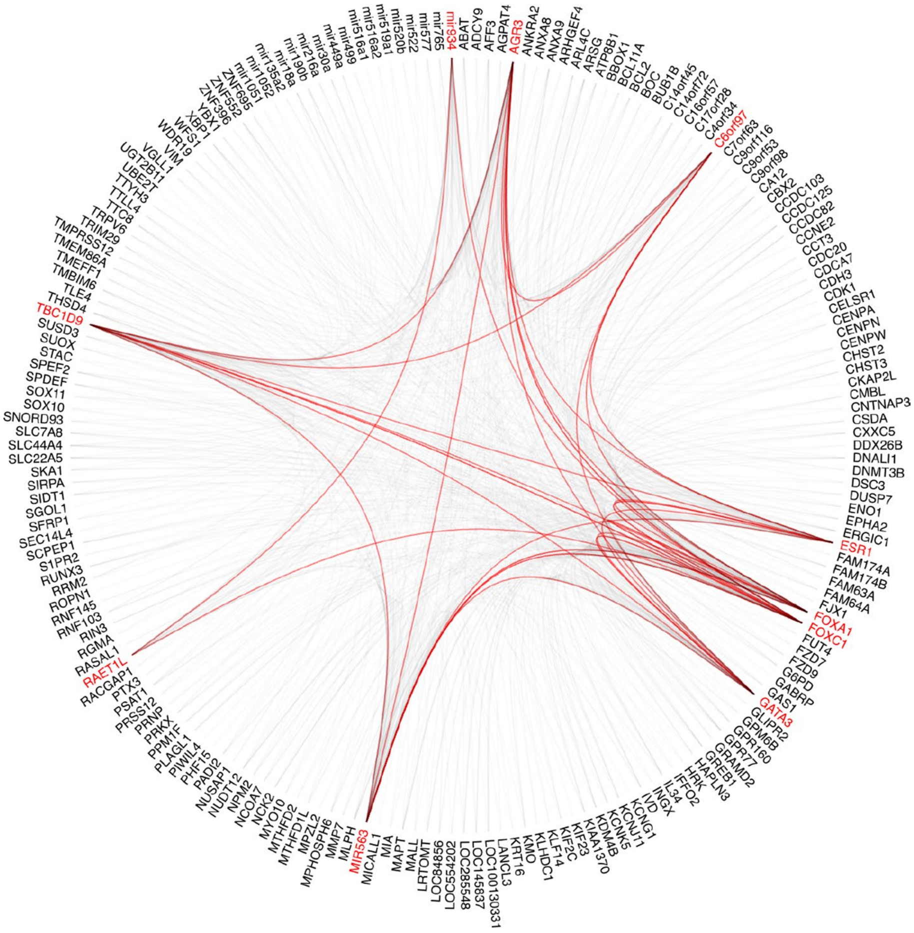


Fig. 8 BRCA top 10 genes attention score in relation to top 200 genes. Nodes around the circles represented individual genes and the edges indicated the relationship between the nodes with non-zero attention score. Edges were highlighted in red to show that there are connections (non-zero attention score) within the top 10 genes

biomarkers, further reinforcing its biological significance. Additionally, independent validation using external datasets from NCBI GEO confirmed that most selected biomarkers exhibited significant differential expression across subtypes, particularly in BRCA and KIPAN. However, some biomarkers could not be tested due to dataset limitations, and a few were found to be statistically non-significant. Despite the limitations, other promising findings suggest that AMOGEL can serve as a robust framework for enhancing cancer subtype classification and for identifying biologically meaningful biomarkers, which ultimately contribute to more effective and personalized treatment strategies.

The model's performance is currently influenced by the need for careful hyperparameter tuning, particularly the information-edge filter threshold, to prevent the creation of overly dense graph structures. Future research could focus on developing auto hyperparameter tuning methods and refining ARM algorithms to produce more efficient information-based graphs. While this study focused on the BRCA and KIPAN cancer subtypes, future research should explore the application of AMOGEL to other cancer subtypes to validate its broader applicability.

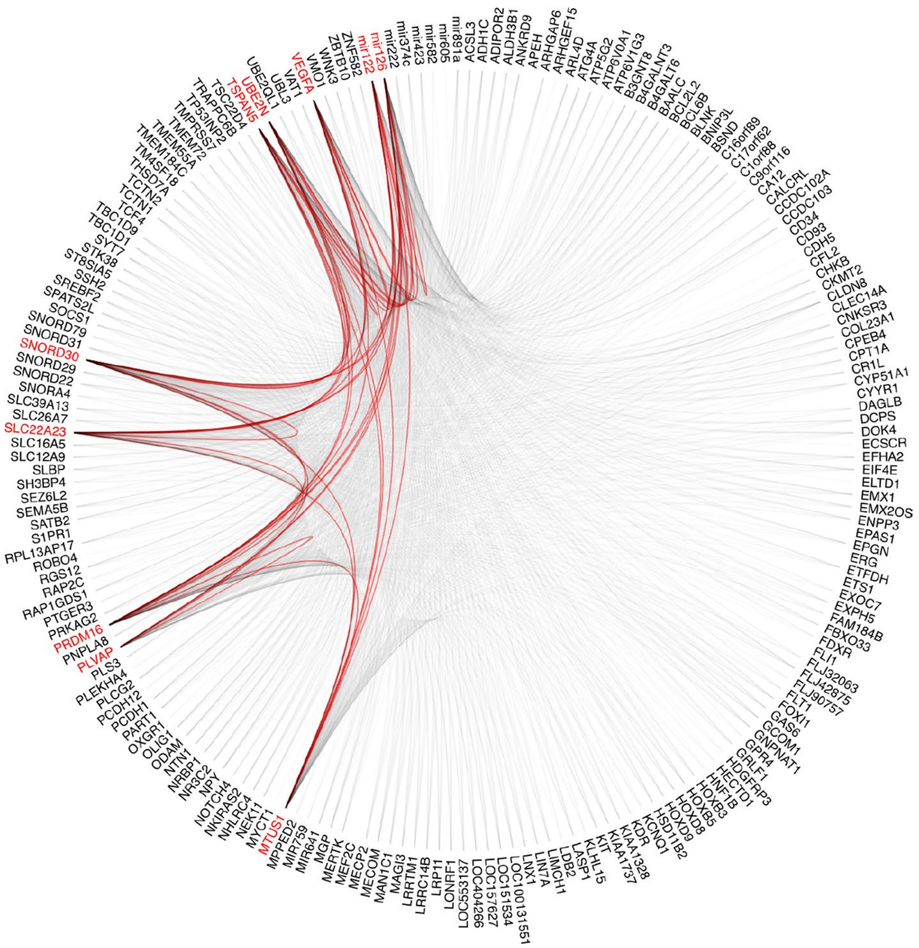


Fig. 9 KIPAN top 10 genes attention score in relation to top 200 genes. Nodes around the circles represented individual genes and the edges indicated the relationship between the nodes with non-zero attention score. Edges were highlighted in red to show that there are connections (non-zero attention score) within the top 10 genes

Author contributions

Design study and methodology, Chia Yan Tan, Huey Fang Ong, Chern Hong Lim and Mei Sze Tan; Data preparation and experiment, Chia Yan Tan; Writing-original draft, Chia Yan Tan; Writing review and editing, Huey Fang Ong, Chern Hong Lim, Mei Sze Tan, Ean Hin Ooi and KokSheik Wong.

Funding

This work is funded by the Ministry of Higher Education (MOHE), Malaysia, under the Fundamental Research Grant Scheme (FRGS) (Ref: FRGS/1/2022/ICT06/MUSM/03/1).

Availibility of data and materials

Breast cancer subtype, Breast Invasive Carcinoma (BRCA), and kidney cancer subtype, Pan-kidney (KIPAN), dataset can be downloaded from TCGAbiolink (https://bioconductor.org/packages/release/bioc/html/TCGAbiolinks.html) and Broad GDAC Firehouse (https://gdac.broadinstitute.org/). PPI dataset can be downloaded from STRING database https://string-db.org/, while KEGG pathway and GO were retrieved from Database for Annotation, Visualization and Integrated Discovery (DAVID). https://davidbioinformatics.nih.gov/. The source code for AMOGEL can be downloaded from https://github.com/tchiayan/amogel.

Declarations

Competing interests

The authors declare that they have no conflict of interest.

Received: 29 November 2024 Accepted: 10 March 2025
Published online: 28 March 2025

References

1. Ding X-L, Su Y-G, Yu L, Bai Z-L, Bai X-H, Chen X-Z, Yang X, Zhao R, He J-X, Wang Y-Y. Clinical characteristics and patient outcomes of molecular subtypes of small cell lung cancer (sclc). World J Surg Oncol. 2022;20:54.
2. Mohammed AA. The clinical behavior of different molecular subtypes of breast cancer. Cancer Treatment Res Commun. 2021;29: 100469.
3. Li J, Huang G, Ren C, Wang N, Sui S, Zhao Z, Li M. Identification of differentially expressed genes-related prognostic risk model for survival prediction in breast carcinoma patients. Aging. 2021;13:16577–99.
4. Li S, Yang Y, Wang X, Li J, Yu J, Li X, Wong K-C. Colorectal cancer subtype identification from differential gene expression levels using minimalist deep learning. BioData Mining. 2022;15(1):12. https://doi.org/10.1186/s13040-022-00295-w.
5. Wang T, Shao W, Huang Z, Tang H, Zhang J, Ding Z, Huang K. Mogonet integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification. Nat Commun. 2021;12(1):3445–3445. https://doi.org/10.1038/s41467-021-23774-w.
6. 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. https://doi.org/10.3390/ijms25052788.
7. 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;27:4591–600.
8. Li B, Nabavi S. A multimodal graph neural network framework for cancer molecular subtype classification. BMC Bioinform. 2024;25(1):27. https://doi.org/10.1186/s12859-023-05622-4.
9. Li B, Wang T, Nabavi S. Cancer molecular subtype classification by graph convolutional networks on multi-omics data. In: Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics. BCB '21. Association for Computing Machinery, New York, NY, USA. 2021. https://doi.org/10.1145/3459930.3469542 .
10. Yao X, Jiang X, Luo H, Liang H, Ye X, Wei Y, Cong S. Mocat: multi-omics integration with auxiliary classifiers enhanced autoencoder. BioData Mining. 2024;17(1):9. https://doi.org/10.1186/s13040-024-00360-6.
11. Ouyang D, Liang Y, Li L, Ai N, Lu S, Yu M, Liu X, Xie S. Integration of multi-omics data using adaptive graph learning and attention mechanism for patient classification and biomarker identification. Comput Biol Med. 2023;164: 107303. https://doi.org/10.1016/j.compbiomed.2023.107303.
12. Bellman R. Adaptive control processes: a guided tour. Princeton University Press, ??? 1961. http://www.jstor.org/stable/j.ctt183ph6v
13. Piatetsky-Shapiro G, Tamayo P. Microarray data mining: facing the challenges. SIGKDD Explor Newsl. 2003;5(2):1–5. https://doi.org/10.1145/980972.980974.
14. Ong HF, Mustapha N, Hamdan H, Rosli R, Mustapha A. Informative top-k class associative rule for cancer biomarker discovery on microarray data. Expert Syst Appl. 2020;146: 113169. https://doi.org/10.1016/j.eswa.2019.113169.
15. AbdElNabi MLR, Wajeeh Jasim M, EL-Bakry HM, Hamed N, Taha M, Khalifa NEM. Breast and colon cancer classification from gene expression profiles using data mining techniques. Symmetry. 2020;12(3):408. https://doi.org/10.3390/sym12030408.
16. Boln-Canedo V, Snchez-Maroo N, Alonso-Betanzos A. Feature selection for high-dimensional data, 1st edn. Springer, ??? (2015)
17. Sowan B, Eshtay M, Dahal K, Qattous H, Zhang L. Hybrid pso feature selection-based association classification approach for breast cancer detection. Neural Comput & Applic. 2023;35(7):5291–317. https://doi.org/10.1007/s00521-022-07950-7.
18. Hou X, Hou J, Huang G. Bi-dimensional principal gene feature selection from big gene expression data. PLoS One. 2022;17(12):0278583. https://doi.org/10.1371/journal.pone.0278583.
19. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521:436–44. https://doi.org/10.1038/nature14539.
20. Liang B, Gong H, Lu L, Xu J. Risk stratification and pathway analysis based on graph neural network and interpretable algorithm. BMC Bioinform. 2022;23(1):394. https://doi.org/10.1186/s12859-022-04950-1.
21. Szklarczyk D, Kirsch R, Koutrouli M, Nastou K, Mehryary F, Hachilif R, Gable AL, Fang T, Doncheva NT, Pyysalo S, Bork P, Jensen LJ, Mering C. The string database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. Nucleic Acids Res. 2023;51:638–46.
22. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using David bioinformatics resources. Nat Protoc. 2009;4:44–57.
23. Sherman BT, Hao M, Qiu J, Jiao X, Baseler MW, Lane HC, Imamichi T, Chang W. David: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). Nucleic Acids Res. 2022;50:216–21.
24. Kaur M, Kang S. Market basket analysis: identify the changing trends of market data using association rule mining. Procedia Comput Sci. 2016;85:78–85. https://doi.org/10.1016/j.procs.2016.05.180.
25. Sen D, Paladhi S, Frnda J, Chatterjee S, Banerjee S, Nedoma J. Associative classifier coupled with unsupervised feature reduction for dengue fever classification using gene expression data. IEEE Access. 2022;10:1–1. https://doi.org/10.1109/ACCESS.2022.3198937.
26. Alagukumar S, Lawrance R. Classification of microarray gene expression data using associative classification. In: 2016 International Conference on Computing Technologies and Intelligent Data Engineering (ICCTIDE'16). 2016; pp. 1–8. https://doi.org/10.1109/ICCTIDE.2016.7725362
27. Mallik S, Mukhopadhyay A, Maulik* U. Ranwar: Rank-based weighted association rule mining from gene expression and methylation data. IEEE Trans NanoBiosci. 2015;14(1):59–66 https://doi.org/10.1109/TNB.2014.2359494

28. Pati B, Panigrahi CR, Buyya R, Li K-C. Boolean association rule mining on microarray gene expression data. vol. 1082, pp. 99–111. Singapore: Springer Singapore Pte. Limited, Singapore (2020). https://doi.org/10.1007/978-981-15-1081-6_9
29. Li H, Sheu PC-Y. A scalable association rule learning and recommendation algorithm for large-scale microarray datasets. J Big Data-Ger. 2022;9(1):1–25. https://doi.org/10.1186/s40537-022-00577-4.
30. Borgelt C, Yang X, Nogales-Cadenas R, Carmona-Saez P, Pascual-Montano A. Finding closed frequent item sets by intersecting transactions. In: Proceedings of the 14th International Conference on Extending Database Technology. EDBT/ICDT '11, pp. 367–376. Association for Computing Machinery, New York, NY, USA (2011). https://doi.org/10.1145/1951365.1951410 .
31. Ong HF, Neoh CYM, Vijayaraj, VK., Low, YX. Information-based rule ranking for associative classification. In: 2022 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS), 2022;pp. 1–4. https://doi.org/10.1109/ISPACS57703.2022.10082812.
32. Veličković P, Cucurull G, Casanova A, Romero A, Liò P, Bengio Y. Graph attention networks. 2018. https://arxiv.org/abs/1710.10903.
33. Liu B, Hsu W, Ma Y. Integrating classification and association rule mining. In: Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining. KDD'98, pp. 80–86. AAAI Press, New York, NY 1998.
34. Jin Y, Han B, Chen J, Wiedemeyer R, Orsulic S, Bose S, Zhang X, Karlan BY, Giuliano AE, Cui Y, Cui X. Foxc1 is a critical mediator of egfr function in human basal-like breast cancer. Ann Surg Oncol. 2014;21(Suppl 4):758–66.
35. Lei JT, Gou X, Seker S, Ellis MJ. Esr1 alterations and metastasis in estrogen receptor positive breast cancer. J Cancer Metastasis Treatment. 2019;5:38.
36. Chou J, Provot S, Werb Z. Gata3 in development and cancer differentiation: cells gata have it! J Cell Physiol. 2010;222:42–9.
37. Zhang G, Zhao Y, Liu Y, Kao L-P, Wang X, Skerry B, Li Z. Foxa1 defines cancer cell specificity. Sci Adv. 2016;2:1501473.
38. Yamamoto-Ibusuki M, Yamamoto Y, Fujiwara S, Sueta A, Yamamoto S, Hayashi M, Tomiguchi M, Takeshita T, Iwase H. C6orf97-esr1 breast cancer susceptibility locus: influence on progression and survival in breast cancer patients. Eur J Human Genet: EJHG. 2015;23:949–56.
39. Xu Q, Shao Y, Zhang J, Zhang H, Zhao Y, Liu X, Guo Z, Chong W, Gu F, Ma Y. Anterior gradient 3 promotes breast cancer development and chemotherapy response. Cancer Res Treat. 2020;52:218–45.
40. Lu Y, Hu X, Yang X. mir-934 promotes breast cancer metastasis by regulation of pten and epithelial-mesenchymal transition. Tissue & cell. 2021;71: 101581.
41. Kothari C, Clemenceau A, Ouellette G, Ennour-Idrissi K, Michaud A, C-Gaudreault R, Diorio C, Durocher F. Tbc1d9: an important modulator of tumorigenesis in breast cancer. Cancers. 2021;13:3557.
42. Jalil AT, Abdulhadi MA, Al-Ameer LR, Abbas HA, Merza MS, Zabibah RS, Fadhil AA. The emerging role of microrna-126 as a potential therapeutic target in cancer: a comprehensive review. Pathol-Res Practice. 2023;248: 154631. https://doi.org/10.1016/j.prp.2023.154631.
43. Guillaume Z, Auvray M, Vano Y, Oudard S, Helley D, Mauge L. Renal carcinoma and angiogenesis: therapeutic target and biomarkers of response in current therapies. Cancers. 2022;14:6167.
44. Faramin Lashkarian M, Hashemipour N, Niaraki N, Soghala S, Moradi A, Sarhangi S, Hatami M, Aghaei-Zarch F, Khosravifar M, Mohammadzadeh A, Najafi S, Majidpoor J, Farnia P, Aghaei-Zarch SM. Microrna-122 in human cancers: from mechanistic to clinical perspectives. Cancer Cell Int. 2023;23:29.
45. Sim J, Wi YC, Park HY, Park SY, Yoon YE, Bang S, Kim Y, Jang K, Paik SS, Shin S-J. Clinicopathological significance of mtus1 expression in patients with renal cell carcinoma. Anticancer Res. 2020;40:2961–7.
46. Xu Y, Miller CP, Xue J, Zheng Y, Warren EH, Tykodi SS, Akilesh S. Single cell atlas of kidney cancer endothelial cells reveals distinct expression profiles and phenotypes. BJC Rep. 2024;2(1):23. https://doi.org/10.1038/s44276-024-00047-9.
47. Kundu A, Kho E-Y, Shelar SB, Nam H, Brinkley G, Darshan S, Tang Y, Kirkman R, Crossman DK, Varambally S, Rowe GC, Wei S, Buckhaults P, Sudarshan S. Abstract 4483: functional implications of prdm16 loss in kidney cancer. Cancer Res. 2018;78:4483–4483. https://doi.org/10.1158/1538-7445.AM2018-4483.
48. Whisenant TC, Nigam SK. Organic anion transporters (oat) and other slc22 transporters in progression of renal cell carcinoma. Cancers. 2022;14:4772.
49. Edgar R, Domrachev M, Lash AE. Gene expression omnibus: Ncbi gene expression and hybridization array data repository. Nucleic Acids Res. 2002;30:207–10.

Publisher's Note

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