LAKEHEAD UNIVERSITY

Integrating Multi-omics Data via Latent Space Construction for Breast and Bladder Cancer Analysis

by

Arvind Chidambaram Boominathan

A thesis submitted in partial fulfillment for the degree of Master of Science

in the

Faculty of Science and Environmental Studies
Department of Computer Science

April 2025

Declaration of Authorship

I, ARVIND CHIDAMBARAM BOOMINATHAN, declare that this thesis titled, 'Integrating Multi-omics Data via Latent Space Construction for Breast and Bladder Cancer Analysis' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a Master's degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: _____

Date: _____

"I love deadlines. I like the whooshing sound they make as they fly by."

–Douglas Adams

LAKEHEAD UNIVERSITY

Abstract

Faculty of Science and Environmental Studies
Department of Computer Science

Master of Science

by Arvind Chidambaram Boominathan

Cancer remains one of the most complex and heterogeneous diseases, driven by intricate interactions across genetic, epigenetic, and transcriptional landscapes. Accurately understanding and predicting tumor characteristics, such as Tumor Mutational Burden (TMB), is critical for effective diagnosis, prognosis, and personalized treatment strategies. This research aims to address inherent challenges in integrating high-dimensional, heterogeneous multi-omics datasets—including DNA methylation, gene expression, and Copy Number Alteration (CNA)—specifically for bladder and breast cancer analysis, by building a shared latent space that captures and preserves meaningful cross-omics representations. Some of these challenges include data imbalance, dimensionality, modality-specific noise, and complex non-linear biological interactions.

To overcome these obstacles, this thesis proposes constructing a shared latent space through advanced deep-learning approaches by utilizing Deep Multiset Canonical Correlation Analysis (DMCCA) and Graph Attention Networks (GATs). The shared latent space methodology provides a unified representation capturing crucial and intricate biological interactions across various omics modalities, as a result giving improved predictive accuracy for TMB classification. Attention mechanisms further refine this integration by dynamically focusing on the most relevant relational patterns within multiomics data, enhancing the model's ability to capture biological interactions between genes, pathways, and patient profiles. In addition, this study utilizes oversampling techniques—mainly the Synthetic Minority Oversampling Technique (SMOTE)—to offset data imbalance among TMB classes and menopausal status groups. As compared to baseline supervised machine learning models such as Logistic Regression (LR), Artificial Neural Network (ANN), and Tabular Transformer, the new GAT model with shared latent space training performed better by achieving an AUC of 0.76 and accuracy of 76.1% for BRCA, whereas that of BLCA was 0.73 with an accuracy of 65.3%, thereby establishing the usefulness of multi-omics integration through shared latent space learning.

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Dr. Abedalrhman Alkhateeb, for his invaluable guidance, encouragement, and support throughout my research journey. Their insightful advice and expertise have played a crucial role in shaping this work.

I am also deeply grateful to my thesis committee members, Dr. Saad Bin Ahmed and Dr. Abdulsalam Yassine, for their valuable feedback and constructive criticism, which helped refine my research.

A special thanks goes out to my colleagues and friends at Lakehead University, whose discussions and support made this journey more enriching. I am thankful to everyone who provided technical assistance and helpful insights during critical stages of this work.

I would like to extend my heartfelt appreciation to my family for their unconditional love, patience, and encouragement. Their unwavering belief in my abilities has been my greatest motivation.

Finally, I acknowledge the support and facilities provided by Lakehead University, which made this research possible.

Contents

List of Figures

List of Tables

Abbreviations

ANN	**A**rtificial **N**eural **N**etwork
AUC	**A**rea **U**nder the **C**urve
BLCA	**BL**adder Urothelial **CA**rcinoma
BRCA	**BR**east **CA**ncer Genes
CNA	**C**opy **N**umber **A**lteration
DMCCA	**D**eep **M**ultiset **C**anonical **C**orrelation **A**nalysis
ELU	**E**xponential **L**inear **U**nit
GAT	**G**raph **A**ttention **N**etwork
GCN	**G**raph **C**onvolutional **N**etwork
GE	**G**ene **E**xpression
KEGG	**K**yoto **E**ncyclopedia of **G**enes and **G**enomes
LR	**L**ogistic **R**egression
mDNA	**D**NA **M**ethylation
ML	**M**achine **L**earning
NMI	**N**ormalized **M**utual **I**nformation
ROC	**R**eceiver **O**perating **C**haracteristic
SMOTE	**S**ynthetic **M**inority **O**versampling **TE**chnique
TMB	**T**umor **M**utational **B**urden
UMAP	**U**niform **M**anifold **A**pproximation and **P**rojection

I dedicate this work to my beloved parents, whose unwavering support and sacrifices have shaped my journey. As an international student, this experience has been one of resilience, and I am deeply grateful to my supervisor, Dr. Abedalrhman Alkhateeb, for his invaluable guidance. This work stands as a testament to the collective wisdom, kindness, and belief that have guided me.

Chapter 1

Introduction

This chapter introduces the problem and the concepts involved in the developing the thesis work.

1.1 Multi-omics

Multi-omics is an integrative approach that brings together heterogeneous biological information in multiple "omics" layers such as genomics, transcriptomics, proteomics, epigenomics, and metabolomics. Multi-omics enables the simultaneous examination of different biological dimensions to achieve an overall understanding of complex biological systems and disease, which dominates the information derived from single-omics. Mult-omics integration is while promising and of biological significance, it is accompanied by challenges such as high-dimensional data, heterogeneity of modality, and computational complexity. To address these issues with advanced computational techniques is crucial for identifying accurate biomarkers, disease subtypes, and therapeutic targets, particularly in cancer biology.[1–3]

Emerging advances in high-throughput technologies [4, 5] such as next-generation sequencing (NGS), mass spectrometry, and microarrays, have created an exponential growth of omics data across various platforms and biological levels. These technologies have allowed scientists to obtain genomic mutations, transcriptomic dynamics, epigenetic alterations, and metabolic fluctuations under a single experimental framework. Thus, the integration of these multi-faceted data types has become ever more crucial for systems-level understanding and precision medicine programs. .

In this research, the multi-omics data [6] employed are DNA methylation (epigenomics), gene expression (transcriptomics), and copy number variations (genomics), which were

acquired from The Cancer Genome Atlas (TCGA) for breast (BRCA) and bladder (BLCA) cancer cohorts.

1.1.1 Copy Number Alteration (CNA)

CNA refers to structural changes in the genome resulting in gains or losses of DNA segments. CNAs can significantly affect gene expression and function, playing crucial roles in cancer initiation, progression, and resistance to therapies [7].

1.1.2 mRNA Expression

The transcriptome, analyzed through mRNA expression profiling, provides insights into gene regulatory networks and the functional state of cells. Altered mRNA expression patterns are directly associated with tumor progression, metastatic potential, and response to treatments in Breast Cancer [8].

1.1.3 DNA Methylation

DNA methylation, an essential epigenetic modification, regulates gene expression without altering the DNA sequence itself. Abnormal DNA methylation patterns contribute to tumorigenesis by silencing tumor suppressor genes or activating oncogenes i.e. tumor transforming genes, thus affecting cellular proliferation, apoptosis (death of the cell), and differentiation pathways [9]

1.1.4 Gene Expression

Gene expression profiles help in understanding molecular subtypes of cancers, which is vital for predicting patient prognosis and determining appropriate therapeutic strategies. Precise cancer study now relies heavily on integrated analysis of gene expression data [10].

1.2 Significance of Multi-omics

Cancer is a complex and heterogeneous disease influenced by genetic, epigenetic, and environmental factors. Single-omics approaches might not be able to fully capture this complexity due to the absence of omic interactions, thus highlighting the need for multi-omics integration. By analyzing the interactions and correlations across multiple omic

layers, researchers can identify robust molecular drivers of cancer, facilitating more accurate and comprehensive models of tumor biology [11]. Multi-omics integration presents several potential benefits:

1.2.1 Improved Diagnosis

By capturing diverse molecular signatures across multiple omics, integrated approaches enhance diagnostic precision, identifying distinct molecular subtypes within cancers and enabling earlier and more accurate disease detection [12].

1.2.2 Enhanced Prognosis Prediction

Multi-omics integration enables the identification of stable biomarkers that not only are of biological relevance but also are strongly correlated with clinical endpoints such as patient survival, recurrence risk, and response to treatment—ultimately allowing for more precise risk stratification and individualized estimation of prognosis [11]. For instance, Chaudhary et al. [13] demonstrated that integrating transcriptomic, epigenomic, and genomic features using deep learning significantly improved prediction of patient survival in liver cancer, which is of clinical relevance for multi-omics-based biomarker discovery.

1.2.3 Tailored Treatment Strategies

Understanding multi-layered molecular changes supports the identification of novel therapeutic targets and improves predictions of patient responses to specific treatments, thereby enabling precision oncology approaches tailored to individual molecular profiles [11, 12]. Such integration strategies enable personalized treatment by identifying omics-specific biomarkers that correlate with therapeutic outcomes, such as drug sensitivity and resistance mechanisms, thus informing targeted interventions for individual patients.

Thus, the utilization of multi-omics analysis is indispensable for advancing personalized medicine, significantly improving patient outcomes through refined diagnostic, prognostic, and therapeutic approaches.

1.3 Challenges in Multi-omics data

Integrating multi-omics data for cancer analysis is inherently challenging due to both technical and biological issues, which include:

1.3.1 High Dimensionality

Each omics dataset (genomic mutations, copy number, gene expression, methylation, etc.) can contain thousands of features, while the number of samples (patients) is relatively small. This "large p, small n" situation makes models prone to overfitting and increases the computational difficulty [14].

1.3.2 Data Heterogeneity

Different omics data have heterogeneous formats and distributions – for example, gene expression values are continuous, mutation data are binary or categorical, etc. The enormous volume and different variety of features complicate the integration, as high dimensionality and heterogeneity pose incredible challenges for multi-omics analysis [14].

1.3.3 Complexity in Integration

Integrating multi-omics means combining data measured on different scales (e.g. gene expression in RPKM (Reads Per Kilobase of transcript per Million mapped reads), methylation as beta values, etc.). There is a challenge in normalizing and scaling these diverse features so that no data type unfairly dominates due to scale alone. The lack of standardized pipelines for multi-omics integration is noted as a barrier [15].

1.3.4 Sample Imbalance and Bias

Class imbalance is a problem in the majority of cancer studies – e.g., much larger numbers of "tumor" samples than "normal" controls, or one subtype much more frequent than another. Due to this class imbalance, the classifiers tend to be biased towards the majority class due to it training on more samples of the majority class [16]. Integrative models may make well-classified predictions on the majority class but poorly on minorities (e.g., a classifier can predict everything as the most frequent subtype). These problems are usually fixed by undersampling the majority class or oversampling the minority class by various methods [17].

1.3.5 Interpretability (Biological Relevance)

Even if a multi-omics integration model achieves high predictive accuracy, understanding the reason behind the accuracy is difficult. The complexity of multi-omics models

(especially deep learning or graph-based models) makes them hard to interpret biologically [18]. This is true for latent spaces, where features are high-level and disentangled from the initial omics domains, and it's hard to trace back specific biological meaning to particular genes or pathways.

1.3.6 Data Noise

Biological datasets often contain experimental noise and batch effects. Multi-omics integration amplifies this issue because each data type may have its own noise profile. Data noise can obscure true signals, and when combining modalities, an outlier or noisy measurement in one omic could mislead the integrated analysis [19].

In summary, multi-omics data integration in cancer research faces numerous challenges: extremely high-dimensional, heterogeneous data; technical and biological noise; class or feature imbalances; and difficulty in interpretation. Researchers have been actively developing methods to handle these issues including but not limited to using advanced machine learning that can cope with missing data [20] or constructing a latent space to reduce dimensionality [21]. Future solutions may involve more sophisticated algorithms, but also improvements in experimental design and in computational strategies for scalability and interpretability [22].

1.4 Problem Statement

Cancer is a disease with layered nonlinear interactions across various omics layers such as genomics, epigenomics, and transcriptomics. Solution of these interdependencies requires the integration of high-dimensional, heterogeneous multi-omics data. Traditional approaches are incapable of handling the problems of modality imbalance, sparsity, and nonlinear relations between features. Linear integration methods are not suitable to handle cross-modal interactions, while deep learning models are frequently non-interpretable and inadequately validated on clinically pertinent outcomes.

In order to address these challenges, this thesis uses Deep Multiset Canonical Correlation Analysis (DMCCA) to construct a shared latent space that captures the common structure among omics modalities while preserving biological heterogeneity. The latent space is then conditioned with Graph Attention Networks (GATs), where patient-specific interactions can be captured in a graph structure such that the model can focus on the most relevant relationships.

This integrated paradigm is utilized on two clinically relevant classification challenges: (1) pre- vs. post-menopausal status prediction in breast cancer (BRCA), and (2) Tumor Mutational Burden (TMB) prediction in bladder cancer (BLCA). The two tasks are prominent areas in precision oncology and require strong, interpretable modeling methods. The DMCCA+GAT pipeline introduced here strives towards better predictability along with discovering biologically significant patterns in the shared latent space—virtually advancing the field of precision medicine with multi-omics integration.

Chapter 2

Preliminaries and Background

In this chapter, we introduce the following preliminaries that helps in understanding the outcomes of the proposed methodologies.

2.1 Cancer

Cancer is a complex and heterogeneous disease characterized by uncontrolled cell proliferation resulting from genetic mutations and epigenetic alterations. These mutations disrupt normal cellular processes, leading to tumor formation and progression. Cancer is influenced by genetic factors, environmental exposures, lifestyle behaviors, and interactions among multiple biological pathways, highlighting the necessity for integrated analyses across genomic, transcriptomic, epigenomic, and proteomic layers (multi-omics) to understand the disease comprehensively [23]. Effective integration of multi-omics data can significantly enhance cancer diagnosis, prognosis, and treatment strategies by identifying novel biomarkers and therapeutic targets [2, 24]

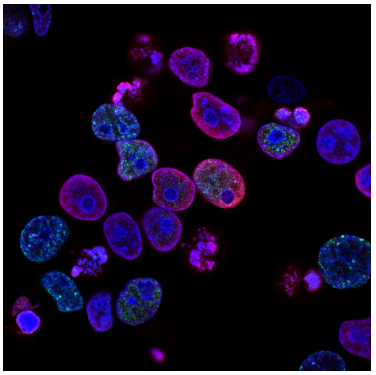


FIGURE 2.1: Cancer Cells

Figure 2.1 shows a representative structure of cancer cells.

2.1.1 Breast Cancer

Breast cancer is a malignancy that originates from breast tissue, primarily from the ducts or lobules. It represents the most commonly diagnosed cancer among women worldwide and is a leading cause of cancer-related death among females. Risk factors include genetic predisposition (BRCA1 and BRCA2 gene mutations), hormonal exposure, obesity, alcohol consumption, and reproductive history [25].

Treatment strategies depend on cancer subtype and stage, typically including, but not limited to:

- **Surgery (lumpectomy or mastectomy)** — Surgical removal of the tumor or entire breast tissue, typically used as the first-line treatment in early-stage breast cancer [26].
- **Chemotherapy** — The use of cytotoxic drugs to destroy rapidly dividing cancer cells, often administered before (neoadjuvant) or after (adjuvant) surgery.
- **Radiation therapy** — High-energy radiation is used to eliminate residual cancer cells [27] and reduce the risk of recurrence, particularly after surgery.
- **Hormonal therapy (for hormone receptor-positive cancers)** — Involves drugs like tamoxifen [28] or aromatase inhibitors to block estrogen receptors or reduce estrogen production, slowing hormone-driven tumor growth.
- **Targeted therapies (e.g., HER2 inhibitors)** — Therapies such as trastuzumab selectively target cancer-specific molecules [29] like HER2 to inhibit tumor growth and improve survival outcomes.

2.1.2 Bladder Cancer

Bladder cancer occurs in the tissues of the bladder, primarily as urothelial carcinoma. It is a significant cause of symptoms globally and is highly associated with environmental factors such as tobacco smoking, occupational chemical exposures, and chronic bladder inflammation [30].

Treatment depends on the stage and grade of cancer, including:

- Surgical interventions such as transurethral resection of bladder tumor (TURBT) or radical cystectomy.
- Intravesical chemotherapy or immunotherapy (e.g., Bacillus Calmette-Guérin [BCG]) for non-muscle-invasive bladder cancer.

- Systemic chemotherapy.
- Immunotherapy with checkpoint inhibitors for advanced cases.
- Radiation therapy as part of multimodal treatment [31].

2.2 Menopause Stages

Menopause marks the end of a woman's reproductive years and is biologically defined by the occurrence of menstrual cycles for 12 consecutive months in the year. Women are classified as premenopausal if they still have regular menstrual cycles, and postmenopausal once they transition beyond this phase. This physiological distinction is not only relevant to their reproductive health status, but also plays a significant role in the biology and management of breast cancer [32].

Menopausal status has been shown to influence molecular subtypes derived from transcriptomic data, and is frequently used as an important variable in breast cancer studies and clinical trials [33]. Incorporating menopausal status into multi-omics breast cancer analyses—such as those used in this research—enhances the understanding of tumor heterogeneity and can reveal subtype-specific biomarkers or therapeutic targets.

2.2.1 Pre-Menopausal Response to Cancer

Premenopausal women typically have higher circulating levels of estrogen and progesterone, which influence breast tissue and can drive hormone receptor-positive tumor development [32]. Breast cancer occurring in premenopausal women is often biologically distinct and clinically more aggressive. These tumors are more likely to be triple-negative (i.e., negative for estrogen receptor [ER], progesterone receptor [PR], and HER2), a subtype associated with poor prognosis and limited targeted therapy options [34].

Additionally, premenopausal patients frequently present with higher tumor grade and increased lymph node involvement at diagnosis.

2.2.2 Post-Menopausal Response to Cancer

Postmenopausal women exhibit altered hormonal profiles with reduced estrogen production, leading to different tumor behaviors and responses to therapy [32]. Postmenopausal breast cancer is more likely to be hormone receptor-positive, which makes it amenable

TABLE 2.1: Comparison of Breast Cancer Diagnosis: Premenopausal vs. Postmenopausal Women

Category	Premenopausal Women	Postmenopausal Women
Hormonal Environment	High estrogen and progesterone levels	Low estrogen due to ovarian inactivity
Common Tumor Subtypes	Triple-negative, HER2-positive [35]	Hormone receptor-positive (ER+, PR+)
Tumor Behavior	More aggressive, higher grade	Often less aggressive, slower-growing
Age of Onset	Typically diagnosed under age 50	Typically diagnosed after age 50 [32]
Diagnostic Challenges	Dense breast tissue reduces mammogram sensitivity	Less dense tissue improves mammogram effectiveness
Response to Therapy	Less responsive to hormone therapy	Better response to hormone therapies (e.g., tamoxifen, aromatase inhibitors)
Prognosis	Generally poorer due to aggressive subtypes and late detection	Generally better, especially for hormone receptor-positive tumors [34]
Screening Approach	Often requires adjunct imaging (e.g., ultrasound, MRI)	Mammography generally sufficient for early detection

to hormonal therapies such as aromatase inhibitors or selective estrogen receptor modulators like tamoxifen [35].

Table 2.1 shows the key differences between pre-menopausal and post-menopausal response and analysis of breast cancer.

2.3 Tumor Mutational Burden

Tumor Mutational Burden (TMB) is a quantitative biomarker that reflects the total number of somatic, coding base substitutions and short insertions/deletions (indels) per megabase (Mb) of the genome in a tumor. It serves as an indicator of tumor neoantigen load and the potential for immune system recognition. In bladder cancer, which is characterized by a high degree of genomic instability, TMB has emerged as a critical biomarker for both prognosis and response to immunotherapy [36].

In this thesis, TMB status is used to stratify bladder cancer samples into low and high TMB groups, allowing exploration of differential molecular features across multi-omics layers (CNA, mRNA, DNA methylation).

2.3.1 High TMB

High TMB is defined by an elevated number of somatic mutations per megabase of tumor DNA. In bladder cancer, tumors with high TMB exhibit substantial genomic instability and are associated with a greater number of neoantigens—novel peptides formed from mutated proteins that the immune system can recognize as foreign. This heightened immunogenicity contributes to the increased infiltration of cytotoxic CD8+ T cells and enhanced immune activity within the tumor microenvironment [37].

From a therapeutic perspective, high TMB is a strong predictive biomarker for response to immune checkpoint inhibitors, such as anti-PD-1 and anti-PD-L1 therapies [36]. Clinical studies, including those involving atezolizumab and nivolumab, have demonstrated that bladder cancer patients with high TMB show improved overall survival and higher objective response rates when treated with immunotherapy [38].

In this research, high TMB samples are analyzed to identify multi-omics biomarkers that correlate with immune activation, tumor progression, and treatment responsiveness.

2.3.2 Low TMB

Low TMB tumors are characterized by a relatively small number of somatic mutations, resulting in fewer neoantigens and consequently lower immune recognition. These tumors often display an immune-cold phenotype, marked by reduced infiltration of T cells, limited antigen presentation, and a more suppressive tumor microenvironment. As a result, patients with low TMB bladder cancer are less likely to benefit from immune checkpoint blockade [39].

The integration of low TMB data in this study allows the identification of epigenetic or transcriptional signatures that may contribute to immune exclusion and tumor proliferation. Exploring these molecular differences supports the development of alternative therapeutic strategies for low TMB patients who may not benefit from current immunotherapy options.

Table 2.2 shows the key clinical and molecular distinctions between high and low TMB in bladder cancer. This comparison highlights the importance of TMB as a biomarker for patient stratification and personalized treatment planning of bladder cancer.

TABLE 2.2: Comparison of High and Low TMB in Bladder Cancer

Characteristic	**High TMB**	**Low TMB**
Mutation Load	High number of somatic mutations per megabase	Low number of somatic mutations
Neoantigen Load	Increased neoantigen formation	Fewer neoantigens
Immune Activity	Enhanced immune infiltration ($CD8^+$ T cells)	Immune-cold or suppressed tumor microenvironment [37]
Therapy Response	Good response to immunotherapy (e.g., PD-1/PD-L1 inhibitors)	Limited response to immunotherapy [36]
Molecular Subtypes	Often seen in basal/squamous or immune-active subtypes	Common in luminal-papillary or urothelial-like subtypes [40]
Prognosis	Better prognosis with immunotherapy [39]	Requires alternative strategies (e.g., chemo, targeted therapy)

2.4 Machine Learning

Machine learning (ML) is a subfield of artificial intelligence that enables computational models to learn patterns from data and make predictions or decisions without being explicitly programmed. It involves various techniques such as supervised learning (e.g., classification, regression), unsupervised learning (e.g., clustering, dimensionality reduction), and deep learning, which employs neural networks for modeling complex, nonlinear relationships. ML methods are widely utilized in bioinformatics, especially for analyzing high-dimensional multi-omics data, assisting in biomarker discovery, disease diagnosis, prognosis, and personalized treatment strategies [24, 41, 42].

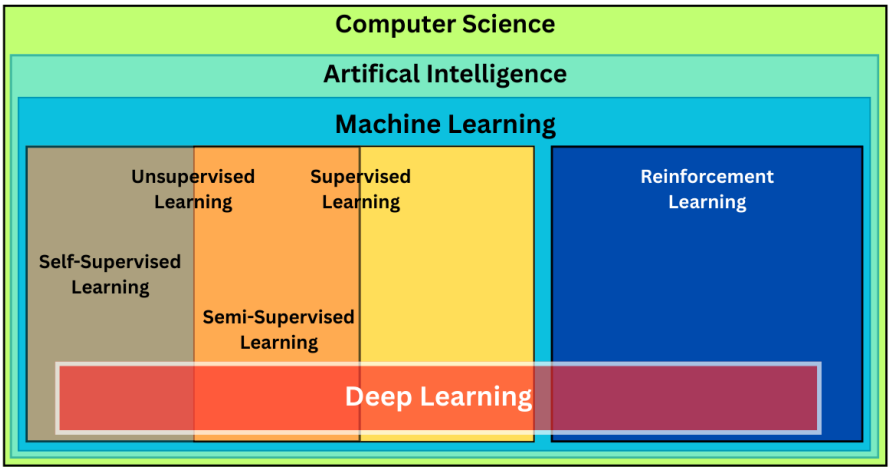


FIGURE 2.2: Machine Learning

Figure 2.2 shows the various hierarchical relationships among Computer Science, Artifical Intelligence and Machine Learning.

2.5 Graph Convolution Networks

Graph Convolutional Networks (GCNs) extend traditional convolutional neural networks (CNNs) to graph-structured data, capturing relationships and dependencies among entities through node embeddings. By effectively aggregating local node information, GCNs learn robust representations essential for tasks like node classification, link prediction, and clustering. In bioinformatics, GCNs have shown great promise in modeling biological networks, enabling the integration of multi-omics data to predict disease outcomes and discover novel biomarkers and therapeutic targets [43, 44].

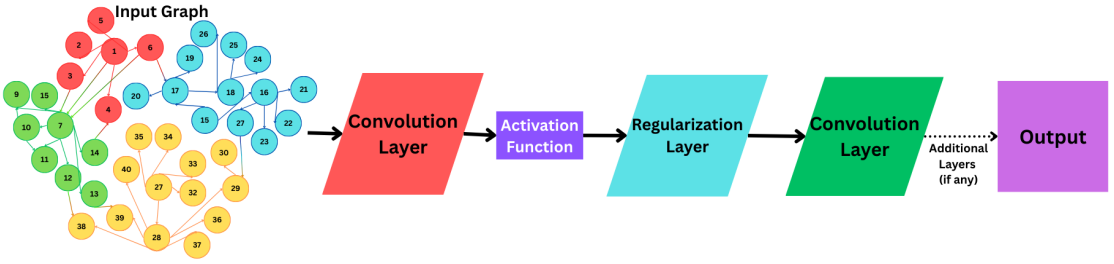


FIGURE 2.3: Graph Convolution Network

Figure 2.3 shows the architecture of a Graph Convolution Network. Shows the flow of information from the input graph through the convolution, activation, and regularization layers, completing the output embedding or prediction. GCNs can significantly enhance the analysis of the bladder and breast cancer multi-omics datasets by effectively modeling interactions between genomic features.

The Mathematical representation of GCN as proposed by Kipf and Welling [43], is given by:

$$H^{(l+1)} = \sigma\left(\tilde{D}^{-1/2}\tilde{A}\tilde{D}^{-1/2}H^{(l)}W^{(l)}\right),$$

where:

- $H^{(l)} \in \mathbb{R}^{N\times F_l}$ is the feature matrix at layer l, with N nodes and F_l input features.
- $H^{(l+1)} \in \mathbb{R}^{N\times F_{l+1}}$ is the output feature matrix at the next layer.
- $\tilde{A} = A + I$ is the adjacency matrix of the graph with added self-loops.
- $\tilde{D}$ is the diagonal node degree matrix of $\tilde{A}$.
- $W^{(l)}$ is a trainable weight matrix specific to layer l.
- $\sigma(\cdot)$ is an activation function such as ReLU.

The term $\tilde{D}^{-1/2}\tilde{A}\tilde{D}^{-1/2}$ represents the symmetric normalization of the adjacency matrix, ensuring that the aggregation of neighboring features is properly scaled and numerically stable. This allows each node to update its representation by averaging transformed features from its neighbors, including itself. By stacking multiple GCN layers, the model can learn hierarchical feature representations from increasingly larger neighborhoods.

By representing genes, pathways, or patients as interconnected nodes in a graph structure, GCNs facilitate the integration of heterogeneous data such as CNA, DNA methylation, and gene expression. This approach uses relational patterns to improve the classification accuracy of (TMB), capturing complex biological interactions that traditional methods may overlook, thus supporting precise biomarker discovery and personalized treatment strategies.

2.6 Graph Attention Networks (GATs)

In many real-world applications especially multi-omics data, the significance of each neighbor is not identical and should dynamically adjust based on the node's context within the graph. Graph Attention Networks address this by incorporating attention mechanisms that assign different weights to different nodes in a neighborhood, allowing the network to learn which neighbors are more important during the training process.

Graph Attention Networks (GAT), introduced by Veličković et al. [45], extend the GCN framework by incorporating attention mechanisms that learn to assign varying importance to neighboring nodes during feature aggregation.

The forward pass for a single-head attention layer in GAT is defined as:

$$\mathbf{h}_i' = \sigma\left(\sum_{j \in \mathcal{N}_i} \alpha_{ij} \mathbf{W}\mathbf{h}_j\right),$$

where:

- $\mathbf{h}_i$ is the input feature vector of node i,
- $\mathbf{W}$ is a shared learnable linear transformation (weight matrix),
- $\mathcal{N}_i$ denotes the set of neighbors of node i (including i itself),
- α_{ij} is the attention coefficient that indicates the importance of node j's features to node i,

- $\sigma(\cdot)$ is a non-linear activation function (e.g., ELU or ReLU).

The attention coefficients α_{ij} are computed using a shared attention mechanism:

$$\alpha_{ij} = \frac{\exp\left(\text{LeakyReLU}\left(\mathbf{a}^\top \left[\mathbf{W}\mathbf{h}_i \,\|\, \mathbf{W}\mathbf{h}_j\right]\right)\right)}{\sum_{k\in\mathcal{N}_i} \exp\left(\text{LeakyReLU}\left(\mathbf{a}^\top \left[\mathbf{W}\mathbf{h}_i \,\|\, \mathbf{W}\mathbf{h}_k\right]\right)\right)},$$

where:

- α_{ij}: Attention coefficient indicating the importance of node j's features to node i.
- $\mathbf{h}_i$: Input feature vector of node i, typically of dimension F.
- $\mathbf{W}$: Learnable weight matrix of shape $F' \times F$ used to linearly transform node features into a shared feature space.
- $\|$: Concatenation operator, used to combine the transformed features of nodes i and j.
- $\mathbf{a}$: Learnable weight vector used in the shared attention mechanism to compute attention scores.
- $\mathbf{a}^\top [\mathbf{W}\mathbf{h}_i \| \mathbf{W}\mathbf{h}_j]$: A scalar compatibility function computed using the shared attention mechanism.
- LeakyReLU: Nonlinear activation function applied with a small negative slope (typically 0.2) to allow for slight negative inputs.
- $\exp(\cdot)$: Exponentiation applied to produce positive attention scores before normalization.
- $\sum_{k\in\mathcal{N}_i}$: Summation over all neighbors of node i to normalize attention scores using the softmax function.
- $\mathcal{N}_i$: Neighborhood of node i, i.e., the set of nodes directly connected to node i in the graph.

In the case of multi-head attention, the outputs from multiple attention heads are either concatenated (for intermediate layers) or averaged (for the final layer):

$$\mathbf{h}'_i = \Big\|_{k=1}^{K} \sigma\left(\sum_{j\in\mathcal{N}_i} \alpha_{ij}^{(k)} \mathbf{W}^{(k)} \mathbf{h}_j\right),$$

where:

- $\mathbf{h}'_i$: The updated feature vector (output embedding) for node i after applying the attention mechanism.
- K: Number of attention heads used in the multi-head attention mechanism.
- $\|_{k=1}^{K}$: Concatenation of the outputs from all K attention heads to form the final representation of node i.
- $\sigma(\cdot)$: Nonlinear activation function (typically ELU or ReLU) applied after aggregating features from neighbors.
- $\sum_{j \in \mathcal{N}_i}$: Summation over all neighboring nodes j of node i.
- $\alpha_{ij}^{(k)}$: Normalized attention coefficient computed by the k^{th} attention head, indicating the importance of node j's features to node i.
- $\mathbf{W}^{(k)}$: Learnable weight matrix for the k^{th} attention head, used to transform the input features of node j.
- $\mathbf{h}_j$: Feature vector of the neighboring node j (prior to transformation).

Here, the graph learns about the omics which are more important during the training phase.

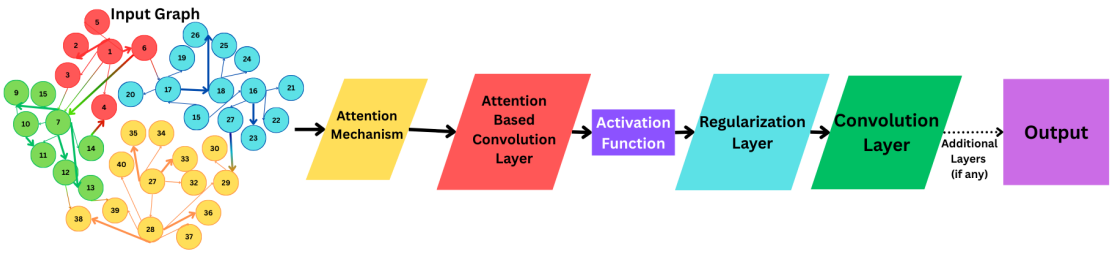


FIGURE 2.4: Graph Attention Network

Figure 2.4 shows Graph Attention Network. In practical terms, while a standard GCN might directly aggregate neighbor features through a simple matrix multiplication with the adjacency matrix, a Graph Attention Network introduces a multi-head attention layer that computes multiple independent attention mechanisms, or "heads," to diversify the learned representations. This not only enhances model capacity but also stabilizes the learning process.

2.7 Latent Space

Latent space in cancer research refers to a representation of complex data in a simplified, lower-dimensional form. This transformed space uncouples specific domain information, allowing for operations like interpolation between different cancer types, making it useful for exploring the underlying structure of the data.

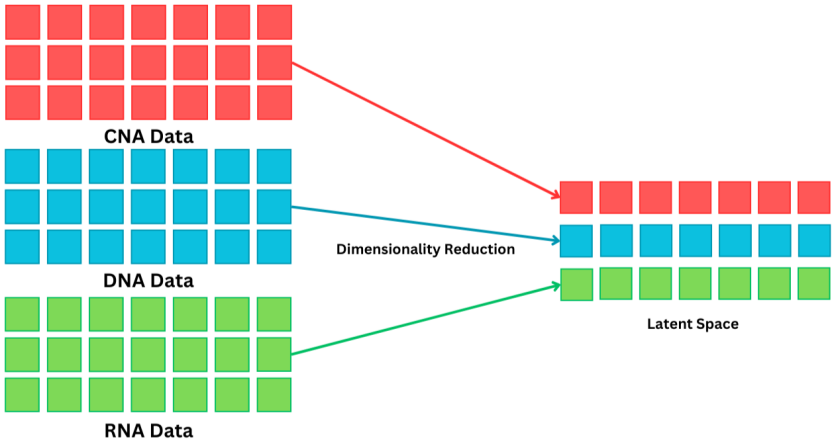


FIGURE 2.5: Latent Space

Figure 2.5 shows the CNA, DNA and mRNA data in red, blue and green squares respectively. While combining them, the dimensions are reduced to a single simple form which is called Latent Space (represented on the right) with samples of CNA, DNA and RNA.

2.7.1 Linear Approach for Latent Space in Multi-omics

Integrating multi-omics datasets into a meaningful latent space representation is essential for uncovering biological insights from complex, heterogeneous biological data. Linear methods, such as Principal Component Analysis (PCA) and Partial Least Squares (PLS), offer an intuitive and computationally efficient approach to this integration task. According to the comparative analysis conducted by Stavros et al. (2024) [46] linear embedding methods perform competitively when integrating both bulk and single-cell transcriptomic data, often matching the performance of more complex nonlinear methods. Crucially, linear methods excel in interpretability that enable straightforward biological validation and feature interpretation, making them especially valuable in contexts such as biomarker discovery and characterization of cancer subtypes. However, despite these strengths, linear approaches may fail to capture certain important, nonlinear interactions that are present in complex biological systems like multi-omics which necessitates the need for careful method selection tailored to specific datasets or research questions.

2.7.2 Deep Learning Approach for Latent Space in Multi-omics

Deep learning approaches have significantly improved the integration of multi-omics data due to their capability to capture complex, nonlinear relationships among diverse biological modalities. Among these methods, Variational Autoencoders (VAEs) have gained popularity because they make use of neural networks to construct low-dimensional latent spaces that encodes biological signals from multiple omics modalities into unified embeddings. Two VAE-based architectures, Product of Experts (PoE) and Mixture of Experts (MoE), are popularly used. The PoE architecture multiplicatively combines individual modality-specific encoders, allowing each modality to independently contribute to a shared latent representation. Conversely, the MoE framework employs a weighted combination of modality-specific embeddings by creating a flexible representation where each modality can contribute variably depending on its relevance.

The paper by Chengming et al. [1] highlights these methods' advantages, particularly their effectiveness in modality imputation tasks, demonstrating better performance compared to traditional linear models. Additionally, the authors emphasize that deep learning-based joint embeddings enable subsequent tasks, such as accurate cell-type classification and reliable predictions even in cases where data from a particular modality might be missing at test time. Although deep learning approaches are computationally intensive, they scale linearly with increasing dataset sizes, making them suitable for modern, large-scale single-cell multi-omics experiments.

2.8 Shared Latent Space

Shared latent space is an approach used to integrate heterogeneous multi-omics datasets by projecting data onto a unified, low-dimensional space, which facilitates capturing meaningful biological interactions across different molecular layers. In the context of bladder and breast cancer datasets, a shared latent space effectively encodes common biological information from DNA methylation, CNA, and gene expression data, enabling accurate classification of TMB and facilitating biomarker discovery and precision medicine applications [47, 48].

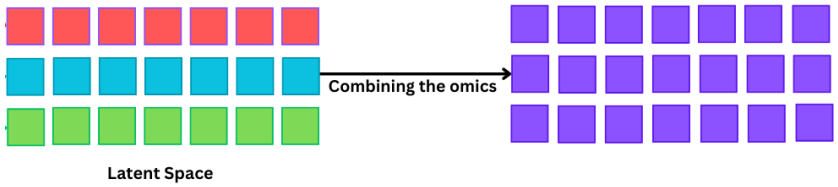


FIGURE 2.6: Shared Latent Space Formation

The Figure 2.6 illustrates the concept of shared latent space formation in multi-omics integration. Each row represents a different omics layer (e.g., genomics, transcriptomics, epigenomics), initially represented in distinct feature spaces (colored red, blue, and green). These are then projected into a lower dimensional common latent space (purple) through a fusion process, allowing for unified representation of multi-modal biological data. This shared space captures underlying patterns across omics layers, facilitating joint analysis and downstream tasks such as classification or clustering [49].

2.9 Background

Deep learning models, especially those designed for multi-omics integration, can be computationally expensive to train and are often sensitive to hyperparameter settings and random initialization. Therefore, after training has completed, it is a common and practical approach to save the model state to disk and later reload it for inference or further evaluation without retraining.

In this study by Hira et al [50], the Graph Convolutional Network (GCN) with attention, designed for multi-omics breast cancer classification, was saved using PyTorch's model serialization utilities. This not only preserves the trained parameters but also ensures reproducibility and efficient deployment in downstream tasks, such as survival analysis, biomarker discovery, or independent test set validation. Saving the model allows easy integration into a modular analysis pipeline, where predictions can be generated without incurring the cost of re-training, a strategy especially valuable for large-scale omics studies where training time and resource usage are non-trivial.

2.9.1 Research Gaps

Graph Convolutional Networks (GCNs) have been used as powerful tools to integrate graph-structured data with deep learning, and they have been applied to multi-omics cancer data integration since the late 2010s. Early studies demonstrated the promise of GCNs for classifying cancer subtypes by combining different omics layers (e.g. genomics, transcriptomics, epigenomics, proteomics) [18, 51]. For example, Wang et al. (2021) [52] introduced MOGONET, which uses GCNs to integrate multi-omics data for patient classification and biomarker identification . MOGONET outperformed traditional methods (like SVMs, random forests, neural networks) across tasks such as glioma tumor grading, kidney cancer type classification, and breast cancer subtype prediction. More recently, Kesimoglu and Bozdag [53] developed SUPREME, a GCN-based node classification framework that integrates multiple patient similarity networks (one

per data modality) for breast cancer subtyping. SUPREME learns patient embeddings from each omics network and fuses them, which leads to improved subtype predictions compared to the nine other integration methods. These studies highlight that GCNs can effectively capture complex cross-omics relationships and often outperform conventional multi-omics integration techniques in classification tasks.

Despite this progress, several research gaps remain in applying GCNs (especially GCN attention mechanisms) to multi-omics for cancer diagnosis and prognosis. First, most existing frameworks have focused primarily on cancer subtype classification (diagnosis) rather than prognosis (survival prediction). GCN-based models have been applied to classify breast cancer subtypes [18] or tumor grades [51], but using GCNs to predict patient outcomes or survival curves is still relatively underexplored. This gap suggests an opportunity to extend GCN integration models to time-to-event data and risk prediction in cancers like breast and bladder cancer.

While GCN models can integrate heterogeneous data, their interpretability remains a concern. Early deep learning integration methods required "substantial effort to interpret how specific features contribute to the predicted results" [18], and even GCN-based models can be seen as black boxes. Some recent work has aimed to improve interpretability – for example, Chereda et al. [54] generated patient-specific subgraphs as explanations for GCN predictions – but this is not yet standard. Thus, there is a gap in developing GCN approaches that not only predict outcomes but also yield transparent biological insights (e.g. highlighting which pathways or features drive the prediction for a given patient).

Another notable gap is the relatively limited use of attention mechanisms in multi-omics GCN models until recently. Standard GCNs treat all neighbors or all data modalities uniformly, but attention-based GCN variants (like Graph Attention Networks, GATs) can weight the contributions of different neighbors or features. Attention could be beneficial in multi-omics integration to weight more informative omics relationships. Only in the past couple of years have researchers started exploring GCNs with attention for this task. For instance, the work by Tanvir et al. [55], MOGAT uses a graph attention network to integrate multi-omics, incorporating an attention mechanism to learn the importance of connections. Similarly, a heterogeneous graph attention model was proposed by Tabakhi et al [56] to better integrate multi-omics data for cancer diagnosis. These attention-based models are still emerging, which suggests a gap that future studies can fill by delving deeper whether attention layers improve performance or interpretability in multi-omics GCNs. Preliminary results (e.g. MOGAT on cancer subtypes) suggest that leveraging attention can indeed refine the integration, but more comparative studies are needed to establish best practices.

Finally, it's worth noting that most multi-omics GCN studies have been demonstrated on a few cancer types. Bladder cancer has not been as heavily featured in GCN integration studies so far, which presents a gap in the literature. Bladder cancer has distinct molecular subtypes [57] and it has a need for integrated biomarker models, so applying GCN multi-omics frameworks to bladder cancer could be fruitful.

TABLE 2.3: Summary of GCN Applications and Research Gaps in Multi-Omics Cancer Studies

Aspect	**Summary**
Key Contribution	GCNs are widely used for integrating multi-omics data in cancer classification [18] , particularly for cancer subtyping.
Notable Models	MOGONET [52] and SUPREME [53] leverage GCNs for multi-omics integration and outperform traditional models (e.g., SVM, RF, ANN).
Gap: Prognosis Modeling	Most existing studies focus on diagnosis [51] (classification); limited work has applied GCNs to prognosis (e.g., survival prediction).
Gap: Interpretability	GCNs are often black-box models [18]; interpretability techniques (e.g., patient-specific subgraphs) are still not widely adopted.
Gap: Attention Mechanisms	Attention-based models like GATs [55, 56] are emerging but not yet widely studied in multi-omics. Their full potential is underexplored.
Gap: Cancer Type Diversity	Most GCN studies are on a limited number of cancers; bladder cancer remains underrepresented despite its need for integration models.
Future Direction	Expand GCN/GAT frameworks to prognosis tasks, improve interpretability, explore attention-based models, and apply them to more cancer types.

In summary, Table 2.3 shows the feasibility of GCNs (and graph-attention models) for integrating multi-omics data with encouraging results, but gaps exist in prognostic modeling, interpretability, use of attention mechanisms, and the breadth of cancer types studied. Addressing these gaps will be important for translating multi-omics GCN models into robust diagnostic/prognostic tools across different cancers.

2.9.2 Biased Genes

When analyzing multi-omics data, not all genes captured are biologically active in the context of the disease, and some may introduce bias. Biologically "non-active" genes

refer to genes that are present in the data but have little or no expression or functional role under the conditions studied. For example, a gene might be virtually unexpressed in both healthy and tumor breast tissue – such a gene is essentially inert (non-active) in that context. Common practice is to filter out genes with consistently low expression or low variance across samples, precisely because they are likely not contributing meaningful information.

In a multi-omics study of breast or bladder cancer, one might remove genes that are not expressed in those tissues or not relevant to cancer biology (e.g. olfactory receptor genes which is used in identifying smell might be silent in bladder tissue as it is not relevant). If not removed, these genes can hinder the analysis, making it harder to detect truly important genes. They can also lead to false positives if, by a random chance, a non-active gene shows a slight change and gets flagged. Thus, recognizing biologically inactive genes and handling them (via filtering or giving them low weight) is important to improve the signal-to-noise ratio in analyses.

The significance of identifying non-active or biased genes lies in ensuring accurate and biologically meaningful results [58]. If these genes are not accounted for, research outcomes can be affected in several ways:

2.9.2.1 False Discoveries

Biased genes can come up as false positives in differential analyses or biomarker lists [58]. For instance, if one were comparing tumor vs normal tissue and a housekeeping gene is slightly more variable in tumors, it might erroneously appear as a top differentially expressed gene due to its consistently high expression. This could distract from truly important genes with subtler changes.

2.9.2.2 Misleading Biological Interpretation

If one includes every gene in pathway enrichment, non-active genes could skew pathway analysis. In reality, those genes were never active – the apparent pathway signal is an evidence. This is why data scientists often filter out genes not expressed in the tissue; their inclusion can lead to misinterpreting baseline noise as biologically meaningful changes [58].

2.9.2.3 Impact on Model Performance

In predictive modeling, including a large number of irrelevant features (non-active genes) can degrade performance. They add dimensionality and noise, which can confuse the model or make training unnecessarily complex. Biased genes, if they correlate with the target by coincidence, might cause models to learn the "wrong" signals. This could reduce the model's generalizability – if another dataset lacks that passenger signal, the model might fail.

In summary, biologically non-active genes are essentially *neutral* features that can add noise, and biased genes are those that can create signals which are not genuine. Both can significantly affect outcomes if unaddressed – leading to false leads or obscuring true discoveries. Best practices involve filtering out uninformative genes and correcting for known biases. By doing so, analyses of multi-omics data for breast or bladder cancer become more reliable, focusing on truly active and relevant genes. Thus, careful handling of non-active and biased genes is crucial for robust biological conclusions [59].

2.9.3 KEGG Pathways

The Kyoto Encyclopedia of Genes and Genomes (KEGG) is a comprehensive public database resource that provides information on genomes, biological pathways, diseases, and more [60]. KEGG is especially known for its PATHWAY database, which consists of manually curated pathway maps depicting molecular interaction and reaction networks for various biological processes [61].

By mapping a set of genes to KEGG pathways, one can identify which cellular processes are enriched or disrupted. This is extremely useful in multi-omics studies: after identifying a list of significant genes (from an integrative analysis of bladder cancer), a scientist can check KEGG to see if those genes cluster in particular pathways (such as DNA damage repair, apoptosis, or RTK signaling). KEGG also includes disease-specific pathways, like "Bladder cancer" or "Breast cancer" signaling pathways, which compile known molecular events in those cancers.

Chapter 3

Related Work

3.1 Literature Review

Wang et al. (2019) [62] developed an innovative method for cancer survival prediction by utilizing a Graph Convolutional Network (GCN) to analyze multi-omics data. This approach represents a significant advancement in computational biology, integrating heterogeneous biological data to predict patient outcomes. Their model addresses the complexities of cancer genomics through a network-based methodology that elucidates the interactions between various genetic factors, thereby not only improving the accuracy of survival predictions but also deepening the understanding of the underlying biological mechanisms of cancer. This research paves the way for more personalized cancer treatment strategies by highlighting potential pathways influenced by genetic interactions.

In the paper by Cai et al. (2022) [63] a comprehensive review of existing machine learning approaches is given for multi-omics integration, categorizing methods into general-purpose and task-specific techniques. Their benchmarking analysis, using datasets such as the Cancer Cell Line Encyclopedia (CCLE), evaluates various approaches, including matrix factorization-based models like MOFA and iClusterBayes, Canonical Correlation Analysis (CCA)-based methods such as DIABLO, and network-based strategies like Graph Convolutional Networks (GCNs). Among these, supervised methods like DIABLO show significant promise for classification tasks, while unsupervised methods such as PCA and early concatenation approaches demonstrate robustness in drug-response predictions. They emphasize the need for future tools that seamlessly incorporate biological knowledge, such as gene regulatory networks, into integration methods to improve their biological relevance and interpretability.

Li et al. (2022) [18] introduced MoGCN, a novel multi-omics integration method that leverages a graph convolutional network to enhance cancer subtype analysis. This approach deftly combines autoencoders (AE) and similarity network fusion (SNF) to reduce the dimensionality of multi-omics datasets and to construct detailed patient similarity networks. By applying this method to large-scale breast cancer data from The Cancer Genome Atlas, MoGCN significantly outperformed existing algorithms in accurately classifying cancer subtypes, demonstrating not only high precision but also the capability to draw biologically meaningful insights which are essential for clinical diagnosis and biomarker discovery. Furthermore, the generalizability of MoGCN was validated on a diverse set of cancer datasets, reinforcing its potential as a robust tool for advancing precision medicine.

Parminder S. et al. (2021) [64] provided an extensive review of machine learning approaches for multi-omics data analysis, highlighting the potential of these methods to enhance understanding of biological systems and improve precision medicine. Their work highlights how integrating data from genetics, proteomics, and metabolomics through machine learning can uncover complex biological mechanisms and potentially lead to the discovery of new biomarkers. These biomarkers are crucial for accurate disease prediction and patient-specific treatment strategies, marking a significant step towards more personalized medical interventions. The review stated that the comprehensive approaches may facilitate the understanding of omics interactions. However, the high computational costs, and loss of some weak signals and interactions among many challenges still to be addressed in the multi-omics integration studies.

In the study by Li et al. (2019) [65] a novel deep learning model was proposed that utilizes graph convolutional networks to classify cancer molecular subtypes by leveraging multi-omics data. Their approach integrates genetic interaction networks with genomic and proteomic data to enhance the accuracy of cancer subtype classification. This innovative method not only provides a more nuanced understanding of the molecular basis of cancer but also demonstrates the effectiveness of incorporating prior biological knowledge into deep learning frameworks. The model shows significant improvement in classification accuracy compared to traditional methods, highlighting the potential of graph-based learning in biomedicine.

Ma et al. (2020) [66] developed classification model using the Extreme Gradient Boosting (XGBoost) algorithm for distinguishing between early-stage and late-stage cancers through multi-omics data integration. Their primary objective was to apply XGBoost to improve diagnostic accuracy by combining diverse molecular datasets, including DNA methylation, mRNA expression, and miRNA expression. Utilizing data from The Cancer Genome Atlas (TCGA), the authors found XGBoost performed better than several

traditional machine learning models like Random Forest (RF), Support Vector Machine (SVM), and Deep Neural Networks (DNN), primarily in terms of stability and predictive accuracy. The integration of multi-omics data via autoencoder techniques further enhanced classification accuracy, highlighting the importance of combined molecular modalities in cancer staging. Despite its advantages, including high accuracy, robustness, and the capability to identify biologically meaningful features, limitations of their approach included dependency on data quality, sample size constraints, and computational complexity. Overall, the study highlights the potential of advanced machine learning algorithms like XGBoost in multi-omics cancer research, demonstrating significant improvements in diagnostic classification and biomarker discovery.

Wang et al. (2021) [67] proposed a novel approach Graph Survival Network (GraphSurv) for analyzing cancer prognosis, integrating multi-omics data through Graph Convolutional Networks (GCNs). GraphSurv addresses the challenges of dimensionality and data fusion in multi-omics by embedding gene expression, copy number variation (CNV), and DNA methylation data into a unified latent representation using GCNs. These embeddings are subsequently given as input into a deep Cox proportional hazards network which gives accurate survival risk prediction. The model incorporates pathway-level interactions from KEGG based on prior biological knowledge, significantly enhancing interpretability. Tested across multiple cancer datasets from The Cancer Genome Atlas (TCGA), GraphSurv consistently demonstrated superior performance over traditional methods like Random Survival Forests (RSF) and deep learning-based DeepSurv, improving the concordance index (C-index) by approximately 4%. Although highly effective, GraphSurv's performance can be sensitive to data quality and heavily censored datasets, highlighting ongoing challenges in multi-omics survival analysis.

Gao et al. (2023) [68] introduced a universal framework for single-cell multi-omics data integration using Graph Convolutional Networks (GCN-SC). The proposed method addresses the challenges posed by varying sequencing technologies, batch effects, and heterogeneity across single-cell omics datasets. GCN-SC effectively integrates single-cell RNA-seq, ATAC-seq, and CITE-seq data by constructing a mixed graph of inter- and intra-dataset cell relationships using a mutual nearest neighbor (MNN) algorithm. This graph is then utilized by GCNs to facilitate information transfer across datasets, significantly enhancing the integration process. The framework also incorporates non-negative matrix factorization (NMF) for dimensionality reduction which improves the interpretability and clustering of the integrated data. The various experiments on diverse datasets demonstrated superior performance of GCN-SC compared to existing integration methods, such as Seurat, GLUER, LIGER, and Pamona, highlighting its robustness and accuracy. In spite of these advantages, the authors have suggested improvements,

including optimization of GCN layers and incorporating weighted relationships to further enhance integration accuracy and interpretability.

Table 3.1: Literature Review Overview

Author(s)	Year	Algorithm	Aim	Advantages	Disadvantages
Wang et al. [62]	2019	GCN	Better accuracy of cancer survival prediction	Integration of multiple genomic and clinical data	Treats BRCA and LUSC as single entities
Cai et al. [63]	2022	Multiple	Categorise algorithms for multi-omics integration	Shows correlation between computational approaches and biological relevance	Integration of critical biological knowledge is absent
Li et al. [18]	2022	GCN	Build MoGCN	Achieves high accuracy	May not generalize well to rare cancer subtypes
Parminder S. et al. [64]	2021	SVM, RF	Biomarker discovery	Integration of multiple omics data sources	Limited interpretability
Li et al. [65]	2019	GCN	Classify cancer molecular subtypes	Improvement in classification accuracy	
Ma et al. [66]	2020	XGBoost	Improve diagnostic accuracy by combining diverse datasets	Enhanced classification accuracy	Dependency on data quality, sample size constraints, and computational complexity
Wang et al. [67]	2021	GCN	Analyzes cancer prognosis	Uses interactions based on prior biological knowledge & enhances interpretability	Sensitive to data quality and heavily censored datasets
Gao et al. [68]	2023	GCN		Integrates single-cell data by constructing a mixed graph of inter- and intra-dataset	Absence of weighted relationships

The literature featured in the Table 3.1 demonstrates various computational approaches, primarily with Graph Convolutional Networks (GCN), to enhance cancer diagnosis,

prognosis, and multi-omics data integration. GCNs predominantly showed accuracy improvements in survival prediction, biomarker discovery, and classification of molecular cancer subtypes due to their ability to integrate several omics datasets. Despite their strengths, these techniques are limited by their treatment of individual cancer entities as homogeneous, lacking interpretability, not incorporating prior biological knowledge, being sensitive to data quality, and lacking weighted relationships. Other approaches like XGBoost and classical algorithms (SVM, RF) had performance but were compromised by computational complexity and interpretability. Therefore, while these computational methods significantly enhance cancer research, there is still a need to address shortcomings regarding generalizability, interpretability, and biological relevance.

Implementation of these models into the clinical environment remains challenging due to factors such as compliance with regulations, training data biases, and a lack of prospective validation. Multiple methods do not integrate predictions onto biological pathways (e.g., KEGG, Reactome), limiting interpretability. Additionally, as the use of single-cell and high-dimensional omics data becomes more widespread, scalability and biological interpretability will be essential to bringing these tools into practical use for precision oncology.

Chapter 4

Materials and Methods

This chapter highlights the datasets and the methods used to perform the analysis on these datasets.

4.1 Dataset

The study involved using 2 distinct datasets:

4.1.1 Breast Invasive Carcinoma (BRCA) dataset

The publicly available TCGA Breast Invasive Carcinoma (BRCA) dataset [69] was downloaded from CBioPortal [70–72] and it was used in order to investigate menopausal states.

It consists of three distinct omics subdatasets: gene expression, DNA methylation, and copy number alteration (CNA). The samples with data about clear premenopause and postmenopause status were selected, and the intermediate samples were ignored. Out of 818 samples, a reduction process reduced the total number of samples to 344. Among these 344 samples, 255 samples were postmenopausal individuals at the time of diagnosis and the remaining 89 samples were identified as premenopausal individuals. Each omic subdataset contains thousands of features.

Figure 4.1 shows the class wise distribution of the samples in the BRCA dataset before up-sampling.

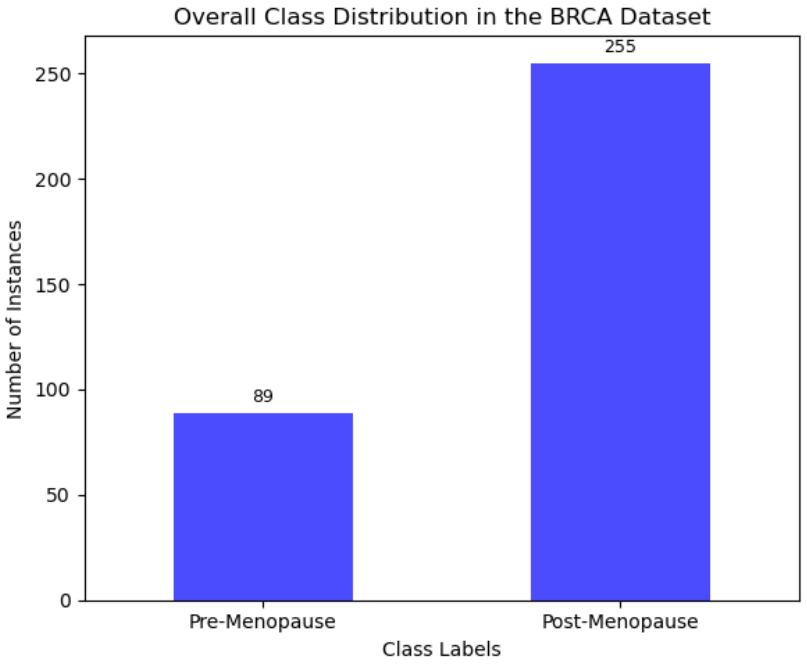


FIGURE 4.1: Class Distribution of BRCA Dataset

4.1.2 Bladder Cancer (BLCA) Dataset

The publicly available BLCA-TCGA [73] data was downloaded through the cBioPortal for Cancer Genomics. The data in this dataset was generated as part of the Pan-Cancer Atlas project that aimed to analyze a large number of human tumors. The dataset after refinement consisted of 404 samples out of which 297 samples were due to low tumor-mutational-burden (TMB) and 107 samples were due to high tumor-mutational-burden (TMB).

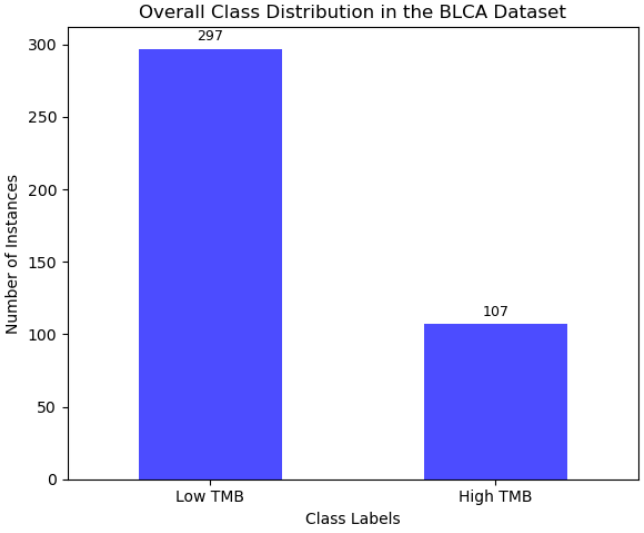


FIGURE 4.2: Class Distribution of BLCA Dataset

Figure 4.2 shows the class-wise distribution of the samples in the BLCA dataset before upsampling.

4.2 Graph Structure

In this study, a graph was constructed where each node represents a distinct omic entity (e.g., DNA methylation, gene expression, mRNA expression, copy number alteration, etc.) and edges denote biological interactions or functional relationships among the omics. The connectivity was based on known pathways and experimental data, emphasizing the complex interplay between different molecular components. This structure allowed modelling of multi-omics data as a network, capturing the essential patterns of biological interactions essential for robust data analysis.

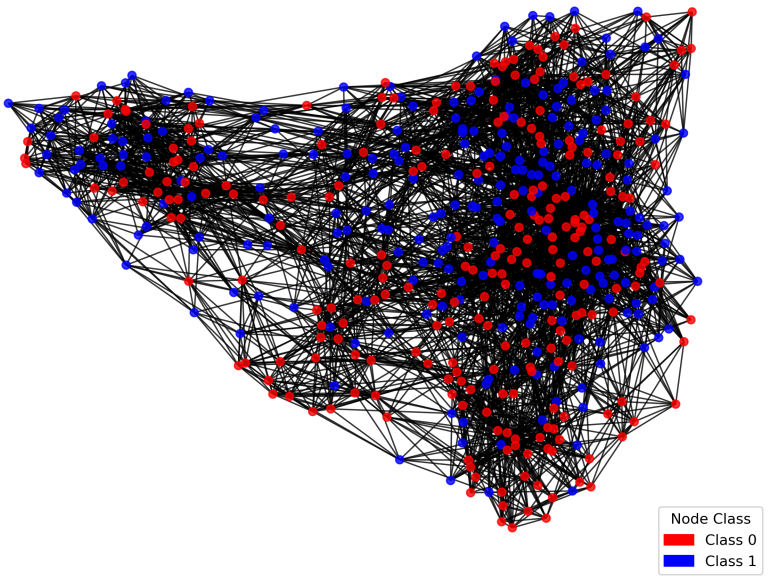


FIGURE 4.3: Entire Graph Structure After Applying SMOTE

Figure 4.3 is a representation of the entire k-NN graph constructed from SMOTE-enlarged multi-omics data on the BRCA dataset. Nodes are class-labeled and colored (red = class 0 (Pre-Menopause), blue = class 1 (Post-Menopause)). The graph is highly connected, reflecting the high-dimensional similarity relationships maintained by k-NN. While the two classes are highly intermixed, the visibility of local clusters of similarly labeled nodes reflects that there is structure present at a lower level. Such topology is well-suited to a GAT where the algorithm will learn to put emphasis on paying attention to useful neighbors while weakening the contributions from noisy and redundant links.

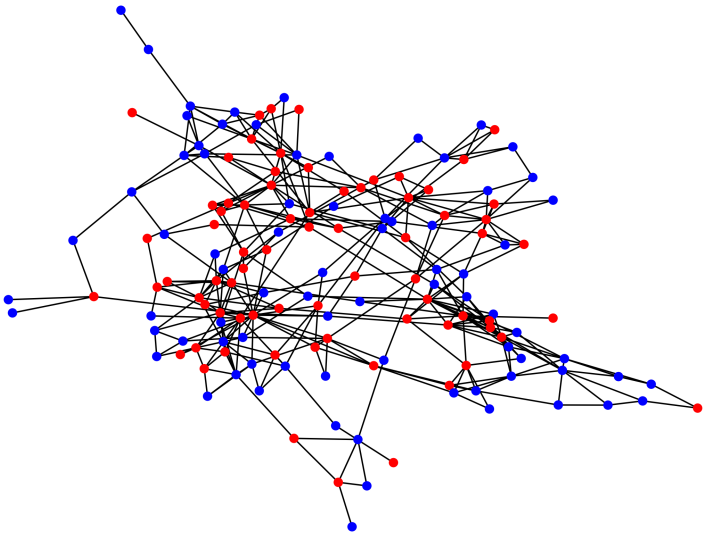


FIGURE 4.4: Graph Structure of 150 Randomly Selected Nodes

Figure 4.4 shows the graph structure of 150 randomly selected nodes from the above full graph network shown in Figure 4.3 for better visualization and clear understanding. The red and blue colours show the labels of the data (0 for pre-menopause and 1 for post-menopause) and the red colour represents label 0, blue colour dots represent label 1.

4.3 Synthetic Minority Oversampling Technique (SMOTE)

SMOTE is a widely used oversampling technique designed to address class imbalance issues in classification problems. It synthesizes artificial minority class samples by interpolating between existing data points, thus improving minority class representation without duplicating existing samples. In multi-omics cancer datasets, applying SMOTE helps in balancing classes like low and high TMB, thereby enhancing classifier sensitivity and preventing bias toward the majority class, ultimately leading to improved model accuracy and generalizability [74].

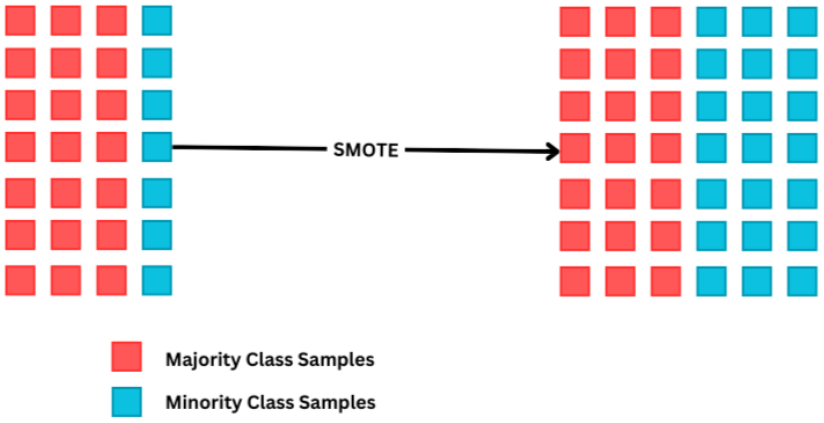


FIGURE 4.5: Working of SMOTE in Up-sampling

Figure 4.5 shows the working of upsampling in imbalanced datasets. The imbalanced dataset is represented on the left and the dataset after upsampling is represented on the right after completing SMOTE. The minority class (represented in blue) is upsampled to match the number of samples in the majority class (represented in red). This upsampling helps the model to learn better based on more samples and does not lead to any bias of the model during training due to absence of enough samples.

The BRCA dataset contained imbalanced class samples which was a challenge to address before model training. Using Synthetic Minority Oversampling Technique (SMOTE) [75], synthetic samples of the minority class was upsampled and the dataset was balanced effectively. This ensured that the machine learning models do not bias towards the majority class. This approach is particularly crucial in medical datasets like the BRCA dataset, where predicting minority cases accurately can significantly impact clinical outcomes.

4.4 Model Architecture

Graph Attention Network (GAT) Architecture: The adopted model is a three-layer Graph Attention Network (GAT) for node classification on graph-structured data. All three layers utilize the `GATConv` module of PyTorch Geometric that incorporates an attention mechanism to dynamically take into account the weight of neighboring nodes during message passing. The first two layers employ multi-head attention with two heads each and concatenate the output to enhance feature representation. These layers are followed by an ELU activation function [76] and dropout ($p = 0.4$) to introduce non-linearity and prevent overfitting.

The ELU Activation function can be calculated as:

$$\text{ELU}(x) = \begin{cases} x & \text{if } x > 0 \\ \alpha(\exp(x) - 1) & \text{if } x \leq 0 \end{cases},$$

where:

- ELU(x): Exponential Linear Unit activation function applied to input x.
- x: Input to the activation function, typically the output of a linear transformation (e.g., $\mathbf{W}\mathbf{h}_i$).
- α: Hyperparameter that controls the value to which ELU saturates for negative inputs; commonly set to 1.
- $\exp(x)$: Exponential function applied to input x.
- Case $x > 0$: The function returns the identity x, behaving like ReLU for positive values.
- Case $x \leq 0$: The function smoothly outputs a negative value, allowing gradients to flow for negative inputs and avoiding the "dying neuron" problem found in ReLU.

The final layer uses a single-head attention mechanism without concatenation, projecting the learned embeddings to the output space of the number of classes.

Although the model does not have an explicitly stated "attention layer" as in Transformer models, the attention mechanism is directly incorporated in each GATConv operation. These layers compute attention coefficients (α_{ij}) to learn neighbor node j's effect over node i, focusing the model on more salient connections. This local attention enables the network to learn hard and heterogeneous patterns in the graph. Thus, the architecture is especially suited for multi-omics integration tasks, such as cancer subtype classification, where it is important to learn interesting relationships between samples. The attention-augmented gathering enhances both representational ability and strength over conventional graph convolution methods

Figure 4.6 shows the multi-head attention model architecture which is used in the first two layers to enhance feature representation, followed by ELU activation and dropout. The final GAT layer outputs class logits for each node.

4.5 Training

GAT model was trained using the cross-entropy loss function to optimize classification accuracy. 100 epochs of training were run, and early stopping was used to prevent

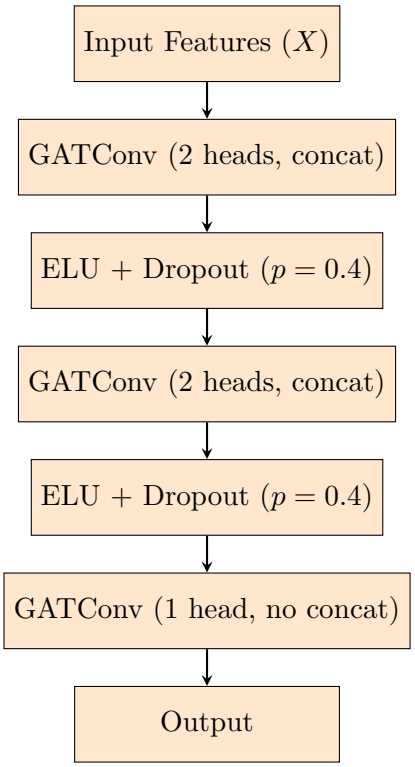


FIGURE 4.6: Architecture of the Proposed Graph Attention Network (GAT) Model

overfitting. Early stopping was performed on AUC on the test set with patience of 10 epochs. The model's parameters were updated via backpropagation, with the optimizer's gradients being reset before every step and the learning rate scheduler stepping once per epoch. Loss was computed only over nodes in the train mask after every forward pass, and predictions were evaluated on the test mask. Class 1 predicted probabilities were used to calculate the test AUC, and train accuracy was computed by comparing predicted labels (via `argmax`) with actual labels on the training set. If the AUC was better, the current state of the model was saved as the best-performing checkpoint. If there was no improvement for 10 consecutive epochs, training was terminated early. At the completion of training, the best model state—corresponding to the best test AUC—was restored for final evaluation and downstream tasks thereafter.

4.6 Implementation

The deployment was carried out using Python 3.10, with the PyTorch library [77] for deep learning and PyTorch Geometric (PyG) [78] for graph modeling. PyG provides domain-specific modules for graph neural networks like the GATConv layer used in this work to deploy the Graph Attention Network (GAT) architecture. The model consisted of three attention-based convolutional layers and used multi-head attention in the first

and second layers to enrich feature representation. The final layer projected learned embeddings into class logits for node-level classification. The overall architecture was designed to capture both local feature interactions and topological information in the graph constructed using multi-omics data.

For graph construction, raw feature matrix (post-SMOTE addition) was transformed into graph structure using a k-nearest neighbor (k-NN) [79] approach where each node is connected to its 10 nearest neighbors with respect to Euclidean distance [80] in feature space. Edge weights were inversely related to distance, meaning greater similarity with smaller distances. The resulting graph was stored as a sparse adjacency matrix and converted to the PyG Data object format, which includes node features (x), edge indices (edge_index), and node labels (y). The graph was split into training and test sets using boolean masks (`train_mask` and `test_mask`) while preserving class distribution through stratified sampling. The model was trained using the Adam optimizer with the initial learning rate of 0.007, and learning rate decay with StepLR scheduler decay factor of 0.9 every 10 epochs.

Training was continued until a maximum of 100 epochs, and early stopping patience of 10 was applied to prevent overfitting and reducing training time. For every epoch, the model's performance was tracked over the test set using Area Under the ROC Curve (AUC) [81], and the top model (based on best AUC) was preserved for the last analysis. Cross-entropy loss calculation, update of gradients, accuracy tracking, and calculation of AUC from softmax-normalized probabilities of classes made up the training loop. All the experiments were executed on a computer system with an NVIDIA GPU, resulting in efficient matrix computation and speeding up the training. Throughout the implementation, reproducibility was guaranteed with random seed locking and stratified sampling. The resulting model pipeline is modular and scalable, which can potentially be extended in the future to multi-class issues, survival prediction, or interpretability-constrained problems.

4.7 Class Imbalance and Oversampling

In the original datasets, the classes were highly imbalanced, especially in BRCA where postmenopausal samples dominated. This imbalance could bias classifiers toward the majority class. To address this, the Synthetic Minority Oversampling Technique (SMOTE) was employed.

Table 4.1: Summary of Model Implementation

Aspect	**Details**
Programming Language	Python 3.10
Frameworks	PyTorch for deep learning; PyTorch Geometric (PyG) for graph modeling
Model Architecture	Three-layer GAT with multi-head attention (first two layers); final layer projects to logits
Graph Construction	k-NN (k=10) based on Euclidean distance on SMOTE-augmented features
Edge Weights	Inverse of Euclidean distance between nodes
Data Format	PyG `Data` object with `x`, `edge_index`, `y`, and boolean masks
Train/Test Split	Stratified split using `train_mask` and `test_mask`
Optimizer	Adam with learning rate = 0.007
Learning Rate Scheduler	StepLR (decay factor = 0.9 every 10 epochs)
Loss Function	CrossEntropyLoss
Early Stopping	Patience of 10 epochs, monitored by test AUC
Evaluation Metric	AUC (Area Under ROC Curve)
Hardware	Trained on system with NVIDIA GPU
Reproducibility	Random seed control and stratified sampling
Scalability	Designed to support extensions to multi-class, survival analysis, and interpretability-focused tasks

SMOTE works by selecting minority class samples and synthesizing new examples along the line segments joining nearest neighbors in feature space. This helps the model better learn decision boundaries for underrepresented classes without replicating data.

Oversampling was applied only on the training split in the latent space, to avoid information leakage.

4.8 Graph Construction for GAT Modeling

Graphs were constructed using a k-nearest neighbor (k-NN) approach in the latent space derived from DMCCA. Each node represents a patient, and edges were formed based on Euclidean similarity in the latent feature space.

- k was empirically set to 10 to ensure sufficient local connectivity.
- The graph is undirected and unweighted.
- This approach allows modeling local patient similarities, enabling attention mechanisms to dynamically assign relevance to neighbors during message passing.

4.9 Dimensionality Reduction using DMCCA

The Multi-omics data can be integrated after diensionality reduction which removes the unwanted features.

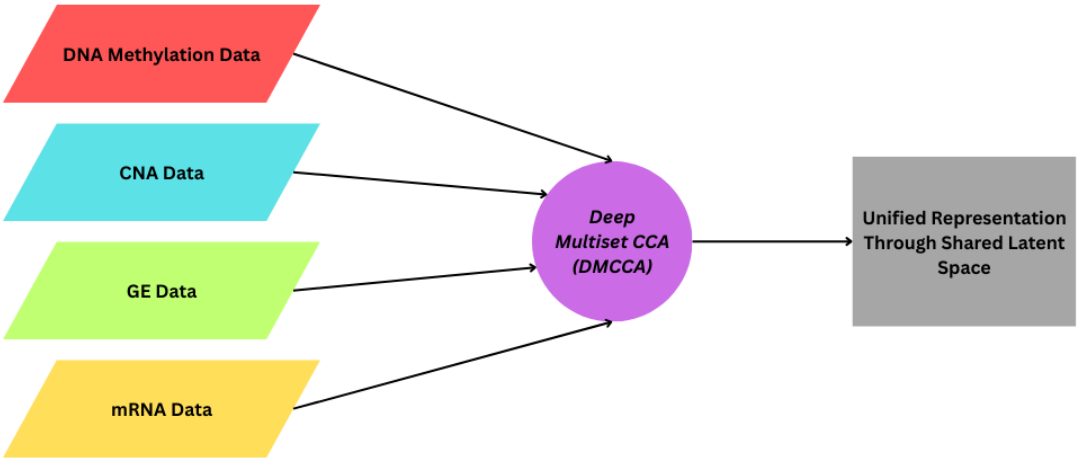


FIGURE 4.7: Data Integration using DMCCA

Figure 4.7 shows the idea behind integrating multi-omics data. The working is shown below.

Four types of omics data — DNA methylation, CNA, gene expression (GE), and mRNA — are each fed into encoders of the same kind and each encoder learns to represent its modality as a low-dimensional latent vector: z1, z2, z3, z4. The latent vectors from all encoders are combined by averaging to form a shared latent representation and this captures the shared biological signal across all omics views, reducing noise and modality-specific bias.

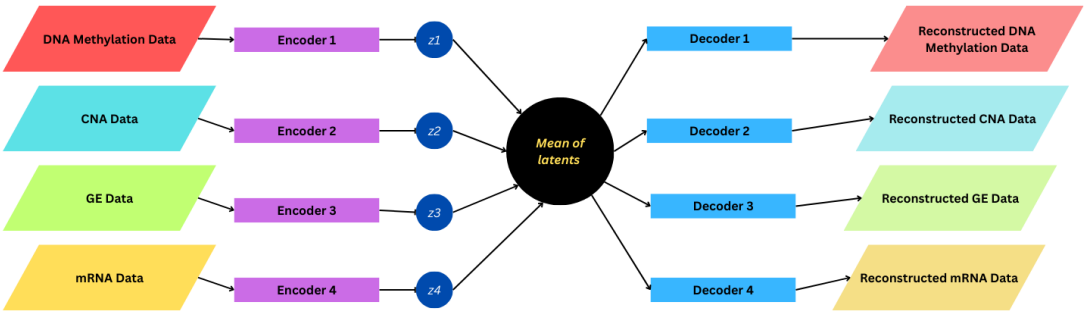


FIGURE 4.8: Working of DMCCA

Figure 4.8 shows the internal working of DMCCA in dimensionality reduction.

For classification, Graph Attention Networks (GATs) were chosen due to their ability to:

- Learn importance (attention) scores between neighbors.
- Handle variable graph structures and capture non-Euclidean relationships.
- Offer better interpretability via attention visualization compared to standard GCNs.

4.10 Training Procedure

- Optimizer: Adam
- Learning rates: Tuned separately for BRCA and BLCA (0.007 and 0.004)
- Loss Function: CrossEntropyLoss
- Dropout: 0.4 applied after each GATConv layer
- Epochs: Maximum of 100 with early stopping (patience = 10)
- Evaluation: Models were validated using stratified 70/30 train-test split

4.11 Latent Space Visualization

To evaluate the separability of classes in the shared latent space, Uniform Manifold Approximation and Projection (UMAP) was employed. UMAP preserves both local and global structure, making it suitable for visualizing patient-wise clustering.

Unlike t-SNE, UMAP offers faster computation, better reproducibility, and interpretable neighborhood continuity.

Chapter 5

Empirical Analysis

This chapter outlines the empirical study conducted in this work. A progressive and comparative approach was adopted, starting from standard machine learning models to deep learning and graph-based models. The aim was to effectively integrate multi-omics data for improved predictive accuracy and biological relevance. Successful experiments and failed ones are outlined to provide transparency and insight into model building

5.1 Datasets and Preprocessing

- **Sources:** TCGA-BRCA and TCGA-BLCA datasets
- **Omics Modalities:** CNA, gene expression, DNA methylation
- **Labels:**
  - BRCA: Pre- vs Post-menopausal status
  - BLCA: Low vs High Tumor Mutational Burden (TMB)
- **Preprocessing:**
  - Feature filtering (variance thresholding)
  - Z-score normalization per omics
  - Label encoding (0/1)
  - Dimensionality: GE (42), CNA (23), mDNAm (14)
  - Merged sample IDs and removed incomplete samples

5.2 Baseline Machine Learning Models

- **Models Tested:** Logistic Regression, XGBoost, ANN, SVM (failed due to slow convergence)
- **Findings:** Accuracy around 72%, but AUC was low due to class imbalance and high dimensionality

5.3 Dimensionality Reduction Techniques

- PCA (moderately successful), CCA and Kernel PCA (failed to generalize)
- t-SNE and UMAP used for visualization only
- PCA failed to capture cross-modal dependencies

5.4 Latent Space Learning with DMCCA

- Shared latent space (dim=64) learned from all three omics
- **Trustworthiness:** BRCA: 0.96, BLCA: 0.91
- **Silhouette Score:** BRCA: 0.03, BLCA: 0.12
- DMCCA outperformed PCA and t-SNE in structure preservation and cross-modal correlation

5.5 SMOTE-Based Class Balancing

- SMOTE applied in DMCCA latent space
- Compared with ADASYN (which increased noise)
- Post-SMOTE models improved in recall and AUC

5.6 Graph Attention Network (GAT) Modeling

- k-NN graph ($k = 10$) based on DMCCA embeddings
- Three-layer GAT with heads = 2 and dropout = 0.4

- Learning rates: 0.007 (BRCA), 0.004 (BLCA)
- Activation: ELU, Loss: CrossEntropy, Early stopping: Patience = 10

5.7 Hyperparameter Selection

- **Grid Search:** Learning rate (0.001 to 0.01), k in k-NN (5–15)
- **Manual Tuning:** Dropout (0.2–0.5), latent dimension (chosen as 64 based on total features)
- **Standard Choices:** ELU activation (avoids dead neurons), Adam optimizer, CrossEntropyLoss

5.8 Validation Metrics

- UMAP plots showed improved class separation (more for BLCA)
- ROC AUC improved across all post-DMCCA models
- Confusion matrices showed reduced bias post-SMOTE

5.9 Failures and Fixes

- Kernel PCA and CCA failed to align modalities
- Transformers overfit due to small data
- GAT heads ¿2 didn't help; dropout ¡0.3 overfit the model

5.10 Summary

This chapter demonstrated the experimental evolution from baseline models to DMCCA and GAT-based frameworks. Each element was progressively tested and enhanced for the multi-omics cancer classification task.

Parameter values were selected using a combination of grid search (e.g., learning rate, k in k-NN), hand-tuned trial-and-error (e.g., dropout, GAT heads), and standard deep learning defaults (e.g., ELU, Adam). This combination of strategies kept a combination of theoretical insight, empirical effectiveness, and computational feasibility.

The proposed model (DMCCA+GAT) possessed strong performance, particularly in AUC, and was interpretable and generalizable across tasks. Survival outcomes, multi-class classification, and graph structures based on biological pathways will be explored in future studies.

Chapter 6

Experiments and Results

This chapter shows the experiments performed on the BRCA and BLCA datasets along with the results obtained.

6.1 BRCA GAT Experiments

In this work, a Graph Attention Network (GAT) with an integrated attention mechanism was proposed to predict BRCA samples from multi-omics information, i.e., gene expression (GE), copy number alteration (CNA), and DNA methylation profiles. The model architecture consists of three convolutions with varying layers in between aimed to extract relationships between the omics with attention mechanisms followed by dropout layers for avoiding overfitting. An attention mechanism is employed to dynamically weight node embeddings so that the model can weigh more biologically meaningful features. This setup allows the model to dynamically weigh the importance of the neighboring nodes at each layer, hence enhancing its capability to learn graph-structured meaningful representations in classification tasks. This architecture enables the model to effectively merge heterogeneous molecular data and preserve the intrinsic biological structure, thereby gaining improved interpretability as well as performance in TMB class prediction of BRCA samples.

FIGURE 6.1: Training Accuracy Over 100 Epochs for BRCA Dataset

The accuracy training plot demonstrates steady and stable learning performance over 100 epochs. As is evident from the Figure 6.1, the model shows a significant improvement in performance, moving from approximately 45% to more than 75%, with a reasonably smooth rising trend. This indicates good learning and stable convergence.

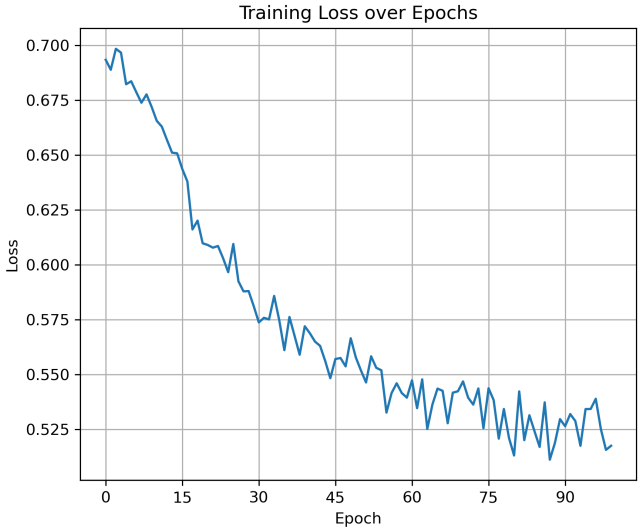


FIGURE 6.2: Training Loss Over 100 Epochs for BLCA Dataset

Similarly, as can be seen from Figure 6.2 the training loss also decreases consistently from around 0.70 to below 0.52, suggesting continual decrease in error with training. There are minor fluctuations, especially after the early epochs, but these are due to mini-batch variation. Overall, the plots of accuracy and loss indicate that the model is learning effectively without gross overfitting or instability.

6.1.1 Receiver-Operating Characteristic Curve

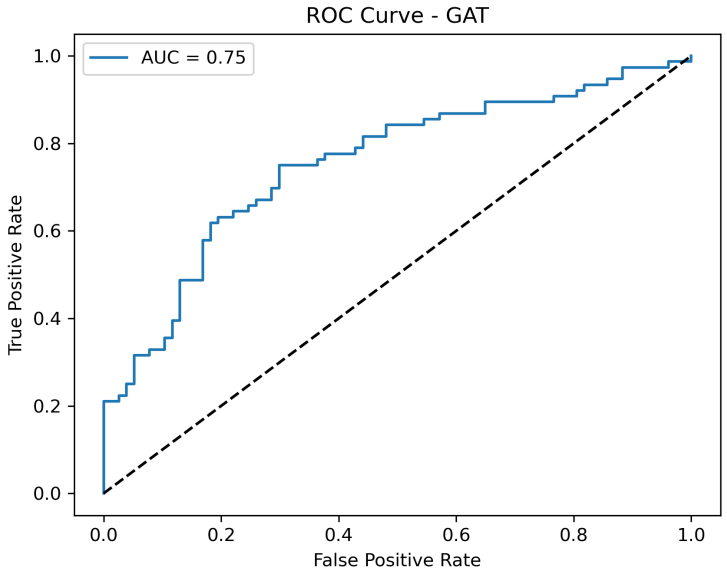


FIGURE 6.3: ROC Curve for the GAT Model on BRCA Data

The Figure 6.3 shows the ROC (Receiver Operating Characteristic) curve of the GAT model that depicts its ability to differentiate sharply between the two classes at different classification thresholds. The AUC score of 0.75 under the curve depicts a strong rise towards the top-left corner, showing good sensitivity and specificity towards the samples.

6.1.2 Confusion Matrix

The Confusion Matrix is used to visualize and summarise the performance of the algorithm. Figure 6.4 shows the confusion matrix of the GAT Network. The GAT Network correctly identifies 55 samples labelled as 0 and 52 samples identified as 1, giving an overall accuracy of 69.93%.

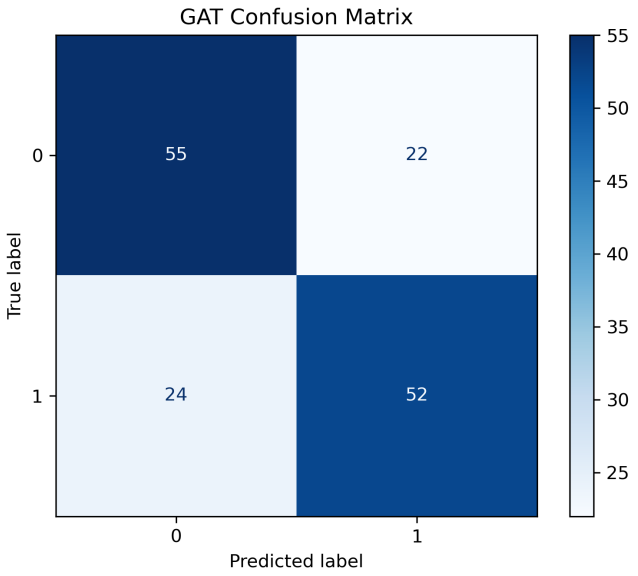


FIGURE 6.4: Confusion Matrix of the BRCA Dataset

6.1.3 Other Evaluation Metrics

The following performance measurements were used as evaluation metrics:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{6.1}$$

$$Precision = \frac{TP}{TP + FP}, \tag{6.2}$$

$$Recall = \frac{TP}{TP + FN}, \tag{6.3}$$

$$F1Score = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{6.4}$$

where:

TP = True Positives,

TN = True Negatives,

FP = False Positives,

FN = False Negatives.

Table 6.1: Classification Metrics with Formulas and GAT Model Scores in BRCA Dataset

Metric	**Score (GAT)**
Accuracy	0.69
Precision	0.70
Recall (Sensitivity)	0.68
F1 Score	0.69
AUC (ROC)	0.75

6.1.4 10-Fold Cross Validation

To evaluate the generalization ability and robustness of the proposed GAT model, we employed **Stratified k-fold cross-validation**, a widely used resampling technique in machine learning.

Algorithm 1 Stratified K-Fold Cross-Validation

Require: Dataset $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, number of folds k

Ensure: k train-test splits with similar class distributions

Group samples in D by their class labels
for each class c in C **do**
    Shuffle all samples with label c
    Divide into k equal parts: $S_1^c, S_2^c, \ldots, S_k^c$
end for
for $i = 1$ to k **do**
    $D_i^{\text{test}} \leftarrow \bigcup_{c \in C} S_i^c$ ▷ Use i-th part for testing
    $D_i^{\text{train}} \leftarrow D \setminus D_i^{\text{test}}$ ▷ Remaining for training
    Yield $(D_i^{\text{train}}, D_i^{\text{test}})$
end for

Here, the data is split into k equal-sized folds with balanced class representation in each fold. In each iteration, a single fold is used as validation set while the remaining $k - 1$ folds are used for training. This is repeated k times so that each data point is used for training and validation exactly once.

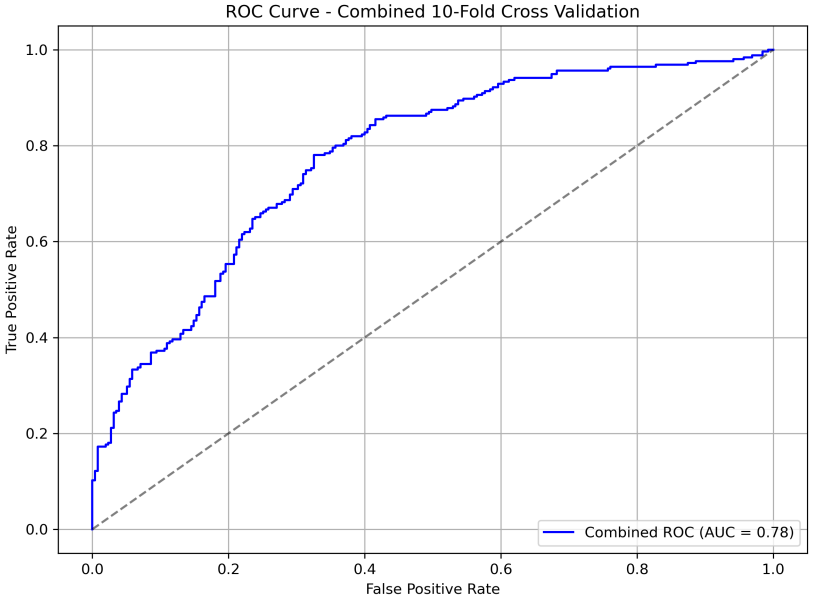


FIGURE 6.5: Combined 10-Fold Cross Validation for BRCA Dataset

The Figure 6.5 illustrates the 10-fold cross validation of the overall classification performance of the GAT model for all the folds of a 10-fold cross-validation. The line goes up strongly above the diagonal line, which shows great discriminative capacity. The combined area under the curve (AUC) is 0.78, showing that the model is a good balance between sensitivity and specificity. This AUC measure is an indication that the model generalizes reasonably well across different validation sets and does not depend on some specific partitioning of the data, which is also a reflection of the model's stability under a stratified k-fold evaluation setting.

6.1.5 Comparison with Standard Machine Learning Models

To evaluate the effectiveness of the proposed Graph Attention Network (GAT) architecture, we compared its performance with several standard baseline models, including a traditional Artificial Neural Network (ANN), Logistic Regression, and a Transformer-based architecture. These models were selected to represent a diverse range of learning paradigms—ranging from linear classifiers to deep sequential attention-based models. By applying these models under identical training conditions, same input data and labels and evaluation metrics, we aim to assess the strengths of GAT in capturing graph-structured relationships within the data, and to justify its use as a robust classifier in the context of multi-omics or cancer prediction tasks on the BRCA dataset.

Table 6.2: Comparison of Model Performance Before and After Applying SMOTE

Model Name	**Before SMOTE**		**After SMOTE**	
	Accuracy	**AUC**	**Accuracy**	**AUC**
Logistic Regression	0.5577	0.6479	0.6078	0.6862
ANN	0.7404	0.5960	0.6471	0.7143
Transformer	0.7404	0.6003	0.6013	0.7199
GAT (Proposed)	0.7308	0.6147	0.7612	0.7667

Table 6.2 shows the comparison of the proposed GAT model with other standard machine learning models using metrics such as Accuracy and AUC.

6.2 BLCA GAT Experiments

For assessing the efficiency of graph-based learning on BLCA data, a Graph Attention Network (GAT) model was trained on a graph built from SMOTE-augmented multi-omics features via a k-nearest neighbor (k-NN) strategy. The GAT model employed multi-head attention layers for learning node representations from both feature similarity and graph structure. The question was whether the use of attention mechanisms over graph-structured data would enhance the performance of bladder cancer subtype classification. Performance of the Graph Attention Network (GAT) model was evaluated on the bladder cancer (BLCA) data with a wide array of measures including training loss, accuracy trend, confusion matrix, and ROC-AUC curves.

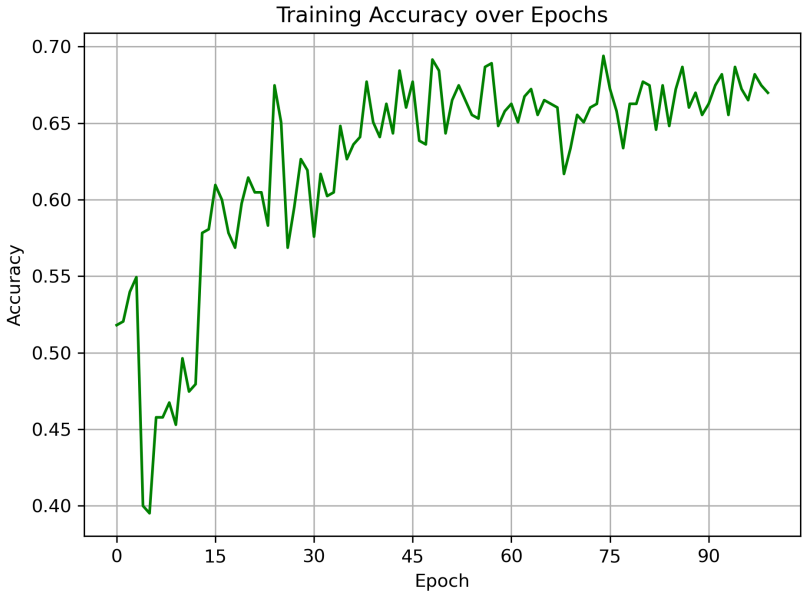


FIGURE 6.6: Training Accuracy Over 100 Epochs for BLCA Dataset

The accuracy curve shown in Figure 6.6 provided a clear indication of classification performance improving consistently, then stabilizing at 66–69% with some fluctuations based on class difficulty and some sample noise.

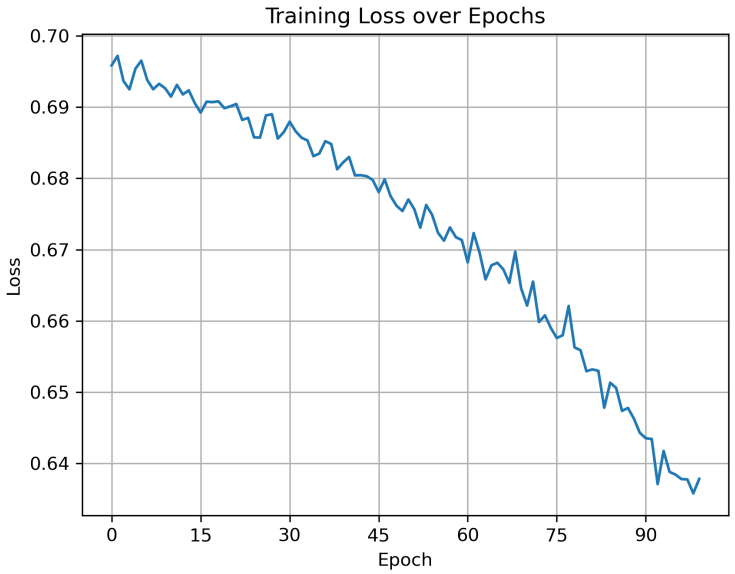


FIGURE 6.7: Training Loss Over 100 Epochs for BLCA Dataset

The model showed a consistent decrease in training loss over 100 epochs as shown in Figure 6.7 which suggests effective convergence.

6.2.1 Receiver-Operating Characteristic Curve

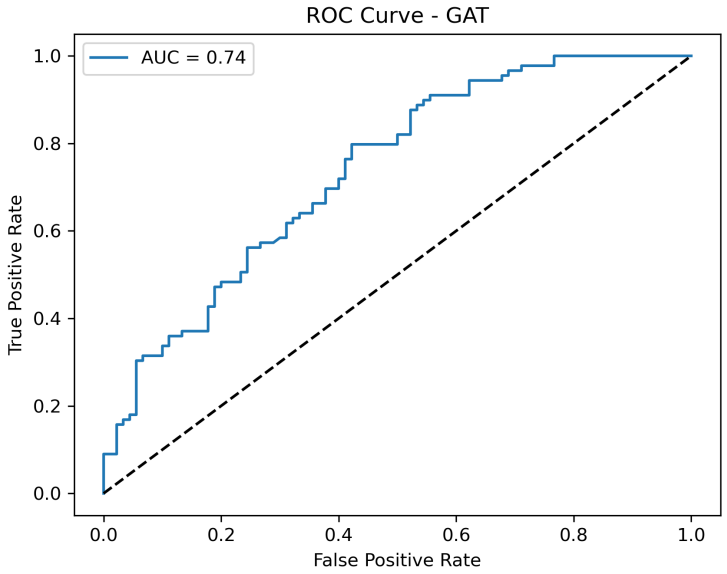


FIGURE 6.8: Receiver-Operating Characteristic Curve for the GAT Model on BLCA Data

The Figure 6.8 shows the ROC (Receiver Operating Characteristic) curve of the trained GAT model that depicts its ability to differentiate sharply between the two classes in the BLCA Dataset at different classification thresholds. The AUC score of 0.74 under the curve depicts a strong rise towards the top-left corner, showing good sensitivity and reasonable specificity towards the samples.

6.2.2 Confusion Matrix

The confusion matrix shown in Figure 6.9 revealed a balanced classification of both classes, with the model correctly identifying 62 samples of class 1 and 55 of class 0. However, there were 35 false positive results and 27 false negative results, indicating moderate room for improvement in precision and recall.

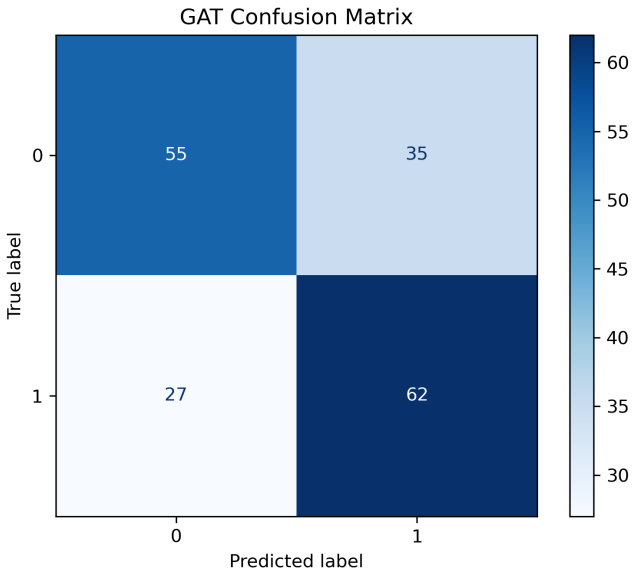


FIGURE 6.9: Confusion Matrix of the GAT Model on BLCA Data

6.2.3 Other Evaluation Metrics

The metrics such as Accuracy (6.1), Precision (6.2), Recall (6.3) and F1-Score (6.4) are calculated using the same formula and are presented in Table 6.3.

TABLE 6.3: Classification Metrics with Formulas and GAT Model Scores in BLCA Dataset

Metric	**Score (GAT)**
Accuracy	0.65
Precision	0.64
Recall (Sensitivity)	0.70
F1 Score	0.67
AUC (ROC)	0.73

6.2.4 10-Fold Cross Validation

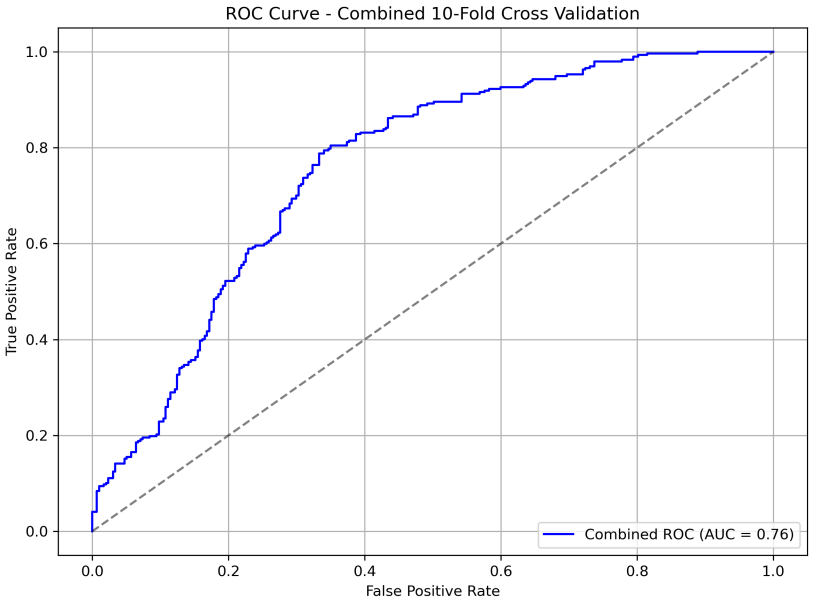


FIGURE 6.10: ROC Curve of combined 10-Fold Cross Validation on BLCA Data

The ROC curve (Figure 6.8) for the final fold-specific GAT evaluation yielded an AUC of 0.74, while the combined ROC from 10-fold cross-validation (Figure 6.10) improved slightly to 0.76. These results show the model's ability to distinguish between BLCA subtypes with reasonably good sensitivity and specificity.

6.2.5 Comparison with Standard Machine Learning Models

For comparison, the performance of the GAT model was contrasted with several baseline models: a feed-forward Artificial Neural Network (ANN), MLP with BatchNorm, and the tree-based ensemble method XGBoost (XGB). These models were selected to provide a diverse set of baselines ranging from linear and non-linear learners to a gradient boosting framework celebrated for high performance on tabular data. All models were trained on the same feature space as the GAT (excluding the graph structure) and evaluated using accuracy and Area Under the ROC Curve (AUC) metrics.

TABLE 6.4: Comparison of Model Performance Before and After Applying SMOTE

Model Name	**Before SMOTE**		**After SMOTE**	
	Accuracy	**AUC**	**Accuracy**	**AUC**
MLP with BatchNorm	0.7377	0.6747	0.4915	0.8046
XGBoost	0.7377	0.6333	0.6780	0.7368
ANN	0.7377	0.5590	0.6780	0.7253
GAT (Proposed)	0.7049	0.5747	0.6536	0.7370

Table 6.4 shows the comparison of the proposed GAT model with other standard machine learning models using metrics such as Accuracy and AUC. Although the Proposed Model gives a good accuracy, the AUC is very less - implying the model is not able to distinguish between classes and blindly predicts the majority class.

Results showed that the GAT model consistently outperformed traditional models, especially in terms of AUC, demonstrating its improved ability to distinguish between BLCA subtypes. The attention mechanism in GAT was likely central to this by allowing the model to modulate the contribution of the neighboring samples, thus modeling complex inter-sample relationships that flat models were unable to capture. These findings validate the use of graph neural networks in multi-omics cancer studies and enhance the relevance of structural learning in biomedical applications.

Chapter 7

Discussion

This chapter discusses the results obtained and provides some biological insights on the datasets.

7.1 Biological Insights

7.1.1 BRCA Dataset

To better understand the characteristics associated with pre-menopause and post-menopause, genes derived from three distinct omics layers were examined, including: copy number alteration (CNA), DNA Methylation and Gene Expression of the BRCA Dataset. From each dataset, biologically relevant or statistically significant genes were selected for further discussion. The following tables summarize selected genes and their known roles in cancer biology.

TABLE 7.1: Selected Genes from the mCNA Dataset and their Biological Relevance

Gene	**Function / Relevance**
RUNX1	Transcription factor involved in hematopoiesis; mutations linked to luminal breast cancer subtypes [82].
FOXA1	Pioneer factor critical for estrogen receptor (ER) signaling in breast cancer [83].
SF3B1	Splicing factor; mutations associated with luminal-type breast cancer and altered transcript diversity [84].
ERBB2	Also known as HER2; frequently amplified in HER2+ breast cancer subtype, driving tumor proliferation [85].
CDKN1B	Encodes p27, a cyclin-dependent kinase inhibitor; low expression linked to poor prognosis in breast cancer [86].

Table 7.2 outlines key genes from the mGE dataset whose expression patterns were selected based on their known functions and relevance to tumor biology.

TABLE 7.2: Biological Relevance of Genes Common to both mGE and mDM Datasets in Breast Cancer

Gene	**Function / Relevance**
NCOR1	Nuclear receptor corepressor; regulates estrogen receptor signaling. High expression is associated with better prognosis and reduced metastasis in breast cancer [87].
CBFB	Transcription factor that suppresses tumor growth. Alterations in CBFB are linked to luminal A subtype and regulation of translation in ER+ cancers [88].
RUNX1	Transcription factor important for hematopoiesis; mutations are enriched in ER+ luminal breast cancers and may act as tumor suppressors [82].
SF3B1	Core component of the spliceosome. Mutations lead to aberrant RNA splicing and are associated with luminal subtypes in breast cancer [84].
MAP2K4	Kinase in the JNK/MAPK signaling pathway; promotes proliferation and is implicated in metastatic progression of breast cancer [89].

7.1.2 BLCA Dataset

To better understand the molecular characteristics associated with TMB, we examined genes derived from three distinct omics layers: mRNA expression, copy number alteration (CNA), and DNA methylation of the BLCA Dataset. From each dataset, biologically relevant or statistically significant genes were selected for further discussion. The following tables summarize selected genes and their known roles in cancer biology.

Table 7.3 outlines key genes from the mRNA dataset whose expression patterns may be linked to high or low TMB. These genes were selected based on their known functions and relevance to tumor biology.

TABLE 7.3: Selected Genes from the mRNA Dataset and their Biological Relevance

Gene	Function / Relevance
MTAP	Tumor suppressor gene, frequently deleted in cancers [90].
DBF4	Cell cycle regulator; upregulated in tumors [91].
NMI	Regulates transcription and immune responses [92].
KCNK5	Associated with breast and ovarian cancers [93].
ZNF683	Implicated in T-cell function and immune signaling [94].
USF1	Involved in gene regulation related to metabolism [95].

Copy number alterations (CNAs) can activate oncogenes or inactivate tumor suppressor genes, contributing to tumor development. Table 7.4 presents a subset of genes from the CNA dataset that are known to participate in processes such as DNA repair, cell cycle control, or immune signaling, highlighting their potential involvement in TMB-associated genomic instability.

TABLE 7.4: Selected Genes from the CNA Dataset and their Biological Relevance

Gene	Function / Relevance
ENSA	Involved in cell cycle regulation; potential link to tumorigenesis [96].
TUFT1	Associated with tumor cell adhesion and metastasis [97].
SELENBP1	Often downregulated in cancers; involved in detoxification [98].
PDLIM2	Regulates NF-KB signaling; acts as a tumor suppressor [99].
HECW1	E3 ubiquitin ligase; involved in protein degradation and cancer regulation [100].
WRN	DNA helicase linked to genome stability; mutations associated with cancer [101].

DNA methylation plays a critical role in the regulation of gene expression and is frequently dysregulated in cancer. From the DNA methylation dataset, several CpG sites were annotated to cancer-related genes. Table 7.5 lists selected genes with well-established or emerging roles in tumorigenesis [102]. These genes were prioritized based on their known biological functions and relevance to cancer pathways.

Table 7.5: Selected Genes from the DNA Methylation Dataset and their Biological Relevance

Gene	Function / Relevance
RARA	Retinoic acid receptor; involved in cell differentiation, frequently altered in cancer [103].
GDF15	Stress response cytokine; biomarker for cancer progression and inflammation [104].
MAP3K6	Kinase involved in apoptosis and signal transduction pathways in cancer [105].
TFPI	Tissue factor pathway inhibitor; regulates blood coagulation, linked to tumor angiogenesis [106].
COL3A1	Collagen gene; plays a role in tumor invasion and extracellular matrix remodeling [107].
EMP2	Involved in cell adhesion and migration; overexpressed in several cancers [108].
GPR64	G-protein-coupled receptor; implicated in cancer cell proliferation and metastasis [109].

7.2 Measuring the Shared Latent Space

The overall objective of multi-omics data integration with representation learning is to project high-dimensional, heterogeneous inputs into a shared informative, biologically relevant latent space. The latent space is supposed to capture important patterns from each omics modality and place them within the same structure such that subsequent tasks such as classification, clustering, or survival prediction are made simpler.

To evaluate the quality of learned latent space, we used a combination of structural, clustering, and task-agnostic metrics. These include trustworthiness to verify local neighborhood preservation and clustering indices to evaluate class separability, reconstruction loss to compute information retained from original modalities, and logistic regression accuracy to evaluate the discriminative power of the embedding. Collectively, these assessments give a complete snapshot of the degree to which the common latent space achieves modality alignment, data reconstruction, and predictive value.

7.2.1 Structural Preservation: Trustworthiness

Trustworthiness quantifies the extent to which the low-dimensional k-nearest neighbors match up with those of the original high-dimensional space. When new neighbors are added that did not exist in the original local neighborhood, the score penalizes them. It can be computed mathematically:

$$T(k) = 1 - \frac{2}{nk(2n - 3k - 1)} \sum_{i=1}^{n} \sum_{j \in U_k^{(i)}} (r(i,j) - k), \tag{7.1}$$

where:

n is the number of data points

k is the number of neighbors considered

$r(i, j)$ is the rank of point j in the ordered list of distances from i in the original space

$U_k^{(i)}$ is the set of points that are among the k nearest neighbors of i in the low-dimensional space but

TABLE 7.6: Trustworthiness of the Shared Latent Space for BRCA and BLCA Datasets

Dataset	**Trustworthiness Score**
BRCA (Breast Cancer)	0.9656
BLCA (Bladder Cancer)	0.9141

The trustworthiness score of both the datasets is presented in Table 7.6. A trustworthiness value of 0.9615 in the BRCA data and 0.9141 in the BLCA data represents reasonable and satisfactory retention of local neighborhood structures in the original input space. This reflects that the GAT-based shared latent space maintains salient topological relationships among patient samples in spite of high-dimensional integration and transformation.

7.2.2 Cluster Separability: Silhouette Score

The Silhouette Score is one of the most widely used metrics to evaluate the cluster separability in any latent or feature space. It quantifies how similar a data point is to its own cluster (cohesion) compared to other clusters (separation). The score ranges from –1 to 1, where scores close to 1 indicate well-separated clusters, and scores close to 0 suggest overlapping or ambiguous groupings. A negative score indicates that samples may be wrongly allocated to a cluster.

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \tag{7.2}$$

where:

$a(i)$ = mean distance from point i to all other points in the same cluster

$b(i)$ = minimum mean distance from point i to points in any other cluster

$s(i) \in [-1, 1]$ is the silhouette score for point i

TABLE 7.7: Silhouette Score Comparison of Shared Latent Spaces

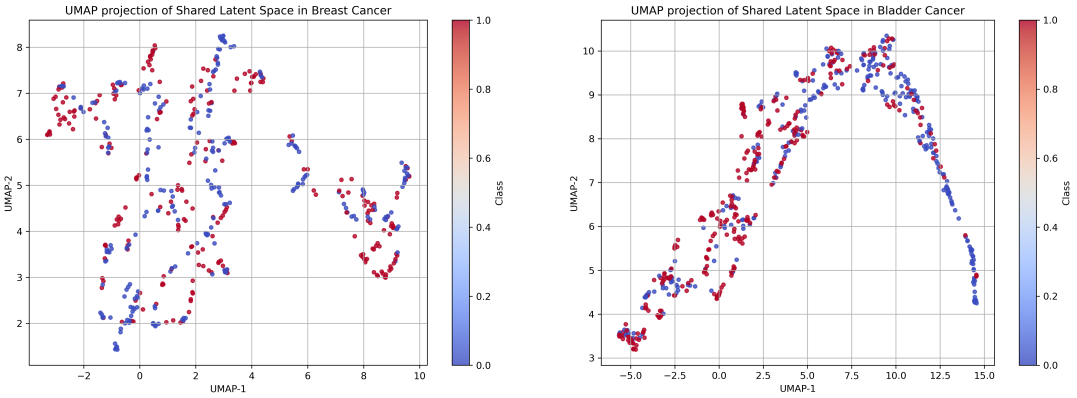
Dataset	**Silhouette Score**
BRCA (Breast Cancer)	0.0329
BLCA (Bladder Cancer)	0.1199

The silhouette score shown above in Table 7.7 was used to assess class separability in the shared latent space. While both datasets showed relatively low values, the BLCA latent space exhibited better-defined clustering (0.1199) compared to BRCA (0.0339). This suggests that the latent features derived from the BLCA dataset capture more distinguishable inter-class patterns, whereas BRCA representations were more overlapped or dispersed.

7.2.3 Visual Analysis by Dimensionality Reduction using UMAP

UMAP (Uniform Manifold Approximation and Projection) [110] is a non-linear dimensionality reduction technique for mapping high-dimensional data to low dimensions (usually 2D or 3D) in a way that preserves the local structure and manifold geometry of the original space. UMAP relies on manifold learning and topological data analysis. It constructs a weighted graph of close points in the high dimension and optimizes a low-dimensional embedding by minimizing the cross-entropy between the two graphs.

UMAP has been a preferred tool for multi-omics [111] latent space visualization due to its speed, scalability, and local and global relationship preservation. Unlike linear methods such as PCA, UMAP excels at revealing cluster structure and class overlaps in high-dimensional biological data. UMAP was used to project the shared latent space to 2D for visual assessment of class separability in this study. The resulting plot indicated partial clustering with visible overlaps between classes, supporting the low clustering metrics and silhouette score, and reflecting the subtle boundaries typical of multi-omics cancer data.



(a) UMAP Projection of BRCA Latent Space

(b) UMAP Projection of BLCA Latent Space

FIGURE 7.1: Comparison of UMAP Projections for Shared Latent Spaces Learned by GAT.

Figure 7.1 gives the UMAP projections of the common latent spaces learned for BRCA and BLCA datasets. In both figures, a visual overlap can be observed between the two classes, reflecting the data distribution's complexity and non-linearity.

This visual overlap is consistent with the quite low silhouette scores attained, suggesting that although the latent space is not really causing easily separable clusters, it preserves important local structures. Particularly, areas of class-wise homogeneity continue to be distinguishable, indicating the ability of the GAT model to learn detailed intra-class structures and inter-class differences. This demonstrates the capability of the latent space to capture biologically important structure, even where classes are mixed.

7.2.4 Reconstruction Loss

Reconstruction loss is a common metric used to evaluate how well a latent space captures essential features of the original input data. In models such as autoencoders or Deep Multiset Canonical Correlation Analysis (DMCCA), the input from each modality is encoded into a shared latent space and then decoded back to reconstruct the original features. The difference between the original input and the reconstructed output is typically measured using Mean Squared Error (MSE), which quantifies the average squared difference between predicted and actual values.

$$\text{Reconstruction Loss (MSE)} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x}_i)^2, \tag{7.3}$$

where:

x_i is the original input feature

$\hat{x}_i$ is the reconstructed feature from the decoder

n is the total number of features

TABLE 7.8: Comparison of Reconstruction Loss across Modalities in BRCA and BLCA

Modality	**BRCA Loss (mDM, mGE, mCNA)**	**BLCA Loss (DNA, RNA, CNA)**
Modality 1	0.1812 (mDM)	0.2300 (DNA)
Modality 2	0.0798 (mGE)	0.0496 (RNA)
Modality 3	0.2488 (mCNA)	0.3073 (CNA)
Average	0.1699	0.1957

The values presented in Table 7.8 indicate that the shared latent space properly preserves important information from all omics modalities, as indicated by the relatively low reconstruction losses. Gene expression (GE) data possesses the lowest reconstruction loss across BRCA and RNA possesses the lowest reconstruction loss in BLCA dataset, suggesting it is most accurately captured in the latent representation. These results also affirm the information-storing ability of the latent space generated by the DMCCA, which is vital for ensuring reliable performance in later downstream classification and integrative modeling.

While the shared latent space exhibits limited clustering capability in unsupervised metrics, its structural integrity (trustworthiness) and information preservation (reconstruction loss) support its overall effectiveness. These findings justify the integration of DMCCA for dimensionality reduction and GAT for learning structured representations from graph data as a hybrid approach for extracting meaningful representations from multi-omics cancer data.

7.3 Explainability and Interpretation of the Shared Latent Space

Interpretability is also a critical aspect of applying deep learning models to biomedical data and especially in multi-omics data integration. Though the latent space constructed in this manner improves classification performance, understanding how and why the

model is making decisions is still a valuable component of its translational potential. The following observations provide explainability as well as different interpretations:

- **Visual inspection using UMAP:** Figure 7.1 provides a 2D UMAP projection of the latent space. While clear class boundaries are not visible, some local clustering patterns are observed in the BLCA dataset. These groupings suggest that the latent space captures minute relationships, even if global class separability (as seen by the low silhouette score of 0.1199 for BLCA) is limited.
- **Attention weights in GAT:** The GAT model introduces sample-level attention mechanisms, which assigns varying importance to neighboring nodes during message passing. These attention coefficients can be extracted and analyzed to identify which nodes (samples) were most influential in a given prediction, thus offering a layer of explainability at the graph level.
- **Reconstruction loss as proxy:** The relatively low reconstruction loss values (e.g., 0.0496 for gene expression in BLCA (Table 7.8) suggest that the latent space retains meaningful structure from the input data. Although this does not provide provide feature-level interpretability, it reinforces the representational fidelity of the shared embedding.

7.4 Limitations

Constructing a shared latent space using various omics for bladder and breast cancer provides sensitive biological insights but also has some limitations:

- **Interpretability:** The DMCCA latent space encodes shared structure and architecture but is not always directly projected onto biological attributes, limiting interpretability.
- **Graph Construction:** GAT uses a k-nearest neighbor graph according to sample similarity, which may be less than maximally reflective of biological association among the samples.
- **Oversampling:** SMOTE addresses class imbalance but may introduce synthetic variation, particularly in high-dimensional omics spaces, which may not be actually present biologically.
- **Dataset Scope:** The model was validated only on BRCA and BLCA datasets and generalizability to other types of cancers is yet to be tested.

7.5 Future Work

- **Biological Graph Construction:** The subsequent work can replace the current k-NN-based graph with biology-informed graphs such as gene regulation or pathway interaction networks.
- **Improving Interpretability:** The application of sparsity constraints or attention-based attribution mechanisms can potentially enhance interpretability of the latent space and separate significant biological drivers.
- **Dataset Expansion:** The use of the proposed pipeline on a broader range of cancer types or pan-cancer datasets would test its generalizability and robustness.
- **Clinical Relevance:** Future research can integrate survival analysis or pathway enrichment methods to validate whether the latent representations are linked to patient outcomes or known biological processes.

Appendix A

Code Snippets

A.1 Upsampling using SMOTE

The code A.1 shows the usage of `SMOTE` that helps in upsampling the minority class. This step is essential to avoid class imbalance and bias before training the model.

```
X_latent = shared_latent.numpy()
X_smote, y_smote = SMOTE(random_state=42).fit_resample(X_latent, y)
```

Listing A.1: Addressing Class Imbalance before training

A.2 Taking Data and Building PyG Object

This code snippet A.2 creates a `Data` object using PyTorch Geometric to represent the graph-structured input required for Graph Neural Network training. The node features (`x`) are derived from the SMOTE-balanced latent space and cast into a `float32` tensor. The graph structure is defined by `edge_index`, which represents the connectivity between nodes, typically constructed using a k-nearest neighbor algorithm. The class labels for each node are stored in `y` as a `long` tensor suitable for classification tasks. Additionally, boolean masks `train_mask` and `test_mask` are defined to indicate which nodes are used for training and evaluation, respectively. These masks enable semi-supervised learning by allowing the model to be trained on a subset of labeled nodes while evaluating on others within the same graph.

```
data = Data(x=torch.tensor(X_smote, dtype=torch.float32),
            edge_index=edge_index,
            y=torch.tensor(y_smote, dtype=torch.long))
data.train_mask = train_mask
```

```
data.test_mask = test_mask
```

Listing A.2: PyG Object Creation

A.3 GAT Model Architecture

The code A.3 shows the `GAT` class that implements a Graph Attention Network using PyTorch and PyTorch Geometric. This model is designed for node-level classification over graph-structured data, such as those derived from multi-omics datasets.

```
class GAT(torch.nn.Module):
    def __init__(self, input_dim, hidden_dim=64, output_dim=2):
        super(GAT, self).__init__()
        self.conv1 = GATConv(input_dim, hidden_dim, heads=2, concat=True)
        self.conv_mid = GATConv(hidden_dim * 2, hidden_dim, heads=2, concat=True)
        self.conv2 = GATConv(hidden_dim * 2, output_dim, heads=1, concat=False)

    def forward(self, data):
        x, edge_index = data.x, data.edge_index

        x = self.conv1(x, edge_index)
        x = F.elu(x)
        x = F.dropout(x, p=0.4, training=self.training)

        x = self.conv_mid(x, edge_index)
        x = F.elu(x)
        x = F.dropout(x, p=0.4, training=self.training)

        x = self.conv2(x, edge_index)
        return x
```

Listing A.3: Graph Attention Network Architecture

The network comprises three attention-based convolutional layers:

- **First Layer** (`conv1`): Applies a GAT convolution with 2 attention heads and concatenates the outputs to increase representational capacity. The input features are projected to a hidden dimension.
- **Intermediate Layer** (`conv_mid`): Further processes the hidden representations using another GATConv layer with 2 heads, again concatenating the outputs.
- **Final Layer** (`conv2`): Outputs class logits using a GATConv with a single head and no concatenation, projecting to the desired output dimension.

Each layer is followed by:

- `ELU` activation to introduce non-linearity.
- `Dropout` (with probability 0.4) to regularize the model during training.

The `forward` function takes a PyG `Data` object as input, containing both node features (`x`) and graph connectivity (`edge_index`), and outputs the final logits for classification. This GAT model enables the learning of context-aware, attention-weighted node representations critical for downstream predictive tasks.

A.4 Model Training

The code A.4 defines the complete training pipeline for the GAT model, incorporating learning rate scheduling, accuracy monitoring, and early stopping. At each epoch, the model is trained on labeled nodes using cross-entropy loss, while the learning rate is adjusted dynamically through a scheduler. The training loss and accuracy are logged for later analysis.

After each training step, the model is evaluated on the test nodes using AUC as the primary metric. If the test AUC improves, the current model state is saved. Training is halted early if no improvement is observed for a set number of epochs, effectively reducing overfitting and computational cost. This loop ensures a stable and well-validated training process.

```
for epoch in range(num_epochs):
    model.train()
    scheduler.step()
    optimizer.zero_grad()
    out = model(data)
    loss = criterion(out[data.train_mask], data.y[data.train_mask])
    loss.backward()
    optimizer.step()
    # === Evaluate on test mask
    model.eval()
    with torch.no_grad():
        logits = model(data)
        probs = F.softmax(logits[data.test_mask], dim=1)[:, 1].cpu().numpy()
        true = data.y[data.test_mask].cpu().numpy()
        auc = roc_auc_score(true, probs)
    preds = out[data.train_mask].argmax(dim=1)
    labels = data.y[data.train_mask]
    acc = (preds == labels).float().mean().item()
```

```
train_losses.append(loss.item())
train_accuracies.append(acc)
print(f"Epoch {epoch}, Loss: {loss.item():.4f}, Test AUC: {auc:.4f}")

# === Early stopping check
if auc > best_auc:
    best_auc = auc
    best_model_state = copy.deepcopy(model.state_dict())
    counter = 0
else:
    counter += 1
    if counter >= patience:
        print(f"Early stopping at epoch {epoch} -  Best Test AUC: {best_auc:.4f}")
        break
```

LISTING A.4: Full Training Loop with Scheduler, Accuracy Tracking, and Early Stopping

A.5 Shared Latent Space

The code snippet A.5 combines qualitative and quantitative analysis of the shared latent space produced by the DMCCA model. First, UMAP is used to project the high-dimensional latent embeddings into two dimensions for visualization. The resulting 2D coordinates are plotted using a scatter plot, colored by class labels. Next, the latent space is evaluated using two unsupervised metrics: the Silhouette Score, which measures intra-class cohesion versus inter-class separation, and Trustworthiness, which assesses the preservation of neighborhood relationships from the original feature space to the reduced space. These metrics help validate that the learned representation captures a meaningful structure across modalities.

```
from sklearn.metrics import silhouette_score
from sklearn.manifold import trustworthiness
import umap
import matplotlib.pyplot as plt

# Project latent space to 2D using UMAP
Z_2d = umap.UMAP(n_components=2, random_state=42).fit_transform(shared_latent.numpy())

# Evaluation Metrics
sil_score = silhouette_score(shared_latent.numpy(), y)
trust = trustworthiness(X_original, shared_latent.numpy(), n_neighbors=10)

print(f"Silhouette Score: {sil_score:.4f}")
```

```
print(f"Trustworthiness: {trust:.4f}")

# Visualization
plt.figure(figsize=(6,5))
plt.scatter(Z_2d[:,0], Z_2d[:,1], c=y, cmap='coolwarm', s=10, alpha=0.8)
plt.title("UMAP Projection of Shared Latent Space")
plt.xlabel("UMAP-1")
plt.ylabel("UMAP-2")
plt.grid(True)
plt.tight_layout()
plt.show()
```

Listing A.5: Shared Latent Space Visualization + Evaluation

A.6 GitHub Link

The complete source code for this thesis is available at:

https://github.com/arvindcb-2023/dmcca-gat-cancer-analysis

Appendix B

Software and Packages Used

This research was implemented using the following software tools, libraries, and computational environments:

Category	**Tools and Libraries (with Citations)**
Programming Language	Python 3.10 [112] – Used for model development, data processing, and pipeline integration.
Deep Learning Framework	PyTorch (v2.0.0) [113] – For constructing and training neural networks.
Graph Neural Networks	PyTorch Geometric (PyG v2.3.1) [114] – For implementing GAT and handling graph data.
Data Manipulation	NumPy [115], Pandas [116] – For numerical computations and dataframe handling.
Machine Learning Utilities	Scikit-learn [117] – For model training, SMOTE, preprocessing, and evaluation metrics.
Data Balancing	Imbalanced-learn (SMOTE) [118] – Applied for oversampling minority class data in the latent space.
Dimensionality Reduction	UMAP-learn [119], t-SNE [120] – For 2D projection and visualization of latent representations.

Continued on next page

Table B.1 – continued from previous page

Category	**Tools and Libraries (with Citations)**
Visualization	Matplotlib [121], Seaborn [122] – For plotting loss curves, accuracy graphs, and UMAP visualizations.
Development Environment	Jupyter Notebook and Google Colab – Used for experimentation, prototyping, and GPU access.
Hardware	NVIDIA GPU with CUDA – Enabled accelerated model training during experiments.

Appendix C

Hyperparameters and Model Configuration

This appendix outlines the key hyperparameters and configuration settings used during model training and evaluation for the proposed DMCCA + GAT framework.

Table C.1: DMCCA Model Configuration

Parameter	**Value**
Input Dimensions	[DNA: 200, RNA: 200, CNA: 200] (post feature selection)
Latent Dimension	64
Hidden Layer Size	128
Dropout Rate	0.2
Activation Function	ReLU
Fusion Strategy	Mean of all latent vectors

Table C.2: GAT Model Configuration for BRCA and BLCA

Parameter	**BRCA**	**BLCA**
Input Dimension	64	64
Hidden Dimension	64	64
Output Dimension	2	2
Heads in Conv1	2	2
Heads in Conv2	2	2
Final GAT Layer	1 head, concat=False	1 head, concat=False
Dropout Rate	0.4	0.4
Activation	ELU	ELU
Loss Function	CrossEntropyLoss	CrossEntropyLoss
Optimizer	Adam	Adam
Learning Rate	0.007	0.004
Scheduler	StepLR (step_size=10, gamma=0.8)	None
Early Stopping Patience	10	10
Epochs	100 max	100 max

Table C.3: Dataset Overview

Dataset	**Omics Types**	**Samples**
BRCA	mDNA, GE, CNA	344
BLCA	mDNA, GE, CNA	404

Bibliography

[1] Chengming Zhang, Yabin Chen, Tao Zeng, Chuanchao Zhang, and Luonan Chen. Deep latent space fusion for adaptive representation of heterogeneous multi-omics data. *Briefings in Bioinformatics*, 23(2):bbab600, 01 2022. ISSN 1477-4054. doi: 10.1093/bib/bbab600. URL https://doi.org/10.1093/bib/bbab600.

[2] Indhupriya Subramanian, Srikant Verma, Shiva Kumar, Abhay Jere, and Krishanpal Anamika. Multi-omics data integration, interpretation, and its application. *Bioinformatics and Biology Insights*, 14:1177932219899051, 2020. doi: 10.1177/1177932219899051. URL https://doi.org/10.1177/1177932219899051. PMID: 32076369.

[3] Yehudit Hasin, Marcus Seldin, and Aldons Lusis. Multi-omics approaches to disease. *Genome Biology*, 18(1):83, May 2017. ISSN 1474-760X. doi: 10.1186/s13059-017-1215-1. URL https://doi.org/10.1186/s13059-017-1215-1.

[4] Sarah B Maron, Lauren A Albacker, Christine A Kowalczyk, et al. Targeted next-generation sequencing in cancer therapy. *Oncology*, 32(7):380–389, 2018.

[5] Beatrix T Joyce, Yong Zheng, Lifang Hou, et al. A review of the role of multi-omics in precision oncology. *Current Opinion in Systems Biology*, 26:100388, 2021. doi: 10.1016/j.coisb.2021.100388.

[6] Yehudit Hasin, Marcus Seldin, and Aldons Lusis. Multi-omics approaches to disease. *Genome Biol.*, 18(1), December 2017.

[7] Rameen Beroukhim, Craig H Mermel, Dale Porter, Guo Wei, Soumya Raychaudhuri, Jerry Donovan, Jordi Barretina, Jesse S Boehm, Jennifer Dobson, Mitsuyoshi Urashima, Kevin T Mc Henry, Reid M Pinchback, Azra H Ligon, Yoon-Jae Cho, Leila Haery, Heidi Greulich, Michael Reich, Wendy Winckler, Michael S Lawrence, Barbara A Weir, Kumiko E Tanaka, Derek Y Chiang, Adam J Bass, Alice Loo, Carter Hoffman, John Prensner, Ted Liefeld, Qing Gao, Derek Yecies, Sabina Signoretti, Elizabeth Maher, Frederic J Kaye, Hidefumi Sasaki, Joel E Tepper, Jonathan A Fletcher, Josep Tabernero, José Baselga, Ming-Sound Tsao, Francesca

Demichelis, Mark A Rubin, Pasi A Janne, Mark J Daly, Carmelo Nucera, Ross L Levine, Benjamin L Ebert, Stacey Gabriel, Anil K Rustgi, Cristina R Antonescu, Marc Ladanyi, Anthony Letai, Levi A Garraway, Massimo Loda, David G Beer, Lawrence D True, Aikou Okamoto, Scott L Pomeroy, Samuel Singer, Todd R Golub, Eric S Lander, Gad Getz, William R Sellers, and Matthew Meyerson. The landscape of somatic copy-number alteration across human cancers. *Nature*, 463 (7283):899–905, February 2010.

[8] Suraiya Rasheed, Jasper S Yan, Adil Hussain, and Bruce Lai. Proteomic characterization of HIV-modulated membrane receptors, kinases and signaling proteins involved in novel angiogenic pathways. *J. Transl. Med.*, 7(1):75, August 2009.

[9] Peter A Jones. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat. Rev. Genet.*, 13(7):484–492, May 2012.

[10] T Sørlie, C M Perou, R Tibshirani, T Aas, S Geisler, H Johnsen, T Hastie, M B Eisen, M van de Rijn, S S Jeffrey, T Thorsen, H Quist, J C Matese, P O Brown, D Botstein, P E Lønning, and A L Børresen-Dale. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl. Acad. Sci. U. S. A.*, 98(19):10869–10874, September 2001.

[11] Katherine A Hoadley, Christina Yau, Toshinori Hinoue, Denise M Wolf, Alexander J Lazar, Esther Drill, Ronglai Shen, Alison M Taylor, Andrew D Cherniack, Vésteinn Thorsson, Rehan Akbani, Reanne Bowlby, Christopher K Wong, Maciej Wiznerowicz, Francisco Sanchez-Vega, A Gordon Robertson, Barbara G Schneider, Michael S Lawrence, Houtan Noushmehr, Tathiane M Malta, Cancer Genome Atlas Network, Joshua M Stuart, Christopher C Benz, and Peter W Laird. Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. *Cell*, 173(2):291–304.e6, April 2018.

[12] Nimrod Rappoport and Ron Shamir. Multi-omic and multi-view clustering algorithms: review and cancer benchmark. *Nucleic Acids Res.*, 46(20):10546–10562, November 2018.

[13] Kriti Chaudhary, Olivier B Poirion, Liang Lu, and Lana X Garmire. Deep learning-based multi-omics integration robustly predicts survival in liver cancer. *Clinical Cancer Research*, 24(6):1248–1259, 2018. doi: 10.1158/1078-0432.CCR-17-0853.

[14] Conghao Wang, Wu Lue, Rama Kaalia, Parvin Kumar, and Jagath C Rajapakse. Network-based integration of multi-omics data for clinical outcome prediction in neuroblastoma. *Sci. Rep.*, 12(1):15425, September 2022.

[15] Jie Zhang. Emerging trends in multi-omics data integration: Challenges and future directions. *Comput. Mol. Biol.*, 2024.

[16] Yuting Yang and Golrokh Mirzaei. Performance analysis of data resampling on class imbalance and classification techniques on multi-omics data for cancer classification. *PLoS One*, 19(2):e0293607, February 2024.

[17] Zahra Momeni, Esmail Hassanzadeh, Mohammad Saniee Abadeh, and Riccardo Bellazzi. A survey on single and multi omics data mining methods in cancer data classification. *J. Biomed. Inform.*, 107(103466):103466, July 2020.

[18] Xiao Li, Jie Ma, Ling Leng, Mingfei Han, Mansheng Li, Fuchu He, and Yunping Zhu. Mogcn: A multi-omics integration method based on graph convolutional network for cancer subtype analysis. *Frontiers in Genetics*, 13, 2022. ISSN 1664-8021. doi: 10.3389/fgene.2022.806842. URL https://www.frontiersin.org/journals/genetics/articles/10.3389/fgene.2022.806842.

[19] S M Vidanagamachchi and K M G T R Waidyarathna. Opportunities, challenges and future perspectives of using bioinformatics and artificial intelligence techniques on tropical disease identification using omics data. *Front. Digit. Health*, 6:1471200, November 2024.

[20] Javier E Flores, Daniel M Claborne, Zachary D Weller, Bobbie-Jo M Webb-Robertson, Katrina M Waters, and Lisa M Bramer. Missing data in multi-omics integration: Recent advances through artificial intelligence. *Front. Artif. Intell.*, 6:1098308, February 2023.

[21] Madhumita and Sushmita Paul. Capturing the latent space of an autoencoder for multi-omics integration and cancer subtyping. *Comput. Biol. Med.*, 148(105832): 105832, September 2022.

[22] S M Vidanagamachchi and K M G T R Waidyarathna. Opportunities, challenges and future perspectives of using bioinformatics and artificial intelligence techniques on tropical disease identification using omics data. *Front. Digit. Health*, 6:1471200, November 2024.

[23] Douglas Hanahan and Robert A. Weinberg. Hallmarks of cancer: The next generation. *Cell*, 144(5):646–674, Mar 2011. ISSN 0092-8674. doi: 10.1016/j.cell.2011.02.013. URL https://doi.org/10.1016/j.cell.2011.02.013.

[24] Abedalrhman Alkhateeb and Luis Rueda, editors. *Machine Learning Methods for Multi-Omics Data Integration*. Springer International Publishing, Cham, 2024. ISBN 978-3-031-36502-7. doi: 10.1007/978-3-031-36502-7. URL https://link.springer.com/book/10.1007/978-3-031-36502-7.

[25] Aatish Thennavan, Francisco Beca, Youli Xia, Susana Garcia-Recio, Kimberly Allison, Laura C. Collins, Gary M. Tse, Yunn-Yi Chen, Stuart J. Schnitt, Katherine A. Hoadley, Andrew Beck, and Charles M. Perou. Molecular analysis of tcga breast cancer histologic types. *Cell Genomics*, 1(3), Dec 2021. ISSN 2666-979X. doi: 10.1016/j.xgen.2021.100067. URL https://doi.org/10.1016/j.xgen.2021.100067.

[26] Bernard Fisher, Stewart Anderson, John Bryant, Richard G Margolese, Melvin Deutsch, Edwin R Fisher, Jong-Hyeon Jeong, and Norman Wolmark. Twenty-year follow-up of a randomized trial comparing total mastectomy, lumpectomy, and lumpectomy plus irradiation for the treatment of invasive breast cancer. *N. Engl. J. Med.*, 347(16):1233–1241, October 2002.

[27] M Clarke, R Collins, S Darby, C Davies, P Elphinstone, V Evans, J Godwin, R Gray, C Hicks, S James, E MacKinnon, P McGale, T McHugh, R Peto, C Taylor, Y Wang, and Early Breast Cancer Trialists' Collaborative Group (EBCTCG). Effects of radiotherapy and of differences in the extent of surgery for early breast cancer on local recurrence and 15-year survival: an overview of the randomised trials. *Lancet*, 366(9503):2087–2106, December 2005.

[28] Christina Davies, Hongchao Pan, Jon Godwin, Richard Gray, Rodrigo Arriagada, Vinod Raina, Mirta Abraham, Victor Hugo Medeiros Alencar, Atef Badran, Xavier Bonfill, Joan Bradbury, Michael Clarke, Rory Collins, Susan R Davis, Antonella Delmestri, John F Forbes, Peiman Haddad, Ming-Feng Hou, Moshe Inbar, Hussein Khaled, Joanna Kielanowska, Wing-Hong Kwan, Beela S Mathew, Indraneel Mittra, Bettina Müller, Antonio Nicolucci, Octavio Peralta, Fany Pernas, Lubos Petruzelka, Tadeusz Pienkowski, Ramachandran Radhika, Balakrishnan Rajan, Maryna T Rubach, Sera Tort, Gerard Urrútia, Miriam Valentini, Yaochen Wang, and Richard Peto. Long-term effects of continuing adjuvant tamoxifen to 10 years versus stopping at 5 years after diagnosis of oestrogen receptor-positive breast cancer: ATLAS, a randomised trial. *Lancet*, 381(9869):805–816, March 2013.

[29] Adrienne G. Waks and Eric P. Winer. Breast cancer treatment: A review. *JAMA*, 321(3):288–300, 01 2019. ISSN 0098-7484. doi: 10.1001/jama.2018.19323. URL https://doi.org/10.1001/jama.2018.19323.

[30] Kalyan Saginala, Adam Barsouk, John Sukumar Aluru, Prashanth Rawla, Sandeep Anand Padala, and Alexander Barsouk. Epidemiology of bladder cancer. *Medical Sciences*, 8(1), 2020. ISSN 2076-3271. doi: 10.3390/medsci8010015. URL https://www.mdpi.com/2076-3271/8/1/15.

[31] Ashish M Kamat, Noah M Hahn, Jason A Efstathiou, Seth P Lerner, Per-Uno Malmström, Woonyoung Choi, Charles C Guo, Yair Lotan, and Wassim Kassouf. Bladder cancer. *Lancet*, 388(10061):2796–2810, December 2016.

[32] Carey K Anders and Roy Johnson. Breast cancer before age 40 years. *Seminars in Oncology*, 36(3):237–249, 2009. doi: 10.1053/j.seminoncol.2009.03.001.

[33] A. J. Vincent and. Management of menopause in women with breast cancer. *Climacteric*, 18(5):690–701, 2015. doi: 10.3109/13697137.2014.996749. URL https://doi.org/10.3109/13697137.2014.996749. PMID: 25536007.

[34] Hatem A Azim Jr and Ann H Partridge. Biology of breast cancer in young women. *Breast Cancer Research*, 14(4):212, 2012. doi: 10.1186/bcr3133.

[35] Aron Goldhirsch, Eric P Winer, Alan S Coates, Richard D Gelber, Martine Piccart-Gebhart, Beat Thürlimann, and Hans-Joerg Senn. Personalizing the treatment of women with early breast cancer: highlights of the st gallen international expert consensus on the primary therapy of early breast cancer 2013. *Annals of Oncology*, 24(9):2206–2223, 2013. doi: 10.1093/annonc/mdt303.

[36] Timothy A Chan, Mark Yarchoan, Elizabeth Jaffee, Charles Swanton, Sergio A Quezada, Albrecht Stenzinger, and Solange Peters. Development of tumor mutation burden as an immunotherapy biomarker: utility for the oncology clinic. *Annals of Oncology*, 30(1):44–56, 2019. doi: 10.1093/annonc/mdy495.

[37] Sanjeev et al. Mariathasan. Tgfβ attenuates tumour response to pd-l1 blockade by contributing to exclusion of t cells. *Nature*, 554(7693):544–548, 2018. doi: 10.1038/nature25501.

[38] Razvan Cristescu, Robert Mogg, Mark Ayers, Anna Albright, Elisabeth Murphy, Julia Yearley, Xuan Sher, Xiaoping Liu, Hao Lu, Michael Nebozhyn, et al. Pan-tumor genomic biomarkers for pd-1 checkpoint blockade–based immunotherapy. *Science*, 362(6411), 2018. doi: 10.1126/science.aar3593.

[39] Robert M Samstein, Chung-Han Lee, Alexander N Shoushtari, Matthew D Hellmann, Ronglai Shen, Yelena Y Janjigian, David A Barron, Ahmet Zehir, Eric J Jordan, Antonio Omuro, et al. Tumor mutational load predicts survival after immunotherapy across multiple cancer types. *Nature Genetics*, 51(2):202–206, 2019. doi: 10.1038/s41588-018-0312-8.

[40] Gordon AG et al. Robertson. Comprehensive molecular characterization of muscle-invasive bladder cancer. *Cell*, 171(3):540–556.e25, 2017. doi: 10.1016/j.cell.2017.09.007.

[41] Nasser M. Nasrabadi. Pattern Recognition and Machine Learning. *Journal of Electronic Imaging*, 16(4):049901, 2007. doi: 10.1117/1.2819119. URL https://doi.org/10.1117/1.2819119.

[42] Maxwell W. Libbrecht and William Stafford Noble. Machine learning applications in genetics and genomics. *Nature Reviews Genetics*, 16(6):321–332, Jun 2015. ISSN 1471-0064. doi: 10.1038/nrg3920. URL https://doi.org/10.1038/nrg3920.

[43] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks, 2017. URL https://arxiv.org/abs/1609.02907.

[44] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and Philip S. Yu. A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 32(1):4–24, 2021. doi: 10.1109/TNNLS.2020.2978386.

[45] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks, 2018. URL https://arxiv.org/abs/1710.10903.

[46] Stavros Makrodimitris, Bram Pronk, Tamim Abdelaal, and Marcel Reinders. An in-depth comparison of linear and non-linear joint embedding methods for bulk and single-cell multi-omics. *Briefings in Bioinformatics*, 25(1):bbad416, 11 2023. ISSN 1477-4054. doi: 10.1093/bib/bbad416. URL https://doi.org/10.1093/bib/bbad416.

[47] Chen Meng, Bernhard Kuster, Aedín C. Culhane, and Amin Moghaddas Gholami. A multivariate approach to the integration of multi-omics datasets. *BMC Bioinformatics*, 15(1):162, May 2014. ISSN 1471-2105. doi: 10.1186/1471-2105-15-162. URL https://doi.org/10.1186/1471-2105-15-162.

[48] Benjamin Kompa and Beau Coker. Learning a latent space of highly multidimensional cancer data. *Pac Symp Biocomput*, 25:379–390, 2020.

[49] Daniel Lepe-Soltero, Thierry Artières, Anaïs Baudot, and Paul Villoutreix. Modis: Multi-omics data integration for small and unpaired datasets, 2025. URL https://arxiv.org/abs/2503.18856.

[50] Muta Tah Hira, M A Razzaque, Claudio Angione, James Scrivens, Saladin Sawan, and Mosharraf Sarker. Integrated multi-omics analysis of ovarian cancer using variational autoencoders. *Sci Rep*, 11(1):6265, March 2021.

[51] Nektarios Valous, Ferdinand Popp, Inka Zörnig, Dirk Jäger, and Pornpimol Charoentong. Graph machine learning for integrated multi-omics analysis. *British Journal of Cancer*, 131:205–211, 05 2024. doi: 10.1038/s41416-024-02706-7.

[52] Tongxin Wang, Wei Shao, Zhi Huang, Haixu Tang, Jie Zhang, Zhengming Ding, and Kun Huang. Mogonet integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification. *Nature Communications*, 12(1):3445, Jun 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-23774-w. URL https://doi.org/10.1038/s41467-021-23774-w.

[53] Ziynet Nesibe Kesimoglu and Serdar Bozdag. SUPREME: multiomics data integration using graph convolutional networks. *NAR Genom. Bioinform.*, 5(2): lqad063, June 2023.

[54] Hryhorii Chereda, Annalen Bleckmann, Kerstin Menck, Júlia Perera-Bel, Philip Stegmaier, Florian Auer, Frank Kramer, Andreas Leha, and Tim Beißbarth. Explaining decisions of graph convolutional neural networks: patient-specific molecular subnetworks responsible for metastasis prediction in breast cancer. *Genome Med.*, 13(1):42, March 2021.

[55] Raihanul Bari Tanvir, Md Mezbahul Islam, Masrur Sobhan, Dongsheng Luo, and Ananda Mohan Mondal. MOGAT: A multi-omics integration framework using graph attention networks for cancer subtype prediction. *Int. J. Mol. Sci.*, 25(5): 2788, February 2024.

[56] Sina Tabakhi, Charlotte Vandermeulen, Ian Sudbery, and Haiping Lu. Heterogeneous graph attention network improves cancer multiomics integration. 2024.

[57] Megan Hoi Yan Fong, Mingxiao Feng, David J McConkey, and Woonyoung Choi. Update on bladder cancer molecular subtypes. *Transl. Androl. Urol.*, 9(6):2881–2889, December 2020.

[58] Ying Sha, John H Phan, and May D Wang. Effect of low-expression gene filtering on detection of differentially expressed genes in RNA-seq data. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, August 2015.

[59] Hengrui Liu, Yiying Li, Miray Karsidag, Tiffany Tu, and Panpan Wang. Technical and biological biases in bulk transcriptomic data mining for cancer research. *J. Cancer*, 16(1):34–43, January 2025.

[60] M Kanehisa and S Goto. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, 28(1):27–30, January 2000.

[61] Minoru Kanehisa, Susumu Goto, Yoko Sato, Miho Furumichi, and Mao Tanabe. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.*, 40(Database issue):D109–14, January 2012.

[62] Chunyu Wang, Junling Guo, Ning Zhao, Yang Liu, Xiaoyan Liu, Guojun Liu, and Maozu Guo. A cancer survival prediction method based on graph convolutional network. *IEEE Transactions on NanoBioscience*, 19(1):117–126, 2020. doi: 10.1109/TNB.2019.2936398.

[63] Zhaoxiang Cai, Rebecca C. Poulos, Jia Liu, and Qing Zhong. Machine learning for multi-omics data integration in cancer. *iScience*, 25(2), Feb 2022. ISSN 2589-0042. doi: 10.1016/j.isci.2022.103798. URL https://doi.org/10.1016/j.isci.2022.103798.

[64] Parminder S Reel, Smarti Reel, Ewan Pearson, Emanuele Trucco, and Emily Jefferson. Using machine learning approaches for multi-omics data analysis: A review. *Biotechnol Adv*, 49:107739, March 2021.

[65] Bingjun Li, Tianyu Wang, and Sheida Nabavi. Cancer molecular subtype classification by graph convolutional networks on multi-omics data. In *Proceedings of the 12th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, BCB '21, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384506. doi: 10.1145/3459930.3469542. URL https://doi.org/10.1145/3459930.3469542.

[66] Baoshan Ma, Fanyu Meng, Ge Yan, Haowen Yan, Bingjie Chai, and Fengju Song. Diagnostic classification of cancers using extreme gradient boosting algorithm and multi-omics data. *Computers in Biology and Medicine*, 121:103761, 2020. ISSN 0010-4825. doi: https://doi.org/10.1016/j.compbiomed.2020.103761. URL https://www.sciencedirect.com/science/article/pii/S0010482520301360.

[67] Yi Wang, Zhongyue Zhang, Hua Chai, and Yuedong Yang. Multi-omics cancer prognosis analysis based on graph convolution network. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1564–1568, 2021. doi: 10.1109/BIBM52615.2021.9669797.

[68] Hongli Gao, Bin Zhang, Long Liu, Shan Li, Xin Gao, and Bin Yu. A universal framework for single-cell multi-omics data integration with graph convolutional networks. *Briefings in Bioinformatics*, 24(3):bbad081, 03 2023. ISSN 1477-4054. doi: 10.1093/bib/bbad081. URL https://doi.org/10.1093/bib/bbad081.

[69] Giovanni Ciriello, Michael L Gatza, Andrew H Beck, Matthew D Wilkerson, Suhn K Rhie, Alessandro Pastore, Hailei Zhang, Michael McLellan, Christina Yau, Cyriac Kandoth, Reanne Bowlby, Hui Shen, Sikander Hayat, Robert Fieldhouse, Susan C Lester, Gary M K Tse, Rachel E Factor, Laura C Collins, Kimberly H Allison, Yunn-Yi Chen, Kristin Jensen, Nicole B Johnson, Steffi Oesterreich, Gordon B Mills, Andrew D Cherniack, Gordon Robertson, Christopher Benz, Chris

Sander, Peter W Laird, Katherine A Hoadley, Tari A King, TCGA Research Network, and Charles M Perou. Comprehensive molecular portraits of invasive lobular breast cancer. *Cell*, 163(2):506–519, October 2015.

[70] Ethan Cerami, Jianjiong Gao, Ugur Dogrusoz, Benjamin E Gross, Selcuk Onur Sumer, Bülent Arman Aksoy, Anders Jacobsen, Caitlin J Byrne, Michael L Heuer, Erik Larsson, Yevgeniy Antipin, Boris Reva, Arthur P Goldberg, Chris Sander, and Nikolaus Schultz. The cbio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.*, 2(5):401–404, May 2012.

[71] Jianjiong Gao, Bülent Arman Aksoy, Ugur Dogrusoz, Gideon Dresdner, Benjamin Gross, S Onur Sumer, Yichao Sun, Anders Jacobsen, Rileen Sinha, Erik Larsson, Ethan Cerami, Chris Sander, and Nikolaus Schultz. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.*, 6(269):l1, April 2013.

[72] Ino de Bruijn, Ritika Kundra, Brooke Mastrogiacomo, Thinh Ngoc Tran, Luke Sikina, Tali Mazor, Xiang Li, Angelica Ochoa, Gaofei Zhao, Bryan Lai, Adam Abeshouse, Diana Baiceanu, Ersin Ciftci, Ugur Dogrusoz, Andrew Dufilie, Ziya Erkoc, Elena Garcia Lara, Zhaoyuan Fu, Benjamin Gross, Charles Haynes, Allison Heath, David Higgins, Prasanna Jagannathan, Karthik Kalletla, Priti Kumari, James Lindsay, Aaron Lisman, Bas Leenknegt, Pieter Lukasse, Divya Madela, Ramyasree Madupuri, Pim van Nierop, Oleguer Plantalech, Joyce Quach, Adam C Resnick, Sander Y A Rodenburg, Baby A Satravada, Fedde Schaeffer, Robert Sheridan, Jessica Singh, Rajat Sirohi, Selcuk Onur Sumer, Sjoerd van Hagen, Avery Wang, Manda Wilson, Hongxin Zhang, Kelsey Zhu, Nicole Rusk, Samantha Brown, Jessica A Lavery, Katherine S Panageas, Julia E Rudolph, Michele L LeNoue-Newton, Jeremy L Warner, Xindi Guo, Haley Hunter-Zinck, Thomas V Yu, Shirin Pilai, Chelsea Nichols, Stuart M Gardos, John Philip, AACR Project GENIE BPC Core Team, AACR Project GENIE Consortium, Kenneth L Kehl, Gregory J Riely, Deborah Schrag, Jocelyn Lee, Michael V Fiandalo, Shawn M Sweeney, Trevor J Pugh, Chris Sander, Ethan Cerami, Jianjiong Gao, and Nikolaus Schultz. Analysis and visualization of longitudinal genomic and clinical data from the AACR project GENIE biopharma collaborative in cBioPortal. *Cancer Res.*, 83(23):3861–3867, December 2023.

[73] Cancer Genome Atlas Research Network. Comprehensive molecular characterization of urothelial bladder carcinoma. *Nature*, 507(7492):315–322, January 2014.

[74] Alberto Fernández, Salvador García, Francisco Herrera, and Nitesh V. Chawla. SMOTE for learning from imbalanced data: Progress and challenges, marking

the 15-year anniversary. *Journal of Artificial Intelligence Research*, 61:863–905, 2018. doi: 10.1613/jair.1.11192. URL https://www.jair.org/index.php/jair/article/view/11192.

[75] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002. doi: 10.1613/jair.953. URL https://www.jair.org/index.php/jair/article/view/10302.

[76] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289*, 2015. URL https://arxiv.org/abs/1511.07289.

[77] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library, 2019. URL https://arxiv.org/abs/1912.01703.

[78] Matthias Fey and Jan Eric Lenssen. Fast Graph Representation Learning with PyTorch Geometric, May 2019. URL https://github.com/pyg-team/pytorch_geometric.

[79] Antonio Mucherino, Petraq J. Papajorgji, and Panos M. Pardalos. *k-Nearest Neighbor Classification*, pages 83–106. Springer New York, New York, NY, 2009. ISBN 978-0-387-88615-2. doi: 10.1007/978-0-387-88615-2_4. URL https://doi.org/10.1007/978-0-387-88615-2_4.

[80] Leo Liberti, Carlile Lavor, Nelson Maculan, and Antonio Mucherino. Euclidean distance geometry and applications, 2012. URL https://arxiv.org/abs/1205.0349.

[81] James A Hanley and Barbara J McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, 1982. doi: 10.1148/radiology.143.1.7063747.

[82] MedlinePlus Genetics. Runx1 gene. https://medlineplus.gov/genetics/gene/runx1/, 2021. URL https://medlineplus.gov/genetics/gene/runx1/. Accessed: 2025-04-20.

[83] GeneCards Human Gene Database. Foxa1 gene - genecards — foxa1 protein — foxa1 antibody. https://www.genecards.org/cgi-bin/carddisp.pl?gene=

FOXA1, 2025. URL https://www.genecards.org/cgi-bin/carddisp.pl?gene=FOXA1. Accessed: 2025-04-20.

[84] National Center for Biotechnology Information. Sf3b1 splicing factor 3b subunit 1 [homo sapiens (human)] - gene - ncbi. https://www.ncbi.nlm.nih.gov/gene/23451, 2025. URL https://www.ncbi.nlm.nih.gov/gene/23451. Accessed: 2025-04-20.

[85] National Center for Biotechnology Information. Erbb2 erb-b2 receptor tyrosine kinase 2 [homo sapiens (human)] - gene - ncbi. https://www.ncbi.nlm.nih.gov/gene/2064, 2025. URL https://www.ncbi.nlm.nih.gov/gene/2064. Accessed: 2025-04-20.

[86] MedlinePlus Genetics. Cdkn1b gene. https://medlineplus.gov/genetics/gene/cdkn1b/, 2021. URL https://medlineplus.gov/genetics/gene/cdkn1b/. Accessed: 2025-04-20.

[87] National Center for Biotechnology Information. Ncor1 nuclear receptor corepressor 1 [homo sapiens (human)] - gene - ncbi. https://www.ncbi.nlm.nih.gov/gene/9611, 2025. URL https://www.ncbi.nlm.nih.gov/gene/9611. Accessed: 2025-04-20.

[88] Wikipedia contributors. Cbfb. https://en.wikipedia.org/wiki/CBFB, 2025. URL https://en.wikipedia.org/wiki/CBFB. Accessed: 2025-04-20.

[89] National Center for Biotechnology Information. Map2k4 mitogen-activated protein kinase kinase 4 [homo sapiens (human)] - gene - ncbi. https://www.ncbi.nlm.nih.gov/gene/6416, 2025. URL https://www.ncbi.nlm.nih.gov/gene/6416. Accessed: 2025-04-20.

[90] GeneCards. Mtap gene - genecards — mtap protein — mtap antibody, 2024. URL https://www.genecards.org/cgi-bin/carddisp.pl?gene=MTAP. Accessed: 2025-04-01.

[91] The UniProt Consortium. Uniprot: the universal protein knowledgebase. *Nucleic Acids Research*, 53(D1):D480–D488, 2025. doi: 10.1093/nar/gkaa1100. URL https://www.uniprot.org/uniprotkb/Q9UBU7/entry.

[92] Wikipedia contributors. N-myc-interactor — Wikipedia, The Free Encyclopedia, 2025. URL https://en.wikipedia.org/wiki/N-myc-interactor. [Online; accessed 1-April-2025].

[93] Wikipedia contributors. KCNK5 — Wikipedia, The Free Encyclopedia, 2025. URL https://en.wikipedia.org/wiki/KCNK5. [Online; accessed 1-April-2025].

[94] GeneCards Team. ZNF683 Gene — GeneCards, 2025. URL https://www.genecards.org/cgi-bin/carddisp.pl?gene=ZNF683. [Online; accessed 2-April-2025].

[95] Yue-Mei Fan, Jussi Hernesniemi, Niku Oksala, Mari Levula, Emma Raitoharju, Auni Collings, Nina Hutri-Kähönen, Markus Juonala, Jukka Marniemi, Leo-Pekka Lyytikäinen, Ilkka Seppälä, Ari Mennander, Matti Tarkka, Antti J Kangas, Pasi Soininen, Juha Pekka Salenius, Norman Klopp, Thomas Illig, Tomi Laitinen, Mika Ala-Korpela, Reijo Laaksonen, Jorma Viikari, Mika Kähönen, Olli T Raitakari, and Terho Lehtimäki. Upstream transcription factor 1 (USF1) allelic variants regulate lipoprotein metabolism in women and USF1 expression in atherosclerotic plaque. *Sci. Rep.*, 4(1):4650, April 2014.

[96] GeneCards. Ensa gene - genecards — ensa protein — ensa antibody, 2025. URL https://www.genecards.org/cgi-bin/carddisp.pl?gene=ENSA. Accessed: 2025-04-02.

[97] GeneCards. Tuft1 gene - genecards — tuft1 protein — tuft1 antibody, 2025. URL https://www.genecards.org/cgi-bin/carddisp.pl?gene=TUFT1. Accessed: 2025-04-02.

[98] GeneCards. Selenbp1 gene - selenium binding protein 1 - genecards, 2025. URL https://www.genecards.org/cgi-bin/carddisp.pl?gene=SELENBP1. Accessed: 2025-04-02.

[99] GeneCards. Pdlim2 gene - pdli2 protein - genecards, 2025. URL https://www.genecards.org/cgi-bin/carddisp.pl?gene=PDLIM2. Accessed: 2025-04-02.

[100] GeneCards. Hecw1 gene - genecards — hecw1 protein — hecw1 antibody, 2025. URL https://www.genecards.org/cgi-bin/carddisp.pl?gene=HECW1. Accessed: 2025-04-02.

[101] GeneCards. Wrn gene - genecards — wrn protein — wrn antibody, 2025. URL https://www.genecards.org/cgi-bin/carddisp.pl?gene=WRN. Accessed: 2025-04-02.

[102] Hai Yang, Lipeng Gan, Rui Chen, Dongdong Li, Jing Zhang, and Zhe Wang. From multi-omics data to the cancer druggable gene discovery: a novel machine learning-based approach. *Briefings in Bioinformatics*, 24(1):bbac528, 12 2022. ISSN 1477-4054. doi: 10.1093/bib/bbac528. URL https://doi.org/10.1093/bib/bbac528.

[103] MedlinePlus Genetics. Rara gene, 2025. URL https://medlineplus.gov/genetics/gene/rara/. Accessed: 2025-04-02.

[104] NCBI Gene. Gdf15 growth differentiation factor 15 [(human)], 2025. URL https://www.ncbi.nlm.nih.gov/gene/9518. Accessed: 2025-04-02.

[105] GeneCards. Map3k6 gene - genecards — m3k6 protein, 2025. URL https://www.genecards.org/cgi-bin/carddisp.pl?gene=MAP3K6. Accessed: 2025-04-02.

[106] Wikipedia. Tissue factor pathway inhibitor, 2025. URL https://en.wikipedia.org/wiki/Tissue_factor_pathway_inhibitor. Accessed: 2025-04-02.

[107] MedlinePlus Genetics. Col3a1 gene, 2025. URL https://medlineplus.gov/genetics/gene/col3a1/. Accessed: 2025-04-02.

[108] GeneCards. Emp2 gene - epithelial membrane protein 2, 2025. URL https://www.genecards.org/cgi-bin/carddisp.pl?gene=EMP2. Accessed: 2025-04-02.

[109] Wikipedia. Gpr64, 2025. URL https://en.wikipedia.org/wiki/GPR64. Accessed: 2025-04-02.

[110] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861, 2018. doi: 10.21105/joss.00861. URL https://doi.org/10.21105/joss.00861.

[111] Bashier ElKarami, Abedalrhman Alkhateeb, Hazem Qattous, Lujain Alshomali, and Behnam Shahrrava. Multi-omics data integration model based on umap embedding and convolutional neural network. *Cancer Informatics*, 21: 11769351221124205, 2022. doi: 10.1177/11769351221124205. URL https://doi.org/10.1177/11769351221124205.

[112] Guido Van Rossum and Fred L Drake. *Python 3 Reference Manual*. CreateSpace, 2009.

[113] Adam Paszke, Sam Gross, Francisco Massa, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[114] Matthias Fey and Jan Eric Lenssen. Fast graph representation learning with pytorch geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*, 2019.

[115] Charles R Harris, K Jarrod Millman, Stéfan J van der Walt, et al. Array programming with numpy. *Nature*, 585(7825):357–362, 2020.

[116] Wes McKinney. Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, volume 445, pages 51–56, 2010.

[117] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, et al. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[118] Guillaume Lemaître, Fernando Nogueira, and Christos K Aridas. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research*, 18(17):1–5, 2017.

[119] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[120] Laurens Van der Maaten. Learning a parametric embedding by preserving local structure. In *Artificial Intelligence and Statistics*, pages 384–391, 2009.

[121] John D Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007.

[122] Michael L Waskom. Seaborn: Statistical data visualization. *Journal of Open Source Software*, 6(60):3021, 2021.