TYPE Review
PUBLISHED 25 July 2024
DOI 10.3389/frai.2024.1408843

Check for updates

OPEN ACCESS

EDITED BY
Francesco Napolitano,
University of Sannio, Italy

REVIEWED BY
Sunyoung Jang,
The Pennsylvania State University,
United States
Muhammad Shahid Iqbal,
Anhui University, China

*CORRESPONDENCE
Asim Waqas
✉ asim.waqas@moffitt.org

†These authors have contributed equally to this work and share first authorship

RECEIVED 28 March 2024
ACCEPTED 09 July 2024
PUBLISHED 25 July 2024

CITATION
Waqas A, Tripathi A, Ramachandran RP, Stewart PA and Rasool G (2024) Multimodal data integration for oncology in the era of deep neural networks: a review. *Front. Artif. Intell.* 7:1408843. doi: 10.3389/frai.2024.1408843

Multimodal data integration for oncology in the era of deep neural networks: a review

Asim Waqas[1,2*†], Aakash Tripathi[1†], Ravi P. Ramachandran[3], Paul A. Stewart[4] and Ghulam Rasool[1]

[1]Department of Machine Learning, Moffitt Cancer Center, Tampa, FL, United States, [2]Department of Cancer Epidemiology, Moffitt Cancer Center, Tampa, FL, United States, [3]Department of Electrical and Computer Engineering, Rowan University, Glassboro, NJ, United States, [4]Department of Biostatistics and Bioinformatics, Moffitt Cancer Center, Tampa, FL, United States

Cancer research encompasses data across various scales, modalities, and resolutions, from screening and diagnostic imaging to digitized histopathology slides to various types of molecular data and clinical records. The integration of these diverse data types for personalized cancer care and predictive modeling holds the promise of enhancing the accuracy and reliability of cancer screening, diagnosis, and treatment. Traditional analytical methods, which often focus on isolated or unimodal information, fall short of capturing the complex and heterogeneous nature of cancer data. The advent of deep neural networks has spurred the development of sophisticated multimodal data fusion techniques capable of extracting and synthesizing information from disparate sources. Among these, Graph Neural Networks (GNNs) and Transformers have emerged as powerful tools for multimodal learning, demonstrating significant success. This review presents the foundational principles of multimodal learning including oncology data modalities, taxonomy of multimodal learning, and fusion strategies. We delve into the recent advancements in GNNs and Transformers for the fusion of multimodal data in oncology, spotlighting key studies and their pivotal findings. We discuss the unique challenges of multimodal learning, such as data heterogeneity and integration complexities, alongside the opportunities it presents for a more nuanced and comprehensive understanding of cancer. Finally, we present some of the latest comprehensive multimodal pan-cancer data sources. By surveying the landscape of multimodal data integration in oncology, our goal is to underline the transformative potential of multimodal GNNs and Transformers. Through technological advancements and the methodological innovations presented in this review, we aim to chart a course for future research in this promising field. This review may be the first that highlights the current state of multimodal modeling applications in cancer using GNNs and transformers, presents comprehensive multimodal oncology data sources, and sets the stage for multimodal evolution, encouraging further exploration and development in personalized cancer care.

KEYWORDS

multimodal, graph neural networks, transformers, oncology, deep learning, cancer, multi-omics, machine learning

1 Introduction

Cancer represents a significant global health challenge, characterized by the uncontrolled growth of abnormal cells, leading to millions of deaths annually. In 2023, the United States had around 1.9 million new cancer diagnoses, with cancer being the second leading cause of death and anticipated to result in approximately 1670 deaths daily (Siegel et al., 2023). However, advancements in oncology research hold the promise of preventing nearly 42% of these cases through early detection and lifestyle modifications. The complexity of cancer, involving intricate changes at both the microscopic and macroscopic levels, requires innovative approaches to its understanding and management. In recent years, the application of machine learning (ML) techniques, especially deep learning (DL), has emerged as a transformative force in oncology. DL employs deep neural networks to analyze vast datasets, offering unprecedented insights into cancer's development and progression (Çalışkan and Tazaki, 2023; Chen et al., 2023; Siam et al., 2023; Muhammad et al., 2024; Talebi et al., 2024). This approach has led to the development of computer-aided diagnostic systems capable of detecting and classifying cancerous tissues in medical images, such as mammograms and MRI scans, with increasing accuracy. Beyond imaging, DL also plays a crucial role in analyzing molecular data, aiding in the prediction of treatment responses, and the identification of new biomarkers (Dera et al., 2019, 2021; Waqas et al., 2021; Barhoumi et al., 2023; Khan et al., 2023; Muhammad and Bria, 2023; Varlamova et al., 2024). DL methods can be categorized based on the level of supervision involved. Supervised learning includes techniques like Convolutional Neural Networks (CNNs) for tumor image classification and Recurrent Neural Networks (RNNs) for predicting patient outcomes, both requiring labeled data (LeCun et al., 2015; Iqbal et al., 2019, 2022). Unsupervised deep learning methods, such as Autoencoders and Generative Adversarial Networks (GANs), learn from unlabeled data to perform tasks like clustering patients based on gene expression profiles or generating synthetic medical images. Semi-supervised deep learning methods, like Semi-Supervised GANs, leverage a mix of labeled and unlabeled data to enhance model performance when labeled medical data is limited. Self-supervised learning methods, such as BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer), use the structure of training data itself for supervision, enabling tasks like predicting patient outcomes or understanding the progression of cancer with limited labeled examples. Reinforcement learning in cancer studies, exemplified by Deep Q-Networks (DQN) and Proximal Policy Optimization (PPO), involves an agent learning optimal treatment strategies through rewards and penalties.

As the volume of oncology data continues to grow, DL stands at the forefront of this field, enhancing our understanding of cancer, improving diagnostic precision, predicting clinical outcomes, and paving the way for innovative treatments. This review explores the latest advancements in DL applications within oncology, highlighting its potential to revolutionize cancer care (Chan et al., 2020; Ibrahim et al., 2022; Ghaffari Laleh et al., 2023; Tripathi et al., 2024a).

Multimodal Learning (MML) enhances task accuracy and reliability by leveraging information from various data sources or modalities (Huang et al., 2021). This approach has witnessed a surge in popularity, as indicated by the growing body of MML-related publications (see Figure 1). By facilitating the fusion of multimodal data, such as radiological images, digitized pathology slides, molecular data, and electronic health records (EHR), MML offers a richer understanding of complex problems (Tripathi et al., 2024c). It enables the extraction and integration of relevant features that might be overlooked when analyzing data modalities separately. Recent advancements in MML, powered by Deep Neural Networks (DNNs), have shown remarkable capability in learning from diverse data sources, including computer vision (CV) and natural language processing (NLP) (Bommasani et al., 2022; Achiam et al., 2023). Prominent multimodal foundation models such as Contrastive Language-Image Pretraining (CLIP) and Generative Pretraining Transformer (GPT-4) by OpenAI have set new benchmarks in the field (Radford et al., 2021; Achiam et al., 2023). Additionally, the Foundational Language And Vision Alignment Model (FLAVA) represents another significant stride, merging vision and language representation learning to facilitate multimodal reasoning (Singh et al., 2022). Within the realm of oncology, innovative applications of MML are emerging. The RadGenNets model, for instance, integrates clinical and genomics data with Positron Emission Tomography (PET) scans and gene mutation data, employing a combination of Convolutional Neural Networks (CNNs) and Dense Neural Networks to predict gene mutations in Non-small cell lung cancer (NSCLC) patients (Tripathi et al., 2022). Moreover, GNNs and Transformers are being explored for a variety of oncology-related tasks, such as tumor classification (Khan et al., 2020), prognosis prediction (Schulz et al., 2021), and assessing treatment response (Joo et al., 2021).

Recent literature has seen an influx of survey and review articles exploring MML (Baltrušaitis et al., 2018; Boehm et al., 2021; Ektefaie et al., 2023; Xu et al., 2023; Hartsock and Rasool, 2024). These works have provided valuable insights into the evolving landscape of MML, charting key trends and challenges within the field. Despite this growing body of knowledge, there remains a notable gap in the literature regarding the application of advanced multimodal DL models, such as Graph Neural Networks (GNNs) and Transformers, in the domain of oncology. Our article aims to fill this gap by offering the following contributions:

1. *Identifying large-scale MML approaches in oncology*. We provide an overview of the state-of-the-art MML with a special focus on GNNs and Transformers for multimodal data fusion in oncology.
2. *Highlighting the challenges and limitations of MML in oncology data fusion*. We discuss the challenges and limitations of implementing multimodal data-fusion models in oncology, including the need for large datasets, the complexity of integrating diverse data types, data alignment, and missing data modalities and samples.
3. *Providing a taxonomy for describing multimodal architectures*. We present a comprehensive taxonomy for describing MML architectures, including both traditional ML and DL, to facilitate future research in this area.
4. *Identifying future directions for multimodal data fusion in oncology*. We identify GNNs and Transformers as potential solutions for comprehensive multimodal integration and present the associated challenges.

By addressing these aspects, our article seeks to advance the understanding of MML's potential in oncology, paving the way for innovative solutions that could revolutionize cancer diagnosis and treatment through comprehensive data integration.

Our paper is organized as follows. Section 2 covers the fundamentals of MML, including data modalities, taxonomy, data

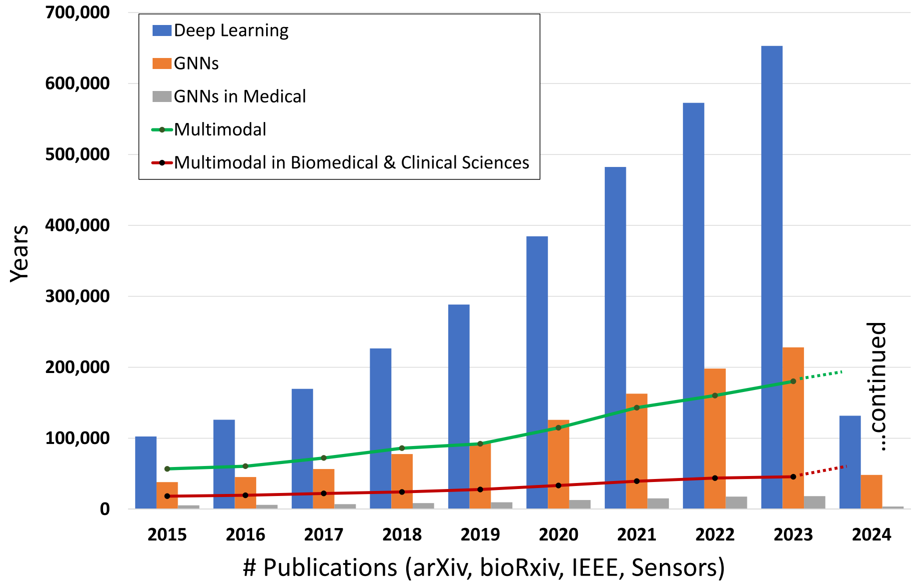

FIGURE 1
Number of publications involving DL, GNNs, GNNs in the medical domain, overall multimodal and multimodal in biomedical and clinical sciences in the period 2015–2024 (Hook et al., 2018).

fusion stages, and neural network architectures. Section 3 focuses on GNNs in MML, explaining graph data, learning on graphs, architectures, and applications to unimodal and multimodal oncology datasets. Section 4 discusses Transformers in MML, including architecture, multimodal Transformers, applications to oncology datasets, and methods of fusing data modalities. Section 5 highlights challenges in MML, such as data availability, alignment, generalization, missing data, explainability, and others. Section 6 provides information on data sources. Finally, we conclude by emphasizing the promise of integrating data across modalities and the need for scalable DL frameworks with desirable properties.

2 Fundamentals of multimodal learning (MML)

2.1 Data modalities in oncology

A data *modality* represents the expression of an entity or a particular form of sensory perception, such as the characters' visual actions, sounds of spoken dialogues, or the background music (Sleeman et al., 2022). A collective notion of these modalities is called *multi-modality* (Baltrušaitis et al., 2018). Traditional data analysis and ML methods to study cancer data use single data modalities [e.g., EHR (Miotto et al., 2016), radiology (Waqas et al., 2021), pathology (Litjens et al., 2017), or molecular, including genomics (Angermueller et al., 2017), transcriptomics (Yousefi et al., 2017), proteomics (Wang et al., 2017), etc.]. However, the data is inherently multimodal, as it includes information from multiple sources or modalities that are related in many ways. Figure 2 provides a view of multiple modalities of cancer at various scales, from the population level to single-cell analysis. Oncology data can be broadly classified into 3 categories: clinical, molecular, and imaging, where each category provides complementary information about the patient's disease. Figure 2 highlights different clinical, molecular, and imaging modalities. Multimodal analysis endeavors to gain holistic insights into the disease process using multimodal data.

2.1.1 Molecular data

Molecular data modalities provide information about the underlying genetic changes and alterations in the cancer cells (Liu et al., 2021). Efforts toward integrating molecular data resulted in the *multi-omics* research field (Waqas et al., 2024a). Two principal areas of molecular analysis in oncology are proteomics and genomics. *Proteomics* is the study of proteins and their changes in response to cancer, and it provides information about the biological processes taking place in cancer cells. *Genomics* is the study of the entire genome of cancer cells, including changes in DNA sequence, gene expression, and structural variations (Boehm et al., 2021). Other molecular modalities include transcriptomics, pathomics, radiomics and their combinations, radiogenomics, and proteogenomics. Many publicly available datasets provide access to molecular data, including the Proteomics Data Commons for proteomics data and the Genome Data Commons for genetic data (Grossman et al., 2016; Thangudu et al., 2020).

2.1.2 Imaging data

Imaging modalities play a crucial role in diagnosing and monitoring cancer. The imaging category can be divided into 2 main categories: (1) radiological imaging and (2) digitized histopathology slides, referred to as Whole Slide Imaging (WSI).

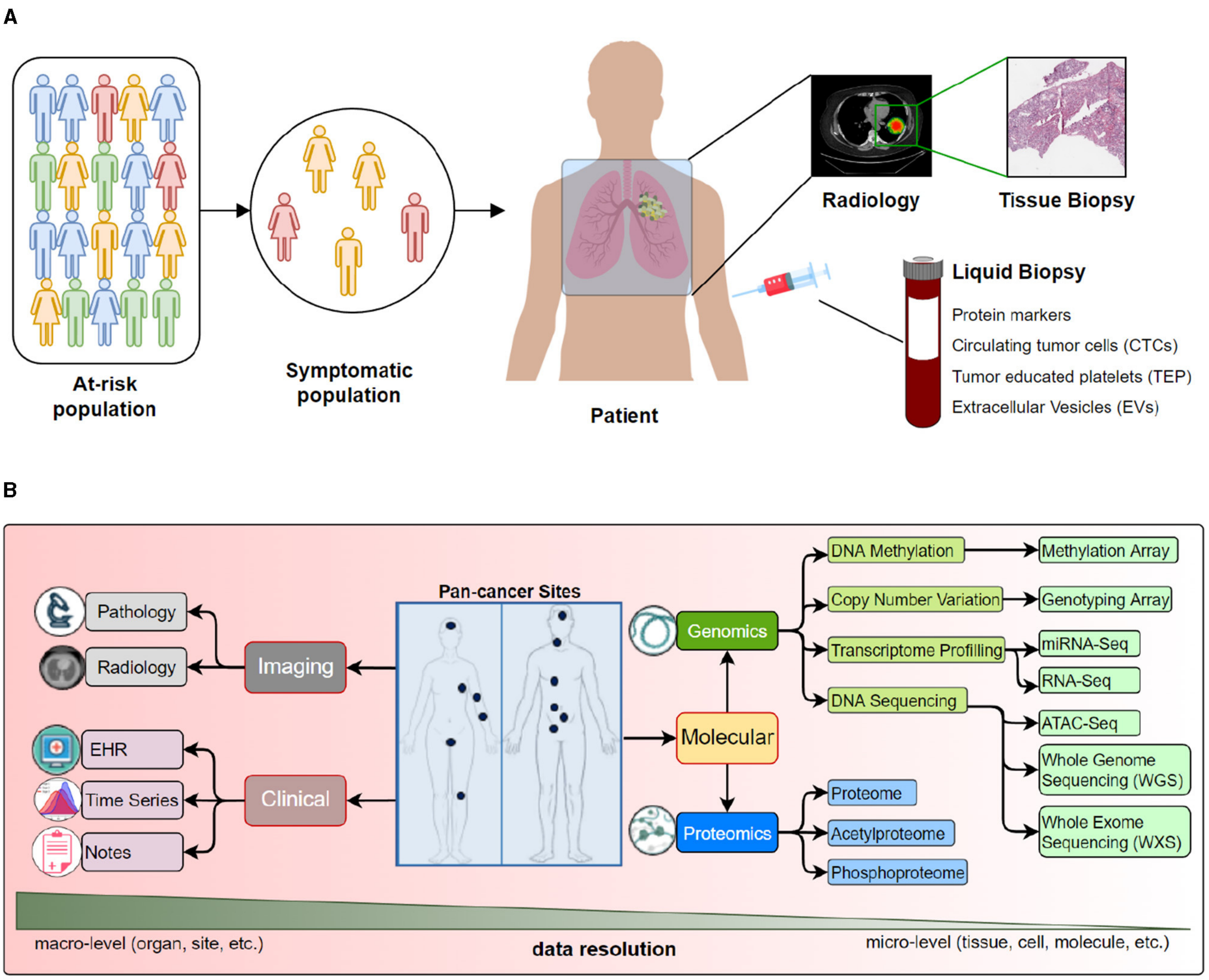

FIGURE 2
We present various data modalities that capture specific aspects of cancer at different scales. For example, radiological images capture organ or sub-organ level abnormalities, while tissue analysis may provide changes in the cellular structure and morphology. On the other hand, various molecular data types may provide insights into genetic mutations and epigenetic changes. **(A)** An overview of data collected from population to a tissue. **(B)** Detailed look into data modalities acquired for cancer care.

Radiological imaging encompasses various techniques such as X-rays, CT scans, MRI, PET, and others, which provide information about the location and extent of cancer within the body. These images can be used to determine the size and shape of a tumor, monitor its growth, and assess the effectiveness of treatments. *Histopathological* imaging is the examination of tissue samples obtained through biopsy or surgery (Rowe and Pomper, 2022; Waqas et al., 2023). Digitized slides, saved as WSIs, provide detailed information about the micro-structural changes in cancer cells and can be used to diagnose cancer and determine its subtype.

2.1.3 Clinical data

Clinical data provides information about the patient's medical history, physical examination, and laboratory results, saved in the patient's electronic health records (EHR) at the clinic. EHR consists of digital records of a patient's health information stored in a centralized database. These records provide a comprehensive view of a patient's medical history, past diagnoses, treatments, laboratory test results, and other information, which helps clinicians understand the disease (Asan et al., 2018). Within EHR, time-series data may refer to the clinical data recorded over time, such as repeated blood tests, lab values, or physical attributes. Such data informs the changes in the patient's condition and monitors the disease progression (Quinn et al., 2019).

2.2 Taxonomy of MML

We follow the taxonomy proposed by Sleeman et al. (2022) (see Figure 3), which defines 5 main stages of multimodal classification: preprocessing, feature extraction, data fusion, primary learner, and final classifier, as given below:

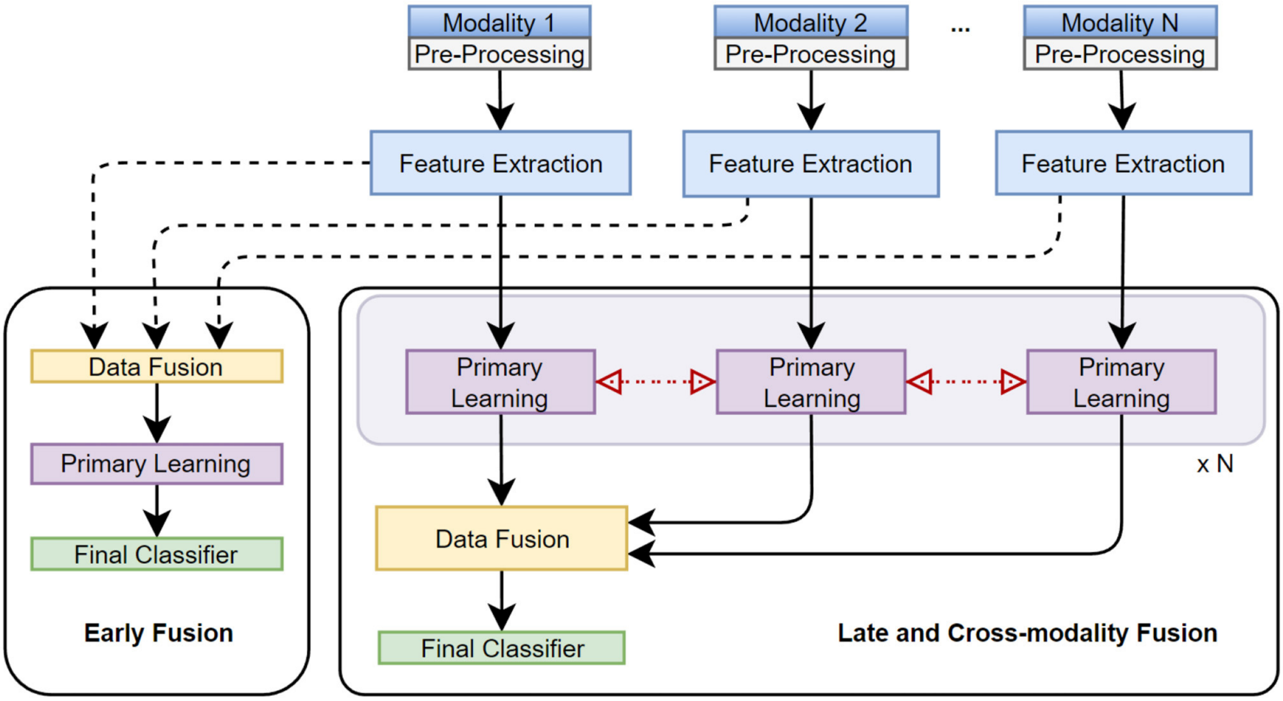

FIGURE 3
Taxonomy, stages, and techniques of multimodal data fusion are presented. *Early*, *late*, *cross-modality* fusion methods integrate individual data modalities (or extracted features) *before*, *after*, or *at* the primary learning step, respectively.

2.2.1 Pre-processing

Pre-processing involves modifying the input data to a suitable format before feeding it into the model for training. It includes data cleaning, normalization, class balancing, and augmentation. Data cleaning removes unwanted noise or bias, errors, and missing data points (Al-jabery et al., 2020). Normalization scales the input data within a specific range to ensure that each modality contributes equally to the training (Gonzalez Zelaya, 2019). Class balancing is done in cases where one class may have a significantly larger number of samples than another, resulting in a model bias toward the dominant class. Data augmentation artificially increases the size of the dataset by generating new samples based on the existing data to improve the model's robustness and generalizability (Al-jabery et al., 2020).

2.2.2 Feature extraction

Different data modalities may have different features, and extracting relevant features may improve model learning. Several manual and automated feature engineering techniques generate representations (or *embeddings*) for each data modality. Feature engineering involves designing features relevant to the task and extracting them from the input data. This can be time-consuming but may allow the model to incorporate prior knowledge about the problem. Text encoding techniques, such as bag-of-words, word embeddings, and topic models (Devlin et al., 2019; Zhuang et al., 2021), transform textual data into a numerical representation, which can be used as input to an ML model (Wang et al., 2020a). In DL, feature extraction is learned automatically during model training (Dara and Tumma, 2018).

2.2.3 Data fusion

Data fusion combines raw features, extracted features, or class prediction vectors from multiple modalities to create a single data representation. Fusion enables the model to use the complementary information provided by each modality and improve its learning. Data fusion can be done using early, late, or intermediate fusion. Section 2.3 discusses these fusion stages. The choice of fusion technique depends on the characteristics of the data and the specific problem being addressed (Jiang et al., 2022).

2.2.4 Primary learner

The primary learner stage is training the model on the pre-processed data or extracted features. Depending on the problem and data, the primary learner can be implemented using various ML techniques. DNNs are a popular choice for primary learners in MML because they can automatically learn high-level representations from the input data and have demonstrated state-of-the-art performance in many applications. CNNs are often used for image and video data, while recurrent neural networks (RNNs) and Transformers are commonly used for text and sequential data. The primary learner can be implemented independently for each modality or shared between modalities, depending on the problem and data.

2.2.5 Final classifier

The final stage of MML is the classifier, which produces category labels or class scores and can be trained on the output of the primary learner or the fused data. The final classifier can be implemented using a shallow neural network, a decision tree, or an ensemble model (Sleeman et al., 2022). Ensemble methods, such as stacking or boosting, are often used to improve and robustify the performance of the final classifier. Stacking involves training multiple models and then combining their predictions at the output stage, while boosting involves repeatedly training weak learners and adjusting their weights based on the errors made by previous learners (Borisov et al., 2022).

2.3 Data fusion strategies

Fusion in MML can be performed at different levels, including early (feature level), intermediate (model level), or late (decision level) stages, as illustrated in Figure 3. Each fusion stage has its advantages and challenges, and the choice of fusion stage depends on the characteristics of the data and the task.

2.3.1 Early fusion

The early fusion involves merging features extracted from different data modalities into a single feature vector before model training. The feature vectors of the different modalities are combined into a single vector, which is used as the input to the ML model (Sleeman et al., 2022). This approach can be used when the modalities have complementary information and can be easily aligned, such as combining visual and audio features in a video analysis application. The main challenge with early fusion is ensuring that the features extracted from different modalities are compatible and provide complementary information.

2.3.2 Intermediate fusion

Intermediate fusion involves training separate models for each data modality and then combining the outputs of these models for inference/prediction (Sleeman et al., 2022). This approach is suitable when the data modalities are independent of each other and cannot be easily combined at the feature level using average, weighted average, or other methods. The main challenge with intermediate fusion is selecting an appropriate method for combining the output of different models.

2.3.3 Late fusion

In late fusion, the output of each modality-specific model is used to make a decision independently. All decisions are later combined to make a final decision. This approach is suitable when the modalities provide complementary information but are not necessarily independent of each other. The main challenge with late fusion is selecting an appropriate method for combining individual predictions. This can be done using majority voting, weighted voting, or employing other ML models.

2.4 MML for oncology datasets

Syed et al. (2021) used a Random Forest classifier to fuse radiology image representations learned from the singular value decomposition method with the textual annotation representation learned from the fastText algorithm for prostate and lung cancer patients. Liu et al. (2022) proposed a hybrid DL framework for combining breast cancer patients' genomic and pathology data using fully-connected (FC) network for genomic data, CNN for radiology data and a Simulated Annealing algorithm for late fusion. Multiview multimodal network (MVMM-Net) (Song J. et al., 2021) combined 2 different modalities (low-energy and dual-energy subtracted) from contrast-enhanced spectral mammography images, each learned through CNN and late-fusion through FC network in breast cancer detection task. Yap et al. (2018) used a late-fusion method to fuse image representations from ResNet50 and clinical representations from a random forest model for a multimodal skin lesion classification task. An award-winning work (Ma and Jia, 2020) on brain tumor grade classification adopted the late-fusion method (concatenation) for fusing outputs from two CNNs (radiology and pathology images). SeNMo, a self-normalizing deep learning model has shown that integrative analysis on 33 cancers having five different molecular (multi-omics) data modalities can improve the patient outcome predictions and primary cancer type classification (Waqas et al., 2024a). Recently, GNNs-based pan-squamous cell carcinoma analysis on lung, bladder, cervicall, esophageal, and head and neck cancers has outperformed different classical and deep learning models (Waqas et al., 2024b).

The single-cell unimodal data alignment is one technique in MML. Jansen et al. (2019) devised an approach (SOMatic) to combine ATAC-seq regions with RNA-seq genes using self-organizing maps. Single-Cell data Integration via Matching (SCIM) matched cells in multiple datasets in low-dimensional latent space using autoencoder (AEs) (Stark et al., 2020). Graph-linked unified embedding (GLUE) model learned regulatory interactions across omics layers and aligned the cells using variational AEs (Cao and Gao, 2022). These aforementioned methods cannot incorporate high-order interactions among cells or different modalities. Single-cell data integration using multiple modalities is mostly based on AEs [scDART (Zhang Z. et al., 2022), Cross-modal Autoencoders (Yang K. D. et al., 2021), Mutual Information Learning for Integration of Single Cell Omics Data (SMILE) (Xu et al., 2022)].

The relevant works discussed in this section is summarized in Table 1.

3 Graph Neural Networks in multimodal learning

Graphs are commonly used to represent the relational connectivity of any system that has interacting entities (Li M. et al., 2022). Graphs have been used in various fields, such as to study brain networks (Farooq et al., 2019), analyze driving maps (Derrow-Pinion et al., 2021), and explore the structure of DNNs themselves (Waqas et al., 2022). GNNs are specifically designed to process data represented as a graph (Waikhom and Patgiri, 2022), which makes them well-suited for analyzing multimodal oncology

TABLE 1 References discussed in Section 2.

Sections		References
Data modalities in oncology	Molecular	Grossman et al., 2016; Thangudu et al., 2020; Boehm et al., 2021; Liu et al., 2021; Waqas et al., 2024a
	Imaging	Rowe and Pomper, 2022; Waqas et al., 2023
	Clinical	Asan et al., 2018; Quinn et al., 2019
Taxonomy of MML		Dara and Tumma, 2018; Devlin et al., 2019; Gonzalez Zelaya, 2019; Al-jabery et al., 2020; Wang et al., 2020a; Zhuang et al., 2021; Borisov et al., 2022; Jiang et al., 2022; Sleeman et al., 2022
Data fusion strategies		Sleeman et al., 2022
MML for oncology datasets		Yap et al., 2018; Jansen et al., 2019; Ma and Jia, 2020; Stark et al., 2020; Song J. et al., 2021; Syed et al., 2021; Yang K. D. et al., 2021; Cao and Gao, 2022; Liu et al., 2022; Xu et al., 2022; Zhang Z. et al., 2022; Waqas et al., 2024a,b

data as each data modality (or sub-modality) can be considered as a single node and the structures/patterns that exist between data modalities can be modeled as edges (Ektefaie et al., 2023).

3.1 The graph data

A graph is represented as $G{=}(V, E)$ having node-set $V{=}\{v_1, v_2, ..., v_n\}$, where node v has feature vector $\mathbf{x}_v$, and edge set $E{=}\{(v_i, v_j) \mid v_i, v_j \in V\}$. The neighborhood of node v is defined as $N(v){=}\{u \mid (u, v) \in E\}$.

3.1.1 Graph types

As illustrated in Figure 4A, the common types of graphs include undirected, directed, homogeneous, heterogeneous, static, dynamic, unattributed, and attributed. *Undirected graphs* comprise undirected edges, i.e., the direction of relation is not important between any ordered pair of nodes. In the *directed graphs*, the nodes have a directional relationship(s). Homogeneous graphs have the same type of nodes, whereas heterogeneous graphs have different types of nodes within a single graph (Yang T. et al., 2021). Static graphs do not change over time with respect to the existence of edges and nodes. In contrast, dynamic graphs change over time, resulting in changes in structure, attributes, and node relationships. *Unattributed graphs* have unweighted edges, indicating that the weighted value for all edges in a graph is the same, i.e., 1 if present, 0 if absent. *Attributed graphs* have different edge weights that capture the strength of relational importance (Waikhom and Patgiri, 2022).

3.1.2 Tasks for graph data

In Figure 4B, we present 3 major types of tasks defined on graphs, including (1) *node-level tasks* - these may include node classification, regression, clustering, attributions, and generation, (2) *edge-level task* - edge classification and prediction (presence or absence) are 2 common edge-level tasks, (3) *graph-level tasks* - these tasks involve predictions on the graph level, such as graph classification and generation.

3.2 ML for graph data

Representing data as graphs can enable capturing and encoding the relationships among entities of the samples (Wu et al., 2020). Based on the way the nodes are encoded, representation learning on graphs can be categorized into the traditional (or shallow) and DNN-based methods, as illustrated in Figure 4C (Wu et al., 2020; Jiao et al., 2022).

3.2.1 Traditional (shallow) methods

These methods usually employ classical ML methods, and their two categories commonly found in the literature are *graph embedding* and *probabilistic methods*. Graph embedding methods represent a graph with low-dimensional vectors (graph embedding and node embedding), preserving the structural properties of the graph. The learning tasks in graph embedding usually involve dimensionality reduction through linear (principal component or discriminant analysis), kernel (nonlinear mapping), or tensor (higher-order structures) methods (Jiao et al., 2022). Probabilistic graphical methods use graph data to represent probability distribution, where nodes are considered random variables, and edges depict the probability relations among nodes (Jiao et al., 2022). Bayesian networks, Markov's networks, variational inference, variable elimination, and others are used in probabilistic methods (Jiao et al., 2022).

3.2.2 DNN-based methods - GNNs

GNNs are gaining popularity in the ML community, as evident from Figure 1. In GNNs, the information aggregation from the neighborhood is fused into a node's representation. Traditional DL methods such as CNNs and their variants have shown remarkable success in processing the data in Euclidean space; however, they fail to perform well when faced with non-Euclidean or relational datasets. Compared to CNNs, where the locality of the nodes in the input is fixed, GNNs have no canonical ordering of the neighborhood of a node. They can learn the given task for any permutation of the input data, as depicted in Figure 5. GNNs often employ a message-passing mechanism in which a node's representation is derived from its neighbors' representations via a

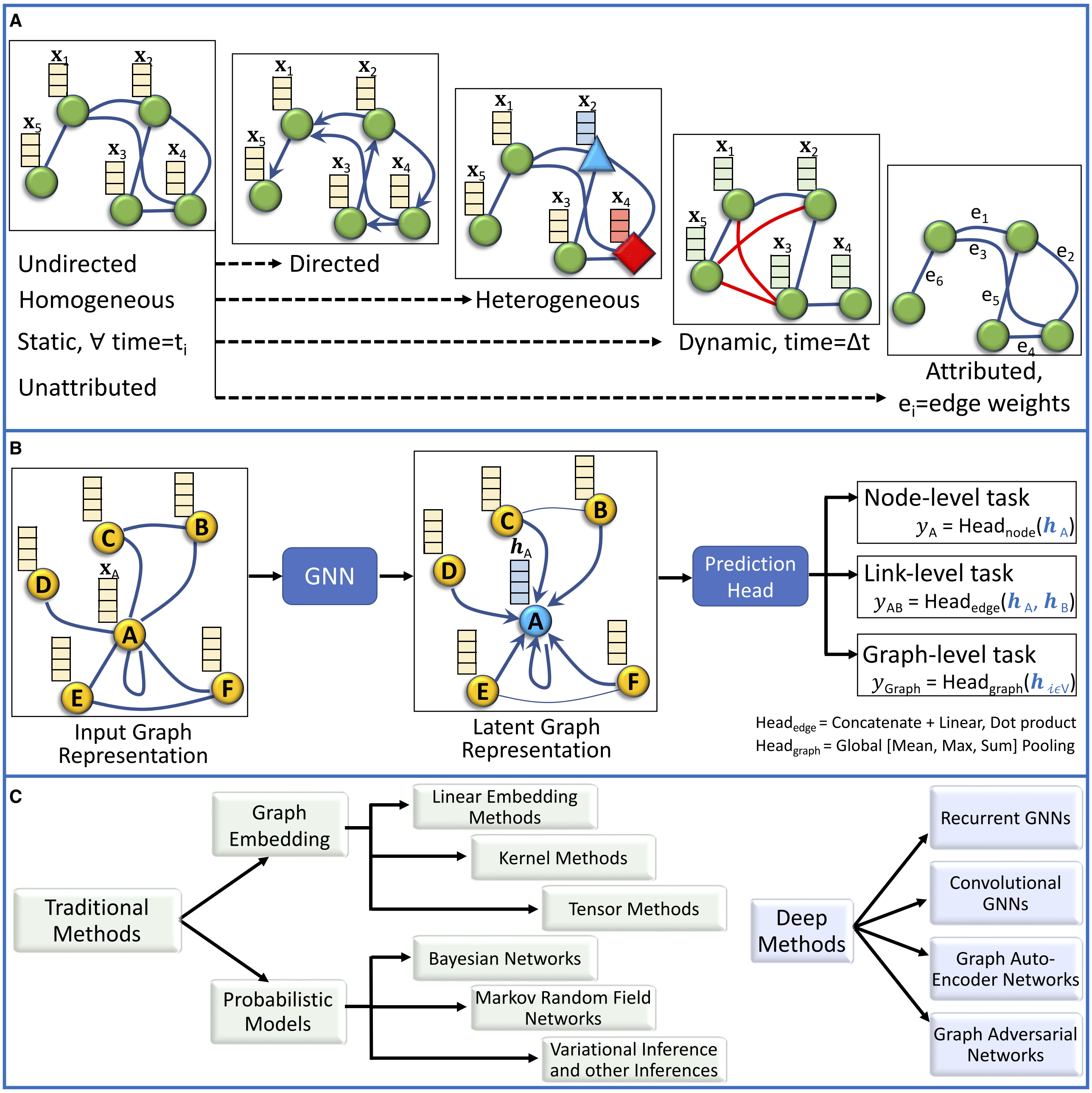

FIGURE 4
(A) The commonly occurring graph types are presented, including (1) undirected and directed, (2) homogeneous and heterogeneous, (3) dynamic and static, (4) attributed (edges) and unattributed. **(B)** Three different types of tasks performed using the graph data are presented and include (1) node-level, (2) link-level, and (3) graph-level analyses. **(C)** Various categories of representation learning for graphs are presented.

recursive computation. The message passing for a GNN is given as follows:

$$\mathbf{h}_v^{(l+1)} = \sigma\left(W_l \sum_{u \in N(v)} \frac{\mathbf{h}_u^{(l)}}{|N(v)|} + B_l \mathbf{h}_v^{(l)}\right) \quad (1)$$

where $h_v^{(l+1)}$ is the updated embedding of node v after l+1 layer, σ is the non-linear function (e.g., rectified linear unit or ReLU), $h_u^{(l)}$ and $h_v^{(l)}$ represent the embeddings of nodes u and v at layer l. W_l and B_l are the trainable weight matrices for neighborhood aggregation and (self)hidden vector transformation, respectively. The message passing can encode high-order structural information in node embedding through multiple aggregation layers. GNNs smooth the features by aggregating neighbors' embedding and filter eigenvalues of graph Laplacian, which provides an extra denoising mechanism (Ma Y. et al., 2021). GNNs comprise multiple permutation equivariant and invariant functions, and they can handle heterogeneous data (Jin et al., 2022). As described earlier, traditional ML models deal with Euclidean data. In oncology data,

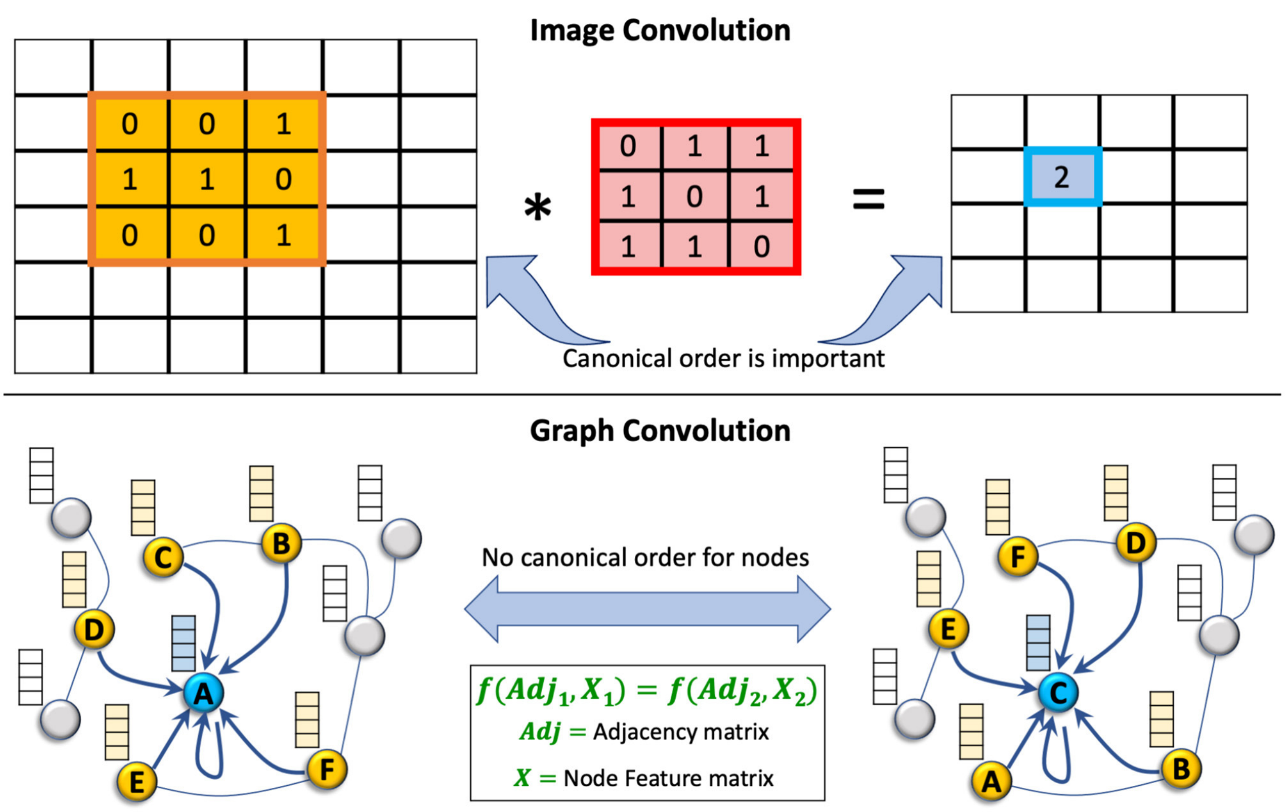
Image Convolution
0 0 1
1 1 0
0 0 1
*
0 1 1
1 0 1
1 1 0
=
2
Canonical order is important
Graph Convolution
No canonical order for nodes
$f(Adj_1, X_1) = f(Adj_2, X_2)$
Adj = Adjacency matrix
X = Node Feature matrix
A
B
C
D
E
F

FIGURE 5
Convolution operation for graphs vs. image data. The canonical order of the input is important in CNNs, whereas in GNNs, the order of the input nodes is not important. From the convolution operation perspective, CNNs can be considered a subset of GNNs (Hamilton, 2020).

the correlations may not exist in Euclidean space; instead, its features may be highly correlated in the non-Euclidean space (Yi et al., 2022). Based on the information fusion and aggregation methodology, GNNs-based deep methods are classified into the following:

3.2.2.1 Recurrent GNNs

RecGNNs are built on top of the standard Recurrent Neural Network (RNN) by combining with GNN. RecGNNs can operate on graphs with variable sizes and topologies. The recurrent component of the RecGNN captures temporal dependencies and learns latent states over time, whereas the GNN component captures the local structure. The information fusion process is repeated a fixed number of times until an equilibrium or the desired state is achieved (Hamilton et al., 2017). RecGNNs employ the model given by:

$$\mathbf{h}_v^{(l+1)} = \text{RecNN}\left(\mathbf{h}_u^{(l)}, \mathbf{Msg}_{N(v)}^{(l)}\right), \tag{2}$$

where, RecNN is any RNN, and $Msg_{N(v)}^{(l)}$ is the neighborhood message-passing at layer l.

3.2.2.2 Convolutional GNNs

ConvGNNs undertake the convolution operation on graphs by aggregating neighboring nodes' embeddings through a stack of multiple layers. ConvGNNs use the symmetric and normalized summation of the neighborhood and self-loops for updating the node embeddings given by:

$$\mathbf{h}_v^{(l+1)} = \sigma\left(W_l \sum_{u \in N(v) \cup v} \frac{\mathbf{h}_v}{\sqrt{|N(v)||N(u)|}}\right). \tag{3}$$

The ConvGNN can be spatial or spectral, depending on the type of convolution they implement. Convolution in spatial ConvGNNs involves taking a weighted average of the neighboring vertices. Examples of spatial ConvGNNs include GraphSAGE (Hamilton et al., 2017), Message Passing Neural Network (MPNN) (Gilmer et al., 2017), and Graph Attention Network (GAT) (Veličković et al., 2017). The spectral ConvGNNs operate in the spectral domain by using the eigendecomposition of the graph Laplacian matrix. The convolution operation is performed on the eigenvalues, which can be high-dimensional. Popular spectral ConvGNNs are ChebNet (Defferrard et al., 2016) and Graph Convolutional Network (GCN) (Kipf and Welling, 2016). An interesting aspect of these approaches is representational containment, which is defined as: convolution ⊆ attention ⊆ message passing.

3.2.2.3 Graph Auto-Encoder Networks

GAEs are unsupervised graph learning networks for dimensionality reduction, anomaly detection, and graph generation. They are built on top of the standard AEs to work with graph data. The encoder component of the GAE maps the input graph to a low-dimensional latent space, while the decoder component maps the latent space back to the original graph (Park et al., 2021).

3.2.2.4 Graph Adversarial Networks

Based on Generative Adversarial Networks, GraphANs are designed to work with graph-structured data and can learn to generate new graphs with similar properties to the input data. The generator component of the GraphAN maps a random noise vector to a new graph, while the discriminator component tries to distinguish between the generated vs. the actual input. The generator generates graphs to fool the discriminator, while the discriminator tries to classify the given graph as real or generated.

3.2.2.5 Other GNNs

Other categories of GNNs may include scalable GNNs (Ma et al., 2019), dynamic GNNs (Sankar et al., 2018), hypergraph GNNs (Bai et al., 2021), heterogeneous GNNs (Wei et al., 2019), and many others (Ma and Tang, 2021).

3.2.3 Graph-based reinforcement learning

GNNs have been combined with Reinforcement Learning (RL) to solve complex problems involving graph-structured data (Jiang et al., 2018). GNNs enable RL agents to effectively process and reason about relational information in environments represented as graphs (Nie et al., 2023). This combination has shown promise in various domains, including multi-agent systems, robotics, and combinatorial optimization (Almasan et al., 2022; Fathinezhad et al., 2023). However, the use of Graph-based RL on cancer data is still less-explored area of research.

3.3 GNNs and ML using unimodal oncology datasets

3.3.1 Pathology datasets

Traditionally, CNN-based models are used to learn features from digital pathology data (Iqbal et al., 2022). However, unlike GNNs, CNNs fail to capture the global contextual information important in the tissue phenotypical and structural micro and macro environment (Ahmedt-Aristizabal et al., 2022). For using histology images in GNNs, the cells, tissue regions, or image patches are depicted as nodes. The relations and interactions among these nodes are represented as (un)weighted edges. Usually, a graph of the patient histology slide is used along with a patient-level label for training a GNN, as illustrated in Figure 6A. Here, we review a few GNN-based pathology publications representative of a trove of works in this field. Histographs (Anand et al., 2020) used breast cancer histology data to distinguish cancerous and non-cancerous images. Pre-trained VGG-UNet was used for nuclei detection, micro-features of the nuclei were used as node features, and Euclidean distance among nuclei was incorporated as edge features. The resulting cell graphs were used to train the GCN-based robust spatial filtering (RSF) model, which performed superior to the CNN-based classification. citewang2020weakly analyzed grade classification in tissue micro-arrays of prostate cancer using the weakly-supervised technique on a variant of GraphSAGE with self-attention pooling (SAGPool). Cell-Graph Signature ($CG_{signature}$) (Wang et al., 2022) predicted patient survival in gastric cancer using cell-graphs of multiplexed immunohistochemistry images processed through two types of GNNs (GCNs and GINs) with two types of pooling (SAGPool, TopKPool). Besides the above-mentioned cell graphs, there is an elaborate review of GNN-based tissue graphs or patch-graphs methods implemented on unimodal pathology cancer data given in Ahmedt-Aristizabal et al. (2022). Instead of individual cell- and tissue-graphs, a combination of the multilevel information in histology slides can help understand the intrinsic features of the disease.

3.3.2 Radiology datasets

GNNs have been used in radiology-based cancer data for segmentation, classification, and prediction tasks, especially on X-rays, mammograms, MRI, PET, and CT scans. Figure 6B illustrates a general pipeline of using radiology-based data to train GNNs. Here we give a non-exhaustive review of GNNs-based works on radiological oncology data as a single modality input. Mo et al. (2020) proposed a framework that improved the liver cancer lesion segmentation in the MRI-T1WI scans through guided learning of MRI-T2WI modality priors. Learned embeddings from fully convolutional networks on separate MRI modalities are projected into the graph domain for learning by GCNs through the co-attention mechanism and finally to get the refined segmentation by re-projection. Radiologists usually review radiology images by zooming into the region of interest (ROIs) on high-resolution monitors. Du et al. (2019) used a hierarchical GNN framework to automatically zoom into the abnormal lesion region of the mammograms and classify breast cancer. The pre-trained CNN model extracts image features, whereas a GAT model is used to classify the nodes for deciding whether to zoom in or not based on whether it is benign or malignant. Based on the established knowledge that lymph nodes (LNs) have connected lymphatic system and LNs cancer cells spread on certain pathways, Chao et al. (2020) proposed a lymph node gross tumor volume learning framework. The framework was able to delineate the LN appearance as well as the inter-LN relationship. The end-to-end learning framework was superior to the state-of-the-art on esophageal cancer radiotherapy dataset. Tian et al. (2020) suggested interactive segmentation of MRI scans of prostate cancer patients through a combination of CNN and two GCNs. CNN model outputs a segmentation feature map of MRI, and the GCNs predict the prostate contour from this feature map. Saueressig et al. (2021) used GNNs to segment brain tumors in 3D MRI images, formed by stacking different modalities of MRI (T1, T2, T1-CE, FLAIR) and representing them as supervoxel graph. The authors reported that GraphSAGE-pool was best for segmenting brain tumors. Besides radiology, a parallel field of radiomics has recently gained attraction. Radiomics is the automated extraction of quantitative features from radiology scans. A survey of radiomics and radiogenomic analysis on brain tumors is presented by Singh et al. (2021).

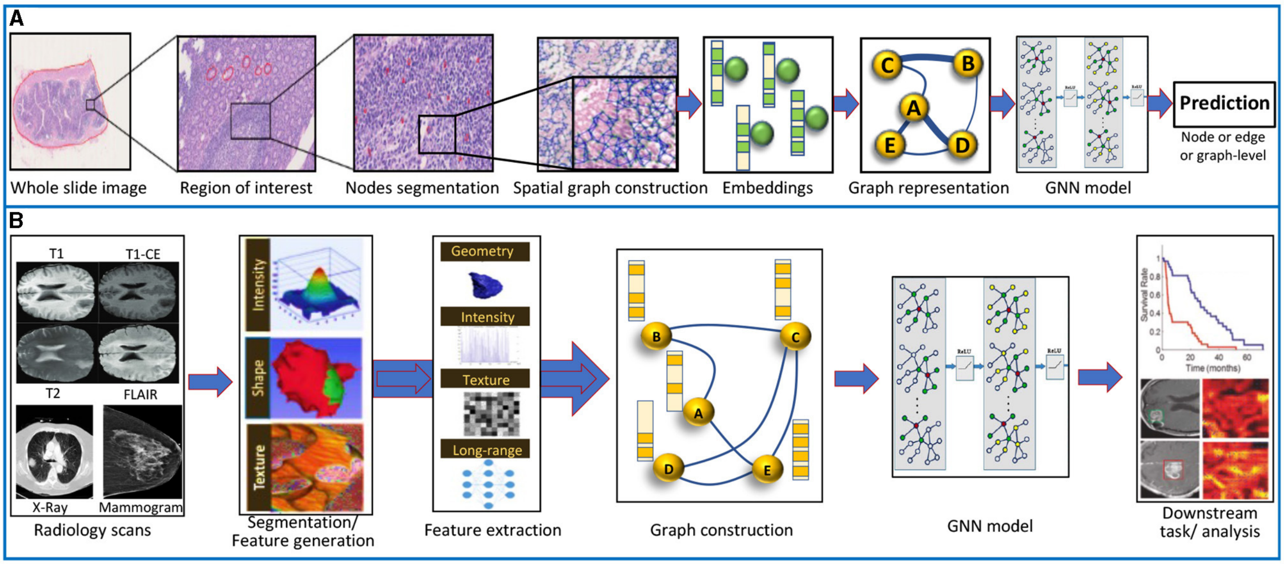

FIGURE 6
(A) Data processing pipeline for histopathology images using GNNs (Chen et al., 2020). **(B)** Graph processing pipeline on radiology data. Adapted from Singh et al. (2021).

3.3.3 Molecular datasets

Graphs are a natural choice for representing molecular data such as omic-centric (DNA, RNA, or proteins) or single-cell centric. Individual modalities are processed separately to generate graph representations that are then processed through GNNs followed by the classifier to predict the downstream task, as illustrated in Figure 7. One method of representing proteins as graphs is to depict the amino acid residue in the protein as the node and the relationship between residues denoted by edge (Fout et al., 2017). The residue information is depicted as node embedding, whereas the relational information between two residues is represented as the edge feature vector. Fout et al. (2017) used spatial ConvGNNs to predict interfaces between proteins which is important in drug discovery problems. Deep predictor of drug-drug interactions (DPDDI) predicted the drug-drug interactions using GCN followed by a 5-layer classical neural network (Feng et al., 2020). Molecular pre-training graph net (MPG) is a powerful framework based on GNN and Bidirectional Encoder Representations from Transformers (BERT) to learn drug-drug and drug-target interactions (Li et al., 2021b). Graph-based Attention Model (GRAM) handled the data inefficiency by supplementing EHRs with hierarchical knowledge in the medical ontology (Choi et al., 2017). A few recent works have applied GNNs to single-cell data. scGCN is a knowledge transfer framework in single-cell omics data such as mRNA or DNA (Song Q. et al., 2021). scGNN processed cell-cell relations through GNNs for the task of missing-data imputation and cell clustering on single-cell RNA sequencing (scRNA-seq) data (Wang J. et al., 2021).

3.4 MML—Data fusion at the pre-learning stage

The first and most primitive form of MML is the pre-learning fusion (see Figure 3), where features extracted from individual modalities of data are merged, and the fused representations are then used for training the multimodal primary learner model. In the context of GNNs being the primary learning model, the extraction step of individual modality representations can be hand-engineered (e.g., dimensionality reduction) or learned by DL models (e.g., CNNs, Transformers). Cui et al. (2021) proposed a GNN-based early fusion framework to learn latent representations from radiological and clinical modalities for Lymph node metastasis (LNM) prediction in esophageal squamous cell carcinoma (ESCC). The extracted features from the two modalities using UNet and CNN-based encoders were fused together with category-wise attention as node representation. The message passing from conventional GAT and correlation-based GAT learned the neighborhood weights. The attention attributes were used to update the final node features before classification by a 3-layer fully connected network. For Autism spectrum disorder, Alzheimer's disease, and ocular diseases, a multimodal learning framework called Edge-Variational GCN (EV-GCN) fuses the radiology features extracted from fMRI images with clinical feature vectors for each patient (Huang and Chung, 2020). An MLP-based pairwise association encoder is used to fuse the input feature vectors and to generate the edge weights of the population graph. The partially labeled population graph is then processed through GCN layers to generate the diagnostic graph of patients.

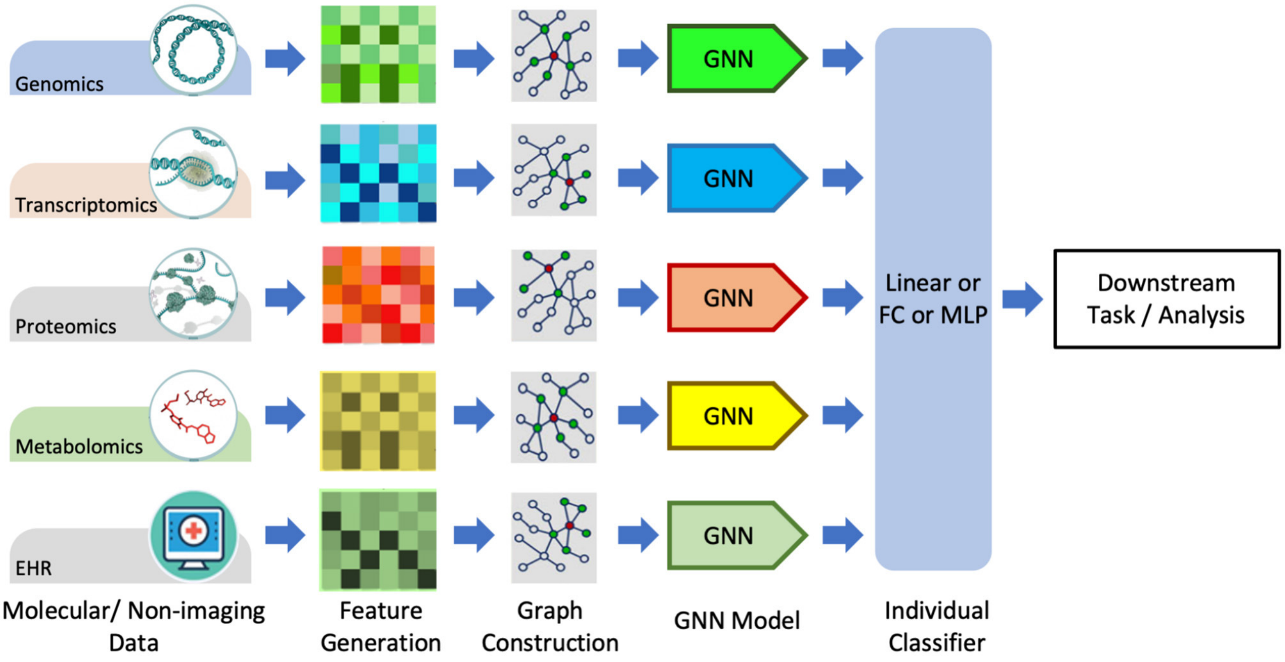

FIGURE 7
Graph data processing pipeline for non-imagery data, including molecular and textual data. Adapted from Wang T. et al. (2021). GNN, graph neural network, FC, Fully-Connected; MLP, Multi-Layer Perception.

3.5 MML—Data fusion using cross-modality learning

Cross-MML involves intermediate fusion and/or cross-links among the models being trained on individual modalities (see Figure 3). For this survey, we consider the GNN-based hierarchical learning mechanisms as the cross-MML methods. Hierarchical frameworks involve learning for one modality and using the learned latent embeddings in tandem with other data modalities sequentially to get the final desired low-dimensional representations. Lian et al. (2022) used a sequential learning framework where tumor features learned from CT images using the ViT model were used as node features of the patient population graph for subsequent processing by the GraphSAGE model. The hierarchical learning from radiological and clinical data using Transformer-GNN outperformed the ResNet-Graph framework in survival prediction of early-stage NSCLC. scMoGNN is the first method to apply GNNs in multimodal single-cell data integration using a cross-learning fusion-based GNN framework (Wen et al., 2022). Officially winning first place in modality prediction task at the NeurIPS 2021 competition, scMoGNN showed superior performance on various tasks by using paired data to generate cell-feature graphs. Hierarchical cell-to-tissue-graph network (HACT-Net) combined the low-level cell-graph features with the high-level tissue-graph features through two hierarchical GINs on breast cancer multi-class prediction (Pati et al., 2020). Data imputation, a method of populating the missing values or false zero counts in single-cell data mostly done using DL autoencoders (AE) architecture, has recently been accomplished using GNNs. scGNN (Wang J. et al., 2021) used imputation AE and graph AE in an iterative manner for imputation, and GraphSCI (Rao et al., 2021) used GCN with AE to impute the single-cell RNA-seq data using the cross-learning fusion between the GCN and the AE networks. Clustering is a method of characterizing cell types within a tissue sample. Graph-SCC clustered cells based on scRNA-seq data through self-supervised cross-learning between GCN and a denoising AE network (Zeng et al., 2020). Recently, a multilayer GNN framework, Explainable Multilayer GNN (EMGNN), has been proposed for cancer gene prediction tasks using multi-omics data from 16 different cancer types (Chatzianastasis et al., 2023).

3.6 MML—Data fusion in post-learning regime

Post-learning fusion methods include processing individual data modalities and later fusing them for the downstream predictive task (Tortora et al., 2023). In the post-learning fusion paradigm, the hand-crafted features perform better than the deep features when the dimensionality of input data is low, and vice versa (Tortora et al., 2023). Many interesting GNN-based works involving the post-learning fusion mechanism have recently been published. Decagon used a multimodal approach on GCNs using proteins and drug interactions to predict exact side effects as a multi-relational link prediction task (Zitnik et al., 2018). Drug-target affinity (DTA) experimented with four different flavors of GNNs (GCN, GAT, GIN, GAT-GCN) along with a CNN to fuse together molecular embeddings and protein sequences for predicting drug-target affinity (Nguyen et al., 2021). PathomicFusion combined the morphological features extracted from image patches (using CNNs), cell-graph features from cell-graphs of histology images (GraphSAGE-based GCNs), and genomic features (using a feed-forward network) for survival prediction on glioma and clear cell renal cell carcinoma (Chen et al., 2020). Shi et al. (2019) proposed a late-fusion technique to study screening of cervical cancer at early stages by using CNNs to extract features from histology images, followed by K-means clustering to generate graphs which are processed through two-layer GCN. BDR-CNN-GCN (batch normalized, dropout, rank-based pooling) used the same mammographic images to extract image-level features using CNN and relation-aware features using GCN (Zhang et al., 2021). The two feature sets are fused using a dot product followed by a trainable linear projection for breast cancer classification. Under the umbrella of multi-omics data, many GNN-based frameworks have been proposed recently. Molecular omics network(MOOMIN), a multi-modal heterogeneous GNN to predict oncology drug combinations, processed molecular structure, protein features, and cell lines through GCN-based encoders, followed by late-fusion using a bipartite drug-protein interaction graph (Rozemberczki et al., 2022). Multi-omics graph convolutional networks (MOGONET) used a GCN-GAN late fusion technique for the classification of four different diseases, including three cancer types: breast, kidney, and glioma (Wang T. et al., 2021). Leng et al. (2022) extended MOGONET to benchmark three multi-omics datasets on two different tasks using sixteen DL networks and concluded that GAT-based GNN had the best classification performance. Multi-Omics Graph Contrastive Learner(MOGCL) used graph structure and contrastive learning information to generate representations for improved downstream classification tasks on the breast cancer multi-omics dataset using late-fusion (Rajadhyaksha and Chitkara, 2023). Similar to MOGCL, Park et al. (2022) developed a GNN-based multi-omics model that integrated mRNA expression, DNA methylation, and DNA sequencing data for NSCLC diagnosis.

The relevant works discussed in this section is summarized in Table 2.

4 Transformers in MML

Transformers are attention-based DNN models originally proposed for NLP (Vaswani et al., 2017). Transformers implement scaled dot-product of the input with itself and can process various types of data in parallel (Vaswani et al., 2017). Transformers can handle sequential data and learn long-range dependencies, making them well-suited for tasks such as language translation, language modeling, question answering, and many more (Otter et al., 2021). Unlike Recurrent Neural Networks (RNNs) and CNNs, Transformers use self-attention operations to weigh the importance of different input tokens (or embeddings) at each time step. This allows them to handle sequences of arbitrary length and to capture dependencies between input tokens that are far apart in the sequence (Vaswani et al., 2017). Transformers can be viewed as a type of GNN (Xu et al., 2023). Transformers are used to process other data types, such as images (Dosovitskiy et al., 2020), audio (Zhang, 2020), and time-series analysis (Ahmed et al., 2022b), resulting in a new wave of multi-modal applications. Transformers can handle input sequences of different modalities in a unified way, using the same self-attention mechanism, which processes the inputs as a fully connected graph (Xu et al., 2023). This allows Transformers to capture complex dependencies between different modalities, such as visual and textual information in visual question-answering (VQA) tasks (Ma J. et al., 2021).

Pre-training Transformers on large amounts of data, using unsupervised or self-supervised learning, and then fine-tuning for specific downstream tasks, has led to the development of foundation models (Boehm et al., 2021), such as BERT (Devlin et al., 2019), GPT (Radford et al., 2018), RoBERTa (Zhuang et al., 2021), CLIP (Radford et al., 2021), T5 (Raffel et al., 2020), BART (Lewis et al., 2019), BLOOM (Scao et al., 2022), ALIGN (Jia et al., 2021), CoCa (Yu et al., 2022) and more. Multimodal Transformers are a recent development in the field of MML, which extends the capabilities of traditional Transformers to handle multiple data modalities. The inter-modality dependencies are captured by the cross-attention mechanism in multimodal Transformers, allowing the model to jointly reason and extract rich data representations. There are various types of multimodal Transformers, such as Unified Transformer (UniT) (Hu and Singh, 2021), Multi-way Multimodal Transformer (MMT) (Tang et al., 2022), CLIP (Radford et al., 2021), Flamingo (Alayrac et al., 2022), CoCa (Yu et al., 2022), Perceiver IO (Jaegle et al., 2021), and GPT-4 (Achiam et al., 2023).

4.1 Model architecture

The original Transformer (Figure 8) was composed of multiple encoder and decoder blocks, each made up of several layers of self-attention and feed-forward neural networks. The encoder takes the input sequence and generates hidden representations, which are then fed to the decoder. The decoder generates the output sequence by attending to the encoder's hidden representations and the previous tokens (i.e., auto-regressive). The self-attention operation (or scaled dot-product) is a crucial component of the Transformer. It determines the significance of each element in the input sequence with respect to the whole input. Self-attention operates by computing a weighted sum of the input sequence's hidden representations, where the weights are determined by the dot product between the *query* vector and the *key* vector, followed by a scaling operation to stabilize the gradients. The resulting weighted sum is multiplied by a *value* vector to obtain the output of the self-attention operation. There has been a tremendous amount of work on various facets of Transformer architecture. The readers are referred to relevant review papers (Galassi et al., 2021; Otter et al., 2021; Han et al., 2023; Xu et al., 2023).

4.2 Multimodal transformers

Self-attention allows a Transformer model to process each input as a fully connected graph and attend to (or equivalently learn from) the global patterns present in the input. This makes Transformers compatible with various data modalities by treating each token (or its embedding) as a node in the graph. To use Transformers for a data modality, we need to tokenize the input and select an embedding space for the tokens. Tokenization and embedding selections are flexible and can be done at multiple

TABLE 2 References Discussed in Section 3.

Sections		References
Graphs and GNNs		Defferrard et al., 2016; Kipf and Welling, 2016; Gilmer et al., 2017; Hamilton et al., 2017; Veličković et al., 2017; Jiang et al., 2018; Sankar et al., 2018; Farooq et al., 2019; Ma et al., 2019; Wei et al., 2019; Wu et al., 2020; Bai et al., 2021; Derrow-Pinion et al., 2021; Ma and Tang, 2021; Ma Y. et al., 2021; Park et al., 2021; Yang T. et al., 2021; Almasan et al., 2022; Jiao et al., 2022; Jin et al., 2022; Li M. et al., 2022; Waikhom and Patgiri, 2022; Waqas et al., 2022; Yi et al., 2022; Ektefaie et al., 2023; Fathinezhad et al., 2023; Nie et al., 2023
GNNs and ML using Unimodal Oncology Datasets	Pathology	Anand et al., 2020; Wang et al., 2020b, 2022; Ahmedt-Aristizabal et al., 2022; Iqbal et al., 2022
	Radiology	Du et al., 2019; Chao et al., 2020; Mo et al., 2020; Tian et al., 2020; Saueressig et al., 2021; Singh et al., 2021
	Molecular	Choi et al., 2017; Fout et al., 2017; Feng et al., 2020; Li et al., 2021b; Song Q. et al., 2021; Wang J. et al., 2021
MML data fusion stages		Zitnik et al., 2018; Shi et al., 2019; Chen et al., 2020; Huang and Chung, 2020; Pati et al., 2020; Zeng et al., 2020; Cui et al., 2021; Nguyen et al., 2021; Rao et al., 2021; Wang J. et al., 2021; Wang T. et al., 2021; Zhang et al., 2021; Leng et al., 2022; Lian et al., 2022; Park et al., 2022; Rozemberczki et al., 2022; Wen et al., 2022; Chatzianastasis et al., 2023; Rajadhyaksha and Chitkara, 2023; Tortora et al., 2023

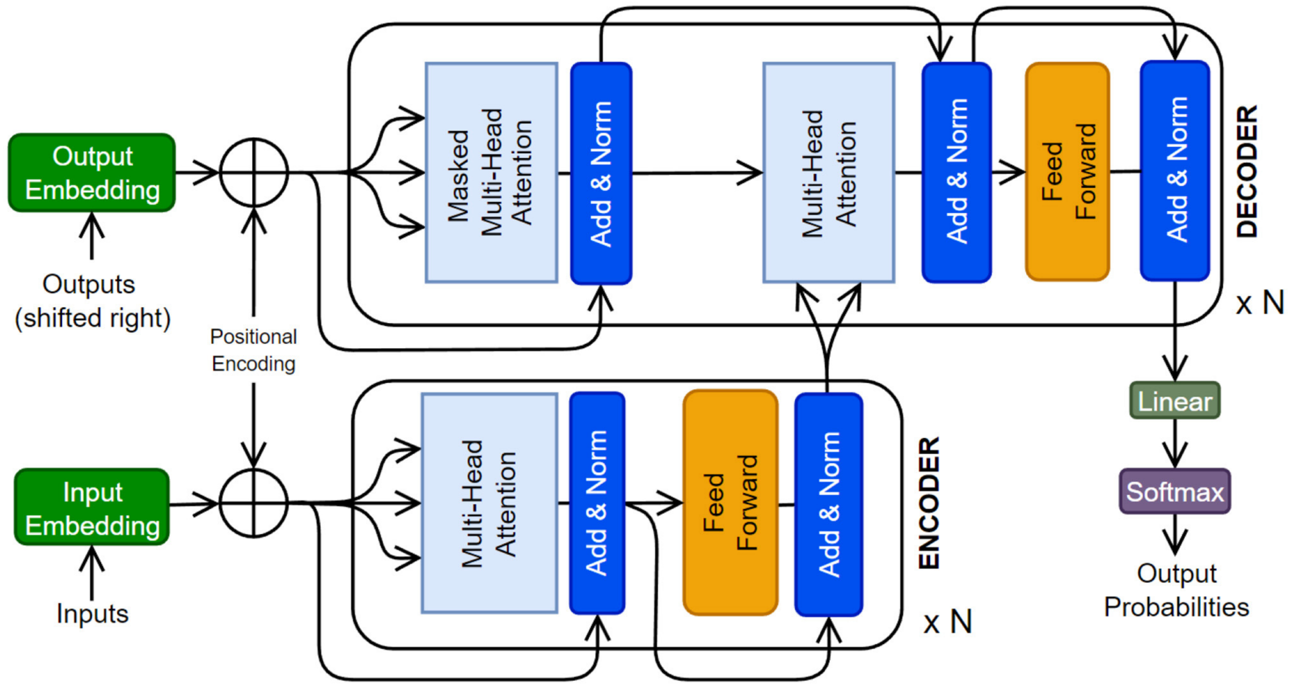

FIGURE 8
The original Transformer architecture is presented (Vaswani et al., 2017). A Transformer can have multiple encoder and decoder blocks, as well as some additional layers.

granularity levels, such as using raw features, ML-extracted features, patches from the input image, or graph nodes. Table 3 summarizes some common practices used for various types of data in cancer data sets. Handling inter-modality interactions is the main challenge in developing multimodal Transformer models. Usually, it is done through one of these fusion methods: *early fusion* of data modalities, *cross-attention*, *hierarchical attention*, and *late fusion*, as illustrated in Figure 9. In the following, we present and compare data processing steps for these four methods using two data modalities as an example. The same analysis can be extended to multiple modalities.

4.2.1 Early fusion

Early fusion is the simplest way to combine data from multiple modalities. The data from different modalities are concatenated to a single input before being fed to the Transformer model, which processes the input as a single entity. Mathematically, the concatenation operation is represented as $x_{cat}=[x_1, x_2]$, where x_1 and x_2 are the inputs from two data modalities, and x_{cat} is the concatenated input to the model. Early fusion is simple and efficient. However, it assumes that all modalities are equally important and relevant for the task at hand (Kalfaoglu et al., 2020), which may not always be practically true (Zhong et al., 2023).

TABLE 3 Oncology data modalities and their respective tokenization and embeddings selection techniques.

Data modalities	Tokenization level	Token embeddings model
Pathology images	Patch	CNNs (Chen et al., 2021)
Radiology images	Patch	CNNs (Xie et al., 2021)
EHR data	ICD code	GNNs (Shang et al., 2019),
		ML models (Rasmy et al., 2021)
-Omics	Graphs	GNNs (Kaczmarek et al., 2021)
	K-mers	ML model (Ji et al., 2020)
Clinical notes	Word	BERT (Devlin et al., 2019)
		RoBERTa (Zhuang et al., 2021)
		BioBERT (Lee et al., 2019)

4.2.2 Cross-attention fusion

Cross-attention is a relatively more flexible approach to modeling the interactions between data modalities and learning their joint representations. The self-attention layers attend to different modalities at different stages of data processing. Cross-attention allows the model to selectively attend to different modalities based on their relevance to the task (Li et al., 2021a) and capture complex interactions between the modalities (Rombach et al., 2022).

4.2.3 Hierarchical fusion

Hierarchical fusion is a complex approach to combining multiple modalities. For instance, the Depth-supervised Fusion Transformer for Salient Object Detection (DFTR) employs hierarchical feature extraction to improve salient object detection performance by fusing low-level spatial features and high-level semantic features from different scales (Zhu et al., 2022). Yang et al. (2020) introduced a hierarchical approach to fine-grained classification using a fusion Transformer. Furthermore, the Hierarchical Multimodal Transformer (HMT) for video summarization can capture global dependencies and multi-hop relationships among video frames (Zhao et al., 2022).

4.2.4 Late fusion

In late fusion, each data modality is processed independently by its own Transformer model, the branch outputs are concatenated and passed through the final classifier. Late fusion allows the model to capture the unique features of each modality while still learning their joint representation. Sun et al. (2021) proposed a multi-modal adaptive late fusion Transformer network for estimating the levels of depression. Their model extracts long-term temporal information from audio and visual data independently and then fuses weights at the end to learn a joint representation of data.

4.3 Transformers for processing oncology datasets

Transformers have been successfully applied to various tasks in oncology, including cancer screening, diagnosis, prognosis, treatment selection, and prediction of clinical variables (Boehm et al., 2021; Chen et al., 2021; Shao et al., 2021; Lian et al., 2022; Liang J. et al., 2022). For instance, a Transformer-based model was used to predict the presence and grade of breast cancer using a combination of imaging and genomics data (Boehm et al., 2021). TransMIL (Shao et al., 2021), a Transformer model, was proposed to process histopathology images using self-attention to learn and classify histopathology slides by overcoming the challenges faced by multi-instance learning (MIL). Recently, a Transformer and convolution parallel network, TransConv (Liang J. et al., 2022), was proposed for automatic brain tumor segmentation using MRI data. Transformers and GNNs have also been combined in MML for early-stage NSCLC prognostic prediction using the patient's clinical and pathological features and by modeling the patient's physiological network (Lian et al., 2022). Similarly, a multimodal co-attention Transformer was proposed for survival prediction using WSIs and genomic sequences (Chen et al., 2021). The authors used a co-attention mechanism to learn the interactions between the two data modalities.

Reinforcement learning with human feedback (RLHF) has emerged as a promising technique to infuse large language models with domain knowledge and human preferences for healthcare applications. Sun et al. (2023) proposed an approach to continuously improve a conversational agent for behavioral interventions by integrating few-shot generation, prompt engineering, and RLHF to leverage human feedback from therapists and clients. Giuffrè et al. (2024) discussed strategies to optimize large language models for digestive disease by using RLHF to infuse domain knowledge through supervised fine-tuning. Basit et al. (2024) introduced MedAide, an on-premise healthcare chatbot that employs RLHF during training to enhance its medical diagnostic capabilities on edge devices. Dai et al. (2023) presented Safe RLHF, a novel algorithm that decouples human preferences for helpfulness and harmlessness during RLHF to improve the safety and value alignment of large language models in sensitive healthcare domains.

The relevant works discussed in this section is summarized in Table 4.

5 MML—Challenges and opportunities

Learning from multimodal oncology data is a complex and rapidly growing field that presents both challenges and opportunities. While MML has shown significant promise, there are many challenges owing to the inductive biases of the ML models (Ektefaie et al., 2023). In this context, we present major challenges of MML in oncology settings

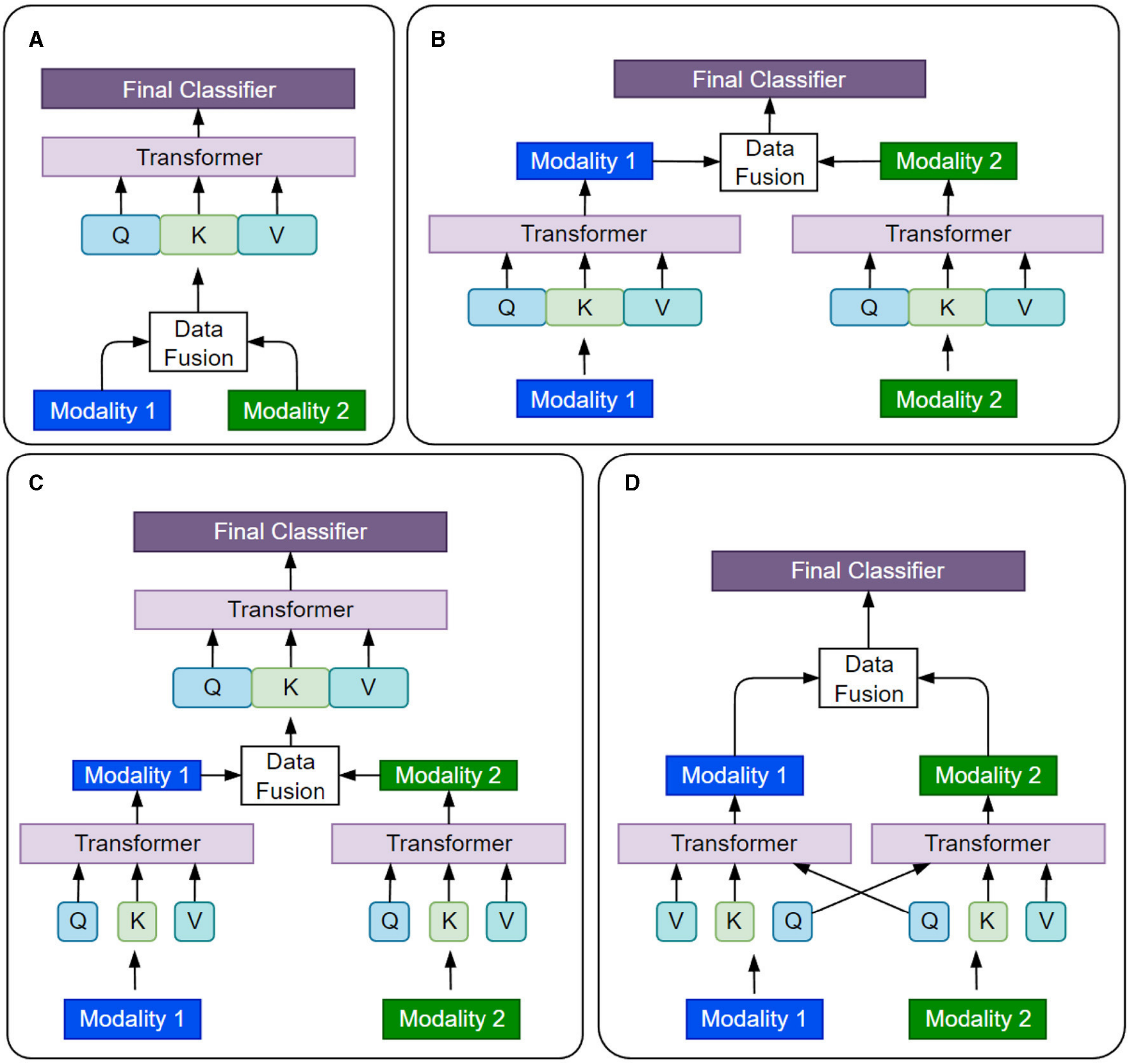

FIGURE 9
Four different strategies of fusing information from various data modalities in multimodal Transformers are presented. **(A)** Early Fusion. **(B)** late Fusion. **(C)** Hierarchial Attention. **(D)** Cross Attention.

TABLE 4 References discussed in Section 4.

Sections	References
Multimodal transformers	Vaswani et al., 2017; Radford et al., 2018, 2021; Devlin et al., 2019; Lewis et al., 2019; Dosovitskiy et al., 2020; Raffel et al., 2020; Zhang, 2020; Boehm et al., 2021; Galassi et al., 2021; Hu and Singh, 2021; Jaegle et al., 2021; Jia et al., 2021; Ma J. et al., 2021; Otter et al., 2021; Zhuang et al., 2021; Ahmed et al., 2022b; Alayrac et al., 2022; Scao et al., 2022; Tang et al., 2022; Yu et al., 2022; Achiam et al., 2023; Han et al., 2023; Xu et al., 2023
MML data fusion stages	Kalfaoglu et al., 2020; Yang et al., 2020; Li et al., 2021a; Sun et al., 2021; Rombach et al., 2022; Zhao et al., 2022; Zhu et al., 2022; Zhong et al., 2023
Transformers for oncology datasets	Boehm et al., 2021; Chen et al., 2021; Shao et al., 2021; Lian et al., 2022; Liang J. et al., 2022; Dai et al., 2023; Sun et al., 2023; Basit et al., 2024; Giuffrè et al., 2024

that, if addressed, could unlock the full potential of this emerging field.

5.1 Large amounts of high-quality data

DL models are traditionally trained on large datasets with enough samples for training, validation, and testing, such as JFT-300M (Sun et al., 2017) and YFCC100M (Thomee et al., 2016), which are not available in the cancer domain. For example, the largest genomics data repository, the Gene Expression Omnibus (GEO) database, has approximately 1.1 million samples with the keyword 'cancer' compared to 3 billion images in JFT-300M (Jiang et al., 2022). Annotating medical and oncology data is a time-consuming and manual process that requires significant expertise in many different areas of medical sciences. Factors like heterogeneity of the disease, noise in data recording, background, and training of medical professionals leading to inter- and intra-operator variability cause lack of reproducibility and inconsistent clinical outcomes (Lipkova et al., 2022).

5.2 Data registration and alignment

Data alignment and registration refer to the process of combining and aligning data from different modalities in a useful manner (Zhao et al., 2023). In multimodal oncology data, this process involves aligning data from multiple modalities such as CT, MRI, PET, and WSIs, as well as genomics, transcriptomics, and clinical records. Data registration involves aligning the data modalities to a common reference frame and may involve identifying common landmarks or fiducial markers. If the data is not registered or aligned correctly, it may be difficult to fuse the information from different modalities (Liang P. P. et al., 2022).

5.3 Pan-cancer generalization and transference

Transference in MML aims to transfer knowledge between modalities and their representations to improve the performance of a model trained on a primary modality (Liang P. P. et al., 2022). Because of the unique characteristics of each cancer type and site, it is challenging to develop models that can generalize across different cancer sites. Furthermore, models trained on a specific modality, such as radiology images, will not perform well with other imaging modalities, such as histopathology slides. Fine-tuning the model on a secondary modality, multimodal co-learning, and model induction are techniques to achieve transference and generalization (Wei et al., 2020). To overcome this challenge, mechanisms for improved universality of ML models need to be devised.

5.4 Missing data samples and modalities

The unavailability of one or more modalities or the absence of samples in a modality affects the model learning, as most of the existing DL models cannot process the “missing information”. This requirement, in turn, constrains the already insufficient size of datasets in oncology. Almost all publicly available oncology datasets have missing data for a large number of samples (Jiang et al., 2022). Various approaches for handling missing data samples and modalities are gradually gaining traction. However, this is still an open challenge (Mirza et al., 2019).

5.5 Imbalanced data

Class imbalance refers to the phenomenon when one class (e.g., cancer negative/positive) is represented significantly more in the data than another class. Class imbalance is common in oncology data (Mirza et al., 2019). DL models struggle to classify underrepresented classes accurately. Techniques such as data augmentation, ensemble, continual learning, and transfer learning are used to counter the class imbalance challenge (Mirza et al., 2019).

5.6 Explainability and trustworthiness

The explainability in DL, e.g., how GNNs and Transformers make a specific decision, is still an area of active research (Li P. et al., 2022; Nielsen et al., 2022). GNNExplainer (Ying et al., 2019), PGM-Explainer (Vu and Thai, 2020), and SubgraphX (Yuan et al., 2021) are some attempts to explain the decision-making process of GNNs. The explainability methods for Transformers have been analyzed in Remmer (2022). Existing efforts and a roadmap to improve the trustworthiness of GNNs have been presented in the latest survey (Zhang H. et al., 2022). However, the explainability and trustworthiness of multimodal GNNs and Transformers is an open challenge.

5.7 Over-smoothing in GNNs

One particular challenge in using GNNs is over-smoothing, which occurs when the GNN is trained for too long, causing the node representations to become almost similar (Wu et al., 2020). This leads to a loss of information, a decrease in the model’s performance, and a lack of generalization (Valsesia et al., 2021). Regularization techniques such as dropout, weight decay, skip-connection, and incorporating higher-order structures, such as motifs and graphlets, have been proposed. However, building deep architectures that can scale and adapt to varying structural patterns of graphs is still an open challenge.

5.8 Modality collapse

Modality collapse is a phenomenon that occurs in MML, where a model trained on multiple modalities may become over-reliant on a single modality, to the point where it ignores or neglects the other modalities (Javaloy et al., 2022). Recent work explored the reasons and theoretical understanding of modality collapse (Huang et al., 2022). However, the counter-actions needed to balance model dependence on data modalities require active investigation by the ML community.

5.9 Dynamic and temporal data

Dynamic and temporal data refers to the data that changes over time (Wu et al., 2020). Tumor surveillance is a well-known technique to study longitudinal cancer growth over multiple data modalities (Waqas et al., 2021). Spatio-temporal methods such as multiple instance learning, GNNs, and hybrid of multiple models can capture complex change in the data relationships over time; however, learning from multimodal dynamic data is very challenging and an active area of research (Fritz et al., 2022).

5.10 Data privacy

Given the sensitive nature of medical data, privacy and security are critical considerations in the development and deployment of MML models for oncology applications. With the increased adoption of MML in healthcare settings, it is essential to adapt these techniques to enable local data processing and protect patient privacy while fostering collaborative research and analysis across different sites and institutions. Federated learning (FL) has emerged as a promising approach to train large multimodal models across various sites without the need for direct data sharing (Pati et al., 2022). In an FL setup, each participating site trains a local model on its own data and shares only the model updates with a central server, which aggregates the updates and sends the updated global model back to the sites. This allows for collaborative model development while keeping the raw data securely within each site's premises.

To further enhance privacy protection in FL and other distributed learning scenarios, differential privacy (DP) can be integrated into the model training process. DP is a rigorous mathematical framework that involves adding carefully calibrated noise to data or model updates before sharing, in order to protect individual privacy while preserving the utility of the data for analysis (Akter et al., 2022; Islam et al., 2022; Nampalle et al., 2023). Secure multi-party computation (SMPC) is another powerful technique for enabling joint analysis and model training on private datasets held by different healthcare providers or research institutions, without revealing the raw data to each other (Şahinbaş and Catak, 2021; Alghamdi et al., 2023; Yogi and Mundru, 2024). SMPC protocols leverage advanced cryptographic techniques to allow multiple parties to compute a function over their combined data inputs securely, such that each party learns only the output of the computation and nothing about the other parties' inputs. In addition to these solutions, implementing appropriate access control and authentication mechanisms is crucial for restricting access to sensitive healthcare data to only authorized individuals and entities (Orii et al., 2024). This involves defining and enforcing strict policies and procedures for granting, managing, and revoking access privileges based on the principle of least privilege and the need-to-know basis. Regular security risk assessments should also be conducted to identify and mitigate potential vulnerabilities proactively, ensuring the ongoing protection of patient data.

5.11 Other challenges

MML requires extensive computational resources to train models on a variety of datasets and tasks. Robustness and failure detection (Ahmed et al., 2022a) are critical aspects of MML, particularly in applications such as oncology. Uncertainty quantification techniques, such as Bayesian neural networks (Dera et al., 2021), are still under-explored avenues in the MML. By addressing these challenges, it is possible to develop MML models that are able to surpass the performance offered by single-modality models.

5.12 Potential future directions

The future of MML in oncology holds immense potential. A critical direction is the integration of large amounts of high-quality data from diverse modalities, such as imaging, genomic, and clinical data, to enhance the accuracy and comprehensiveness of cancer diagnostics and treatment predictions in an end-to-end manner. Overcoming challenges in data registration and alignment is crucial to ensure seamless integration and accurate interpretation of multimodal data. Developing robust models capable of pan-cancer generalization and transference can enable more universal applications across different cancer types. Addressing issues of missing data samples and modalities, and tackling imbalanced datasets, will be essential to improve model robustness and fairness. Enhancing explainability and trustworthiness in these models is vital for clinical adoption, necessitating transparent and interpretable AI systems. Preventing modality collapse is important for maintaining the distinct contributions of each data modality. Moreover, leveraging dynamic and temporal data can offer deeper insights into cancer progression and treatment responses. Ensuring data privacy and ethical considerations will be paramount as the field advances, balancing innovation with the protection of patient information. Lastly, implementing MML applications in clinical settings is crucial to fully realize the benefits of MML in cancer research.

5.13 Limitations of the study

MML is a broad research field that has recently gained traction. In this review, we have focused on the application of MML on oncology data. However, MML is widely being adopted in applications such as autonomous vehicles, education, earth science, climate change, and space exploration (Xiao et al., 2020; Sanders et al., 2023; Hadid et al., 2024; Li et al., 2024). Moreover, beyond GNNs and Transformers, MML has been explored using encoder-decoder methods, constraint-based methods, canonical correlations, Restricted Boltzmann Machines (RBMs), and many more (Qi et al., 2020; Zhao et al., 2024). Each of these topics require an extensive review of the literature in the form of separate articles.

The relevant works discussed in this section is summarized in Table 5.

6 Multimodal oncology data sources

Unifying the various collections of oncology data into central archives necessitates a focused effort. We have assembled a list of datasets from data portals maintained by the National Institute of Health and other organizations, although this list is not exhaustive. The goal of this compilation is to offer machine learning researchers in oncology a consolidated data resource. The collection, which is updated regularly, can be accessed at https://lab-rasool.github.io/pan-cancer-dataset-sources/ (Tripathi et al., 2024a). The compilation of pan-cancer datasets from sources such as The Cancer Imaging Archive (TCIA), Genomic Data Commons (GDC), and Proteomic Data Commons (PDC) serves as a valuable resource for cancer research. By providing a unified view

TABLE 5 References discussed in Section 5.

Sections	References
Large amounts of high-quality data	Thomee et al., 2016; Sun et al., 2017; Lipkova et al., 2022; Ektefaie et al., 2023
Data registration and alignment	Liang P. P. et al., 2022; Zhao et al., 2023
Pan-cancer generalization and transference	Wei et al., 2020; Liang P. P. et al., 2022
Missing data samples and modalities	Mirza et al., 2019; Jiang et al., 2022
Imbalanced Data	Mirza et al., 2019
Explainability and trustworthiness	Ying et al., 2019; Vu and Thai, 2020; Yuan et al., 2021; Li P. et al., 2022; Nielsen et al., 2022; Remmer, 2022; Zhang H. et al., 2022
Over-smoothing in GNNs	Wu et al., 2020; Valsesia et al., 2021
Modality Collapse	Huang et al., 2022; Javaloy et al., 2022
Dynamic and Temporal Data	Wu et al., 2020; Waqas et al., 2021; Fritz et al., 2022
Data Privacy	Şahinbaş and Catak, 2021; Akter et al., 2022; Islam et al., 2022; Pati et al., 2022; Alghamdi et al., 2023; Nampalle et al., 2023; Orii et al., 2024; Yogi and Mundru, 2024
Other Challenges	Dera et al., 2021; Ahmed et al., 2022a
Limitations of the Study	Qi et al., 2020; Xiao et al., 2020; Sanders et al., 2023; Hadid et al., 2024; Li et al., 2024; Zhao et al., 2024

of multimodal data that includes imaging, genomics, proteomics, and clinical records, this compilation facilitates the development of adaptable and scalable datasets specifically designed for machine learning applications in oncology (Tripathi et al., 2024a). The compiled datasets encompass a broad spectrum of data modalities, such as radiology images (CT, MRI, PET), pathology slides, genomic data (DNA, RNA), proteomics, and clinical records. This multimodal nature enables the integration of different data types to capture the intricacies of cancer. Moreover, the compilation covers 32 cancer types, ranging from prevalent cancers like breast, lung, and colorectal to less common forms such as mesothelioma and uveal melanoma. The inclusion of hundreds to thousands of cases for each cancer type provides a substantial resource for training machine learning models, especially deep learning algorithms.

Standardizing the diverse data formats, annotations, and metadata across different sources is essential for creating datasets that are suitable for machine learning. The HoneyBee framework, a modular system designed to streamline the creation of machine learning-ready multimodal oncology datasets from diverse sources, can help address this challenge (Tripathi et al., 2024b). HoneyBee supports data ingestion from various sources, handles different data formats and modalities, and ensures consistent data representation. It also facilitates the integration of multimodal data, enabling the creation of datasets that combine imaging, genomics, proteomics, and clinical data for a holistic view of each patient case. Furthermore, HoneyBee incorporates pre-trained foundational embedding models for different data modalities, such as image encoders, genomic sequence embedders, and clinical text encoders. These embeddings can serve as input features for downstream machine learning models, leveraging transfer learning and reducing the need for extensive labeled data. The framework's scalable and modular architecture allows for efficient processing of large-scale datasets and easy integration of new data sources, preprocessing techniques, and embedding models. By utilizing the HoneyBee framework, researchers can create high-quality, multimodal oncology datasets tailored to their specific research objectives, promoting collaboration and advancing machine learning applications in cancer research.

7 Conclusion

Existing research into the integration of data across various modalities has already yielded promising outcomes, highlighting the potential for significant advancements in cancer research. However, the lack of a comprehensive framework capable of encompassing the full spectrum of cancer dataset modalities presents a notable challenge. The synergy between diverse methodologies and data across different scales could unlock deeper insights into cancer, potentially leading to more accurate prognostic and predictive models than what is possible through single data modalities alone. In our survey, we have explored the landscape of multimodal learning applied to oncology datasets and the specific tasks they can address. Looking ahead, the key to advancing this field lies in the development of robust, deployment-ready MML frameworks. These frameworks must not only scale efficiently across all modalities of cancer data but also incorporate capabilities for uncertainty quantification, interpretability, and generalizability. Such advancements will be critical for effectively integrating oncology data across multiple scales, modalities, and resolutions. The journey toward achieving these goals is complex, yet essential for the next leaps in cancer research. By focusing on these areas, future research has the potential to significantly enhance our understanding of cancer, leading to improved outcomes for patients through more informed and personalized treatment strategies.

Author contributions

AW: Conceptualization, Writing – original draft, Writing – review & editing. AT: Conceptualization, Writing – original draft, Writing – review & editing. RR: Conceptualization, Writing – original draft, Writing – review & editing. PS: Conceptualization, Writing – original draft, Writing – review & editing. GR: Conceptualization, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare financial support was received for the research, authorship, and/or publication of this article. This work was partly supported by the National Science Foundation awards 1903466, 2008690, 2234836, and 2234468, and partly by the Biostatistics and Bioinformatics Shared Resource at the H. Lee

Moffitt Cancer Center & Research Institute, an NCI designated Comprehensive Cancer Center (P30-CA076292).

Conflict of interest

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

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., et al. (2023). Gpt-4 technical report. *arXiv* [preprint] arXiv:2303.08774. doi: 10.48550/arXiv.2303.08774

Ahmed, S., Dera, D., Hassan, S. U., Bouaynaya, N., and Rasool, G. (2022a). Failure detection in deep neural networks for medical imaging. *Front. Med. Technol.* 4:919046. doi: 10.3389/fmedt.2022.919046

Ahmed, S., Nielsen, I. E., Tripathi, A., Siddiqui, S., Rasool, G., and Ramachandran, R. P. (2022b). Transformers in time-series analysis: a tutorial. *arXiv* [preprint] arXiv:2205.01138. doi: 10.1007/s00034-023-02454-8

Ahmedt-Aristizabal, D., Armin, M. A., Denman, S., Fookes, C., and Petersson, L. (2022). A survey on graph-based deep learning for computational histopathology. *Comp. Med. Imag. Graph.* 95:102027. doi: 10.1016/j.compmedimag.2021.102027

Akter, M., Moustafa, N., and Lynar, T. (2022). "Edge intelligence-based privacy protection framework for iot-based smart healthcare systems,"? in *IEEE INFOCOM 2022-IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)* (New York, NY: IEEE), 1–8.

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., et al. (2022). Flamingo: a visual language model for few-shot learning. *Adv. Neural Inf. Process. Syst.* 35, 23716–23736. Available online at: https://proceedings.neurips.cc/paper_files/paper/2022/hash/960a172bc7fbf0177ccccbb411a7d800-Abstract-Conference.html

Alghamdi, W., Salama, R., Sirija, M., Abbas, A. R., and Dilnoza, K. (2023). "Secure multi-party computation for collaborative data analysis,"? in *E3S Web of Conferences* (Les Ulis: EDP Sciences), 04034.

Al-jabery, K. K., Obafemi-Ajayi, T., Olbricht, G. R., and Wunsch, I. I., D. C. (2020). "Data preprocessing,"? in K. K. Al-jabery, T. Obafemi-Ajayi, G. R. Olbricht and D. C. Wunsch *Computational Learning Approaches to Data Analytics in Biomedical Applications* (Cambridge, MA: Academic Press), 7–27.

Almasan, P., Suárez-Varela, J., Rusek, K., Barlet-Ros, P., and Cabellos-Aparicio, A. (2022). Deep reinforcement learning meets graph neural networks: Exploring a routing optimization use case. *Comput. Commun.* 196, 184–194. doi: 10.1016/j.comcom.2022.09.029

Anand, D., Gadiya, S., and Sethi, A. (2020). "Histographs: graphs in histopathology,"? in *Medical Imaging 2020: Digital Pathology* (California: SPIE), 150-155.

Angermueller, C., Lee, H. J., Reik, W., and Stegle, O. (2017). DeepCpG: accurate prediction of single-cell DNA methylation states using deep learning. *Genome Biol.* 18, 1–13. doi: 10.1186/s13059-017-1189-z

Asan, O., Nattinger, A. B., Gurses, A. P., Tyszka, J. T., and Yen, T. W. (2018). Oncologists' views regarding the role of electronic health records in care coordination. *JCO Clini. Cancer Inform.* 2, 1–12. doi: 10.1200/CCI.17.00118

Bai, S., Zhang, F., and Torr, P. H. (2021). Hypergraph convolution and hypergraph attention. *Pattern Recognit.* 110:107637. doi: 10.1016/j.patcog.2020.107637

Baltrušaitis, T., Ahuja, C., and Morency, L.-P. (2018). Multimodal machine learning: a survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.* 41, 423–443. doi: 10.1109/TPAMI.2018.2798607

Barhoumi, Y., Bouaynaya, N. C., and Rasool, G. (2023). Efficient scopeformer: towards scalable and rich feature extraction for intracranial hemorrhage detection. *IEEE Access.* 11, 81656–81671. doi: 10.1109/ACCESS.2023.3301160

Basit, A., Hussain, K., Hanif, M. A., and Shafique, M. (2024). Medaide: Leveraging large language models for on-premise medical assistance on edge devices. *arXiv* [preprint] arXiv:2403.00830. doi: 10.48550/arXiv.2403.00830

Boehm, K. M., Khosravi, P., Vanguri, R., Gao, J., and Shah, S. P. (2021). Harnessing multimodal data integration to advance precision oncology. *Nat. Rev. Cancer* 22, 114–126. doi: 10.1038/s41568-021-00408-3

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., et al. (2022). *On the Opportunities and Risks of Foundation Models*. arXiv preprint.

Borisov, V., Leemann, T., Seßler, K., Haug, J., Pawelczyk, M., and Kasneci, G. (2022). Deep neural networks and tabular data: a survey. *IEEE Trans. Neural Netw. Learn. Syst.* 35, 1–21. doi: 10.1109/TNNLS.2022.3229161

Çalışkan, M., and Tazaki, K. (2023). Ai/ml advances in non-small cell lung cancer biomarker discovery. *Front. Oncol.* 13:1260374. doi: 10.3389/fonc.2023.1260374

Cao, Z.-J., and Gao, G. (2022). Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. *Nat. Biotechnol.* 40, 1458–1466. doi: 10.1038/s41587-022-01284-4

Chan, H.-P., Hadjiiski, L. M., and Samala, R. K. (2020). Computer-aided diagnosis in the era of deep learning. *Med. Phys.* 47, e218–e227. doi: 10.1002/mp.13764

Chao, C.-H., Zhu, Z., Guo, D., Yan, K., Ho, T.-Y., Cai, J., et al. (2020). "Lymph node gross tumor volume detection in oncology imaging via relationship learning using graph neural network," in *Medical Image Computing and Computer Assisted Intervention-MICCAI 2020: 23rd International Conference, Lima, Peru, October 4-8, 2020, Proceedings, Part VII 23* (Cham: Springer), 772–782.

Chatzianastasis, M., Vazirgiannis, M., and Zhang, Z. (2023). Explainable multilayer graph neural network for cancer gene prediction. *arXiv* [preprint] arXiv:2301.08831. doi: 10.1093/bioinformatics/btad643

Chen, B., Jin, J., Liu, H., Yang, Z., Zhu, H., Wang, Y., et al. (2023). Trends and hotspots in research on medical images with deep learning: a bibliometric analysis from 2013 to 2023. *Front. Artif. Intellig.* 6:1289669. doi: 10.3389/frai.2023.1289669

Chen, R. J., Lu, M. Y., Wang, J., Williamson, D. F., Rodig, S. J., Lindeman, N. I., et al. (2020). Pathomic Fusion: an integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis. *IEEE Trans. Med. Imaging* 41, 757–770. doi: 10.1109/TMI.2020.3021387

Chen, R. J., Lu, M. Y., Weng, W.-H., Chen, T. Y., Williamson, D. F., Manz, T., et al. (2021). "Multimodal co-attention transformer for survival prediction in gigapixel whole slide images," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, (Montreal, QC: IEEE), 3995–4005.

Choi, E., Bahadori, M. T., Song, L., Stewart, W. F., and Sun, J. (2017). "GRAM: graph-based attention model for healthcare representation learning," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM), 787–795.

Cui, H., Xuan, P., Jin, Q., Ding, M., Li, B., Zou, B., et al. (2021). "Co-graph attention reasoning based imaging and clinical features integration for lymph node metastasis prediction," in *Medical Image Computing and Computer Assisted Intervention-MICCAI 2021: 24th International Conference, Strasbourg, France, September 27-October 1, 2021, Proceedings, Part V 24* (Cham: Springer), 657–666.

Dai, J., Pan, X., Sun, R., Ji, J., Xu, X., Liu, M., et al. (2023). Safe rlhf: Safe reinforcement learning from human feedback. *arXiv* [preprint] arXiv:2310.12773.

Dara, S., and Tumma, P. (2018). "Feature extraction by using deep learning: a survey," in *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)* (Coimbatore: IEEE).

Defferrard, M., Bresson, X., and Vandergheynst, P. (2016). Convolutional neural networks on graphs with fast localized spectral filtering. *Adv. Neural Inf. Process. Syst.* 29, 3844–3852. Available online at: https://proceedings.neurips.cc/paper_files/paper/2016/file/04df4d434d481c5bb723be1b6df1ee65-Paper.pdf

Dera, D., Bouaynaya, N. C., Rasool, G., Shterenberg, R., and Fathallah-Shaykh, H. M. (2021). PremiUm-CNN: propagating uncertainty towards robust convolutional neural networks. *IEEE Trans. Signal Proc.* 69, 4669–4684. doi: 10.1109/TSP.2021.3096804

Dera, D., Rasool, G., and Bouaynaya, N. (2019). "Extended variational inference for propagating uncertainty in convolutional neural networks," in *2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (MLSP)* (Pittsburgh, PA: IEEE), 1–6.

Derrow-Pinion, A., She, J., Wong, D., Lange, O., Hester, T., Perez, L., et al. (2021). "ETA prediction with graph neural networks in google maps," in *Proceedings of the 30th ACM International Conference on Information & Knowledge Management* (New York, NY: ACM), 3767–3776.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). *Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding*. Minneapolis: North American Chapter of the Association for Computational Linguistics.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2020). *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. Vienna: ICLR.

Du, H., Feng, J., and Feng, M. (2019). Zoom in to where it matters: a hierarchical graph based model for mammogram analysis. *arXiv* [preprint] arXiv:1912.07517. doi: 10.48550/arXiv.1912.07517

Ektefaie, Y., Dasoulas, G., Noori, A., Farhat, M., and Zitnik, M. (2023). Multimodal learning with graphs. *Nat. Mach. Intell.* 5, 340–350. doi: 10.1038/s42256-023-00624-6

Farooq, H., Chen, Y., Georgiou, T. T., Tannenbaum, A., and Lenglet, C. (2019). Network curvature as a hallmark of brain structural connectivity. *Nat. Commun.* 10, 1–11. doi: 10.1038/s41467-019-12915-x

Fathinezhad, F., Adibi, P., Shoushtarian, B., and Chanussot, J. (2023). *Graph Neural Networks and Reinforcement Learning: A Survey*. London: IntechOpen.

Feng, Y.-H., Zhang, S.-W., and Shi, J.-Y. (2020). DPDDI: a deep predictor for drug-drug interactions. *BMC Bioinformat.* 21, 1–15. doi: 10.1186/s12859-020-03724-x

Fout, A., Byrd, J., Shariat, B., and Ben-Hur, A. (2017). "Protein interface prediction using graph convolutional networks," in *Adv. Neural Inf. Process. Syst* (NeurIPS), 30.

Fritz, C., Dorigatti, E., and Rügamer, D. (2022). Combining graph neural networks and spatio-temporal disease models to improve the prediction of weekly covid-19 cases in germany. *Sci. Rep.* 12, 3930. doi: 10.1038/s41598-022-07757-5

Galassi, A., Lippi, M., and Torroni, P. (2021). Attention in natural language processing. *IEEE Trans. Neural Netw. Learn. Syst.* 32, 4291–4308. doi: 10.1109/TNNLS.2020.3019893

Ghaffari Laleh, N., Ligero, M., Perez-Lopez, R., and Kather, J. N. (2023). Facts and hopes on the use of artificial intelligence for predictive immunotherapy biomarkers in cancer. *Clini.Cancer Res.* 29, 316–323. doi: 10.1158/1078-0432.CCR-22-0390

Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., and Dahl, G. E. (2017). "Neural message passing for quantum chemistry," in *International Conference on Machine Learning* (New York: PMLR), 1263–1272..

Giuffrè, M., Kresevic, S., Pugliese, N., You, K., and Shung, D. L. (2024). "Optimizing large language models in digestive disease: strategies and challenges to improve clinical outcomes," in *Liver International* (Wiley).

Gonzalez Zelaya, C. V. (2019). "Towards explaining the effects of data preprocessing on machine learning," in *IEEE 35th International Conference on Data Engineering (ICDE)* (Macao: IEEE), 2086–2090.

Grossman, R. L., Heath, A. P., Ferretti, V., Varmus, H. E., Lowy, D. R., Kibbe, W. A., et al. (2016). Toward a shared vision for cancer genomic data. *New Engl. J. Med.* 375, 1109–1112. doi: 10.1056/NEJMp1607591

Hadid, A., Chakraborty, T., and Busby, D. (2024). When geoscience meets generative ai and large language models: foundations, trends, and future challenges. *Expert Syst.* 2024:e13654. doi: 10.1111/exsy.13654

Hamilton, W., Ying, Z., and Leskovec, J. (2017). "Inductive representation learning on large graphs," in *Adv. Neural Inf. Process. Syst* (NeurIPS), 30.

Hamilton, W. L. (2020). Graph representation learning. *Synth. Lect. Artif. Intellig. Mach. Learn.* 14:5. doi: 10.1007/978-3-031-01588-5

Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., et al. (2023). A survey on vision transformer. *IEEE Trans. Pattern Anal. Mach. Intell.* 45, 87–110. doi: 10.1109/TPAMI.2022.3152247

Hartsock, I., and Rasool, G. (2024). Vision-language models for medical report generation and visual question answering: a review. *arXiv* [preprint] arXiv:2403.02469. doi: 10.48550/arXiv.2403.02469

Hook, D. W., Porter, S. J., and Herzog, C. (2018). Dimensions: building context for search and evaluation. *Front. Res. Metrics Anal.* 3:23. doi: 10.3389/frma.2018.00023

Hu, R., and Singh, A. (2021). "Unit: Multimodal multitask learning with a unified transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (Montreal, QC: IEEE), 1439–1449.

Huang, Y., and Chung, A. C. (2020). "Edge-variational graph convolutional networks for uncertainty-aware disease prediction," in *Medical Image Computing and Computer Assisted Intervention-MICCAI 2020: 23rd International Conference, Lima, Peru, October 4-8, 2020, Proceedings, Part VII 23* (Cham: Springer), 562–572.

Huang, Y., Du, C., Xue, Z., Chen, X., Zhao, H., and Huang, L. (2021). *What Makes Multi-Modal Learning Better Than Single (Provably)*. New Orleans, LA: Advances in Neural Information Processing Systems.

Huang, Y., Lin, J., Zhou, C., Yang, H., and Huang, L. (2022). "Modality competition: What makes joint training of multi-modal network fail in deep learning? (provably)," in *International Conference on Machine Learning* (New York: PMLR), 9226–9259.

Ibrahim, A., Mohamed, H. K., Maher, A., and Zhang, B. (2022). A survey on human cancer categorization based on deep learning. *Front. Artif. Intellig.* 5:884749. doi: 10.3389/frai.2022.884749

Iqbal, M. S., Ahmad, W., Alizadehsani, R., Hussain, S., and Rehman, R. (2022). Breast cancer dataset, classification and detection using deep learning. *Healthcare.* 10:2395. doi: 10.3390/healthcare10122395

Iqbal, M. S., Luo, B., Mehmood, R., Alrige, M. A., and Alharbey, R. (2019). Mitochondrial organelle movement classification (fission and fusion) via convolutional neural network approach. *IEEE Access* 7:86570–86577. doi: 10.1109/ACCESS.2019.2925041

Islam, T. U., Ghasemi, R., and Mohammed, N. (2022). "Privacy-preserving federated learning model for healthcare data," in *2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC)* (Las Vegas, NV: IEEE), 0281–0287.

Jaegle, A., Borgeaud, S., Alayrac, J.-B., Doersch, C., Ionescu, C., Ding, D., et al. (2021). *Perceiver IO: A General Architecture for Structured Inputs* &; *Outputs* (ICLR).

Jansen, C., Ramirez, R. N., El-Ali, N. C., Gomez-Cabrero, D., Tegner, J., Merkenschlager, M., et al. (2019). Building gene regulatory networks from scATAC-seq and scRNA-seq using linked self organizing maps. *PLoS Comput. Biol.* 15:e1006555. doi: 10.1371/journal.pcbi.1006555

Javaloy, A., Meghdadi, M., and Valera, I. (2022). "Mitigating modality collapse in multimodal VAEs via impartial optimization," in *International Conference on Machine Learning* (New York: PMLR), 9938–9964.

Ji, Y., Zhou, Z., Liu, H., and Davuluri, R. V. (2020). DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *bioRxiv*. doi: 10.1101/2020.09.17.301879

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., et al. (2021). "Scaling up visual and vision-language representation learning with noisy text supervision," in *International Conference on Machine Learning* (New York: PMLR), 4904–4916.

Jiang, J., Dun, C., Huang, T., and Lu, Z. (2018). Graph convolutional reinforcement learning. *arXiv* [preprint] arXiv:1810.09202.

Jiang, P., Sinha, S., Aldape, K., Hannenhalli, S., Sahinalp, C., and Ruppin, E. (2022). Big Data in basic and translational cancer research. *Nat. Rev. Cancer* 22, 625–639. doi: 10.1038/s41568-022-00502-0

Jiao, L., Chen, J., Liu, F., Yang, S., You, C., Liu, X., et al. (2022). Graph representation learning meets computer vision: a survey. *IEEE Trans. Artif. Intellig.* 4, 2–22. doi: 10.1109/TAI.2022.3194869

Jin, D., Huo, C., Dang, J., Zhu, P., Zhang, W., Pedrycz, W., et al. (2022). Heterogeneous graph neural networks using self-supervised reciprocally contrastive learning. *arXiv* [preprint] arXiv:2205.00256.

Joo, S., Ko, E., Kwon, S., Jeon, E., Jung, H., Kim, J., et al. (2021). Multimodal deep learning models for the prediction of pathologic response to neoadjuvant chemotherapy in breast cancer. *Sci. Rep.* 11:18800. doi: 10.1038/s41598-021-98408-8

Kaczmarek, E., Jamzad, A., Imtiaz, T., Nanayakkara, J., Renwick, N., and Mousavi, P. (2021). "Multi-omic graph transformers for cancer classification and interpretation," in *Pacific Symposium On Biocomputing 2022* (Singapore: World Scientific), 373–384.

Kalfaoglu, M. E., Kalkan, S., and Alatan, A. A. (2020). "Late Temporal Modeling in 3D CNN Architectures with BERT for Action Recognition," in *Computer Vision-ECCV 2020 Workshops: Glasgow, UK, August 23-28, 2020, Proceedings, Part V 16* (Cham: Springer), 731–747.

Khan, M., Ashraf, I., Alhaisoni, M., Damaševičius, R., Scherer, R., Rehman, A., et al. (2020). Multimodal brain tumor classification using deep learning and robust feature selection: a machine learning application for radiologists. *Diagnostics* 10:565. doi: 10.3390/diagnostics10080565

Khan, S., Ali, H., and Shah, Z. (2023). Identifying the role of vision transformer for skin cancer–a scoping review. *Front. Artif. Intellig.* 6:1202990. doi: 10.3389/frai.2023.1202990

Kipf, T. N., and Welling, M. (2016). Semi-supervised classification with graph convolutional networks. *arXiv* [preprint] arXiv:1609.02907.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444. doi: 10.1038/nature14539

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., et al. (2019). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 36, 1234–1240. doi: 10.1093/bioinformatics/btz682

Leng, D., Zheng, L., Wen, Y., Zhang, Y., Wu, L., Wang, J., et al. (2022). A benchmark study of deep learning-based multi-omics data fusion methods for cancer. *Genome Biol.* 23, 1–32. doi: 10.1186/s13059-022-02739-2

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., et al. (2019). *BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension*. Dublin: Annual Meeting of the Association for Computational Linguistics.

Li, M. M., Huang, K., and Zitnik, M. (2022). Graph representation learning in biomedicine and healthcare. *Nat. Biomed. Eng.* 6, 1353–1369. doi: 10.1038/s41551-022-00942-x

Li, P., Gu, J., Kuen, J., Morariu, V. I., Zhao, H., Jain, R., et al. (2021a). "Selfdoc: self-supervised document representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Nashville, TN: IEEE), 5652–5660.

Li, P., Wang, J., Qiao, Y., Chen, H., Yu, Y., Yao, X., et al. (2021b). An effective self-supervised framework for learning expressive molecular global representations to drug discovery. *Brief. Bioinform.* 22:bbab109. doi: 10.1093/bib/bbab109

Li, P., Yang, Y., Pagnucco, M., and Song, Y. (2022). Explainability in graph neural networks: An experimental survey. *arXiv* [preprint] arXiv:2203.09258. doi: 10.48550/arXiv.2203.09258

Li, Z., Pardos, Z. A., and Ren, C. (2024). Aligning open educational resources to new taxonomies: How ai technologies can help and in which scenarios. *Comp. Educ.* 216:105027. doi: 10.1016/j.compedu.2024.105027

Lian, J., Deng, J., Hui, E. S., Koohi-Moghadam, M., She, Y., Chen, C., et al. (2022). Early stage NSCLS patients' prognostic prediction with multi-information using transformer and graph neural network model. *Elife* 11:e80547. doi: 10.7554/eLife.80547

Liang, J., Yang, C., Zeng, M., and Wang, X. (2022). TransConver: transformer and convolution parallel network for developing automatic brain tumor segmentation in MRI images. *Quant. Imaging Med. Surg.* 12:4. doi: 10.21037/qims-21-919

Liang, P. P., Zadeh, A., and Morency, L.-P. (2022). Foundations and recent trends in multimodal machine learning: principles, challenges, and open questions. *arXiv* [preprint] arXiv:2209.03430. doi: 10.48550/arXiv.2209.03430

Lipkova, J., Chen, R. J., Chen, B., Lu, M. Y., Barbieri, M., Shao, D., et al. (2022). Artificial intelligence for multimodal data integration in oncology. *Cancer Cell* 40:1095–1110. doi: 10.1016/j.ccell.2022.09.012

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., et al. (2017). A survey on deep learning in medical image analysis. *Med. Image Anal.* 42, 60–88. doi: 10.1016/j.media.2017.07.005

Liu, J., Pandya, P., and Afshar, S. (2021). Therapeutic advances in oncology. *Int. J. Mol. Sci.* 22:2008. doi: 10.3390/ijms22042008

Liu, T., Huang, J., Liao, T., Pu, R., Liu, S., and Peng, Y. (2022). A hybrid deep learning model for predicting molecular subtypes of human breast cancer using multimodal data. *IRBM* 43, 62–74. doi: 10.1016/j.irbm.2020.12.002

Ma, J., Liu, J., Lin, Q., Wu, B., Wang, Y., and You, Y. (2021). Multitask learning for visual question answering. *IEEE Trans. Neural Netw. Learn. Syst.* 34, 1380–1394.

Ma, L., Yang, Z., Miao, Y., Xue, J., Wu, M., Zhou, L., et al. (2019). "NeuGraph: parallel deep neural network computation on large graphs," in *USENIX Annual Technical Conference* (USENIX), 443–458.

Ma, X., and Jia, F. (2020). "Brain tumor classification with multimodal MR and pathology images," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries: 5th International Workshop, BrainLes 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 17, 2019, Revised Selected Papers, Part II 5* (Cham: Springer), 343–352.

Ma, Y., Liu, X., Zhao, T., Liu, Y., Tang, J., and Shah, N. (2021). "A unified view on graph neural networks as graph signal denoising," in *Proceedings of the 30th ACM International Conference on Information & Knowledge Management* (New York, NY: ACM), 1202–1211.

Ma, Y., and Tang, J. (2021). *Deep Learning on Graphs*. Cambridge: Cambridge University Press.

Miotto, R., Li, L., Kidd, B. A., and Dudley, J. T. (2016). Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Sci. Rep.* 6, 1–10. doi: 10.1038/srep26094

Mirza, B., Wang, W., Wang, J., Choi, H., Chung, N. C., and Ping, P. (2019). Machine learning and integrative analysis of biomedical big data. *Genes* 10, 87. doi: 10.3390/genes10020087

Mo, S., Cai, M., Lin, L., Tong, R., Chen, Q., Wang, F., et al. (2020). "Multimodal priors guided segmentation of liver lesions in MRI using mutual information based graph co-attention networks," in *Medical Image Computing and Computer Assisted Intervention-MICCAI 2020: 23rd International Conference, Lima, Peru, October 4-8, 2020, Proceedings, Part IV 23* (Cham: Springer), 429–438.

Muhammad, L. J., and Bria, A. (2023). Editorial: Ai applications for diagnosis of breast cancer. *Front. Artif. Intellig.* 6:1247261. doi: 10.3389/frai.2023.1247261

Muhammad, W., Ahmed, S.-,b,-S., Naeem, S., Khan, A. A. M. H., Qureshi, B. M., Hussain, A., et al. (2024). Artificial neural network-assisted prediction of radiobiological indices in head and neck cancer. *Front. Artif. Intellig.* 7:1329737. doi: 10.3389/frai.2024.1329737

Nampalle, K. B., Singh, P., Narayan, U. V., and Raman, B. (2023). Vision through the veil: Differential privacy in federated learning for medical image classification. *arXiv* [preprint] arXiv:2306.17794. doi: 10.48550/arXiv.2306.17794

Nguyen, T., Le, H., Quinn, T. P., Nguyen, T., Le, T. D., and Venkatesh, S. (2021). GraphDTA: predicting drug-target binding affinity with graph neural networks. *Bioinformatics* 37, 1140–1147. doi: 10.1093/bioinformatics/btaa921

Nie, M., Chen, D., and Wang, D. (2023). Reinforcement learning on graphs: A survey. *IEEE Trans. Emerg. Topics Comp. Intellig.* 7, 1065–1082. doi: 10.1109/TETCI.2022.3222545

Nielsen, I. E., Dera, D., Rasool, G., Ramachandran, R. P., and Bouaynaya, N. C. (2022). Robust explainability: a tutorial on gradient-based attribution methods for deep neural networks. *IEEE Signal Process. Mag.* 39, 73–84. doi: 10.1109/MSP.2022.3142719

Orii, L., Feldacker, C., Tweya, H., and Anderson, R. (2024). ehealth data security and privacy: Perspectives from diverse stakeholders in malawi. *Proc. ACM on Human-Comp. Interact.* 8, 1–26. doi: 10.1145/3637323

Otter, D. W., Medina, J. R., and Kalita, J. K. (2021). A survey of the usages of deep learning for natural language processing. *IEEE Trans. Neural Netw. Learn. Syst.* 32, 604–624. doi: 10.1109/TNNLS.2020.2979670

Park, J., Cho, J., Chang, H. J., and Choi, J. Y. (2021). "Unsupervised hyperbolic representation learning via message passing auto-encoders," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (Nashville, TN: IEEE), 5516–5526.

Park, M.-K., Lim, J.-M., Jeong, J., Jang, Y., Lee, J.-W., Lee, J.-C., et al. (2022). Deep-learning algorithm and concomitant biomarker identification for NSCLC prediction using multi-omics data integration. *Biomolecules* 12:1839. doi: 10.3390/biom12121839

Pati, P., Jaume, G., Fernandes, L. A., Foncubierta-Rodríguez, A., Feroce, F., Anniciello, A. M., et al. (2020). "HACT-net: a hierarchical cell-to-tissue graph neural network for histopathological image classification," in *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging, and Graphs in Biomedical Image Analysis: Second International Workshop, UNSURE 2020, and Third International Workshop, GRAIL 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 8, 2020, Proceedings 2* (Cham: Springer), 208–219.

Pati, S., Baid, U., Edwards, B., Sheller, M., Wang, S.-H., Reina, G. A., et al. (2022). Federated learning enables big data for rare cancer boundary detection. *Nat. Commun.* 13:7346. doi: 10.1038/s41467-022-33407-5

Qi, G., Sun, Y., Li, M., and Hou, X. (2020). Development and application of matrix variate restricted boltzmann machine. *IEEE Access* 8:137856–137866. doi: 10.1109/ACCESS.2020.3012603

Quinn, M., Forman, J., Harrod, M., Winter, S., Fowler, K. E., Krein, S. L., et al. (2019). Electronic health records, communication, and data sharing: challenges and opportunities for improving the diagnostic process. *Diagnosis* 6:241–248. doi: 10.1515/dx-2018-0036

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., et al. (2021). "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning* (New York: PMLR).

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I. (2018). *Improving Language Understanding by Generative Pre-Training.* Available online at: https://www.mikecaptain.com/resources/pdf/GPT-1.pdf

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., et al. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* 21:1.

Rajadhyaksha, N., and Chitkara, A. (2023). Graph contrastive learning for multi-omics data. *arXiv* [preprint] arXiv:2301.02242. doi: 10.48550/arXiv.2301.02242

Rao, J., Zhou, X., Lu, Y., Zhao, H., and Yang, Y. (2021). Imputing single-cell RNA-seq data by combining graph convolution and autoencoder neural networks. *Iscience* 24:102393. doi: 10.1016/j.isci.2021.102393

Rasmy, L., Xiang, Y., Xie, Z., Tao, C., and Zhi, D. (2021). Med-BERT: Pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ Digital Med.* 4:1. doi: 10.1038/s41746-021-00455-y

Remmer, E. (2022). *Explainability Methods for Transformer-based Artificial Neural Networks: a Comparative Analysis* (PhD thesis).

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022). "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (IEEE), 10684–10695.

Rowe, S. P., and Pomper, M. G. (2022). Molecular imaging in oncology: current impact and future directions. *CA Cancer J. Clin.* 72, 333–352. doi: 10.3322/caac.21713

Rozemberczki, B., Gogleva, A., Nilsson, S., Edwards, G., Nikolov, A., and Papa, E. (2022). "MOOMIN: deep molecular omics network for anti-cancer drug combination therapy," in *Proceedings of the 31st ACM International Conference on Information & Knowledge Management* (New York, NY: ACM), 3472–3483.

Şahinbaş, K., and Catak, F. O. (2021). Secure multi-party computation based privacy preserving data analysis in healthcare iot systems. *arXiv* [preprint] arXiv:2109.14334. doi: 10.1007/978-3-031-08637-3_3

Sanders, L. M., Scott, R. T., Yang, J. H., Qutub, A. A., Garcia Martin, H., Berrios, D. C., et al. (2023). Biological research and self-driving labs in deep space supported by artificial intelligence. *Nat. Mach. Intellig.* 5, 208–219. doi: 10.1038/s42256-023-00618-4

Sankar, A., Wu, Y., Gou, L., Zhang, W., and Yang, H. (2018). Dynamic graph representation learning via self-attention networks. *arXiv* [preprint] arXiv:1812.09430.

Saueressig, C., Berkley, A., Kang, E., Munbodh, R., and Singh, R. (2021). "Exploring graph-based neural networks for automatic brain tumor segmentation," in *From Data to Models and Back: 9th International Symposium, DataMod 2020, Virtual Event, October 20, 2020, Revised Selected Papers 9* (Cham: Springer), 18–37.

Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., et al. (2022). BLOOM: A 176B-parameter open-access multilingual language model. *arXiv* [preprint] arXiv:2211.05100.

Schulz, S., Woerl, A.-C., Jungmann, F., Glasner, C., Stenzel, P., Strobl, S., et al. (2021). Multimodal deep learning for prognosis prediction in renal cancer. *Front. Oncol.* 11:788740. doi: 10.3389/fonc.2021.788740

Shang, J., Ma, T., Xiao, C., and Sun, J. (2019). Pre-training of graph augmented transformers for medication recommendation. *arXiv* [preprint] arXiv:1906.00346. doi: 10.24963/ijcai.2019/825

Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., et al. (2021). "TransMIL: transformer based correlated multiple instance learning for whole slide image classification," in *Adv. Neural Inf. Process. Syst*, 34.

Shi, J., Wang, R., Zheng, Y., Jiang, Z., and Yu, L. (2019). "Graph convolutional networks for cervical cell classification," in *MICCAI 2019 Computational Pathology Workshop COMPAY* (Shenzhen: Compay).

Siam, A., Alsaify, A. R., Mohammad, B., Biswas, M. R., Ali, H., and Shah, Z. (2023). Multimodal deep learning for liver cancer applications: a scoping review. *Front. Artif. Intellig.* 6:1247195. doi: 10.3389/frai.2023.1247195

Siegel, R. L., Miller, K. D., Wagle, N. S., and Jemal, A. (2023). Cancer statistics, 2023. *CA Cancer J. Clin.* 73, 17–48. doi: 10.3322/caac.21763

Singh, A., Hu, R., Goswami, V., Couairon, G., Galuba, W., Rohrbach, M., et al. (2022). *FLAVA: A Foundational Language and Vision Alignment Model*. Seattle, WA: CVPR. doi: 10.1109/CVPR52688.2022.01519

Singh, G., Manjila, S., Sakla, N., True, A., Wardeh, A. H., Beig, N., et al. (2021). Radiomics and radiogenomics in gliomas: a contemporary update. *Br. J. Cancer* 125, 641–657. doi: 10.1038/s41416-021-01387-w

Sleeman, I. V., W. C., Kapoor, R., and Ghosh, P. (2022). Multimodal classification: current landscape, taxonomy and future directions. *ACM Comp. Surv.* 55, 1–31. doi: 10.1145/3543848

Song, J., Zheng, Y., Zakir Ullah, M., Wang, J., Jiang, Y., Xu, C., et al. (2021). Multiview multimodal network for breast cancer diagnosis in contrast-enhanced spectral mammography images. *Int. J. Comput. Assist. Radiol. Surg.* 16, 979–988. doi: 10.1007/s11548-021-02391-4

Song, Q., Su, J., and Zhang, W. (2021). scGCN is a graph convolutional networks algorithm for knowledge transfer in single cell omics. *Nat. Commun.* 12, 3826. doi: 10.1038/s41467-021-24172-y

Stark, S. G., Ficek, J., Locatello, F., Bonilla, X., Chevrier, S., Singer, F., et al. (2020). SCIM: universal single-cell matching with unpaired feature sets. *Bioinformatics* 36, i919–927. doi: 10.1093/bioinformatics/btaa843

Sun, C., Shrivastava, A., Singh, S., and Gupta, A. (2017). "Revisiting unreasonable effectiveness of data in deep learning era," in *Proceedings of the IEEE International Conference on Computer Vision* (Venice: IEEE), 843–852.

Sun, H., Liu, J., Chai, S., Qiu, Z., Lin, L., Huang, X., et al. (2021). Multi-Modal Adaptive Fusion Transformer Network for the estimation of depression level. *Sensors* 21:4764. doi: 10.3390/s21144764

Sun, X., Bosch, J. A., De Wit, J., and Krahmer, E. (2023). "Human-in-the-loop interaction for continuously improving generative model in conversational agent for behavioral intervention," in *Companion Proceedings of the 28th International Conference on Intelligent User Interfaces* (New York, NY: ACM), 99–101.

Syed, K., Sleeman, I. V., W. C., Hagan, M., Palta, J., Kapoor, R., et al. (2021). Multi-view data integration methods for radiotherapy structure name standardization. *Cancers* 13:1796. doi: 10.3390/cancers13081796

Talebi, R., Celis-Morales, C. A., Akbari, A., Talebi, A., Borumandnia, N., and Pourhoseingholi, M. A. (2024). Machine learning-based classifiers to predict metastasis in colorectal cancer patients. *Front. Artif. Intellig.* 7:1285037. doi: 10.3389/frai.2024.1285037

Tang, J., Li, K., Hou, M., Jin, X., Kong, W., Ding, Y., et al. (2022). "MMT: multi-way multi-modal transformer for multimodal learning," in *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22* (Vienna: International Joint Conferences on Artificial Intelligence Organization), 3458–3465.

Thangudu, R. R., Rudnick, P. A., Holck, M., Singhal, D., MacCoss, M. J., Edwards, N. J., et al. (2020). Abstract lb-242: Proteomic data commons: a resource for proteogenomic analysis. *Cancer Res.* 80:LB-242. doi: 10.1158/1538-7445.AM2020-LB-242

Thomee, B., Shamma, D. A., Friedland, G., Elizalde, B., Ni, K., Poland, D., et al. (2016). YFCC100M: The new data in multimedia research. *Commun. ACM* 59:64–73. doi: 10.1145/2812802

Tian, Z., Li, X., Zheng, Y., Chen, Z., Shi, Z., Liu, L., et al. (2020). Graph-convolutional-network-based interactive prostate segmentation in MR images. *Med. Phys.* 47, 4164–4176. doi: 10.1002/mp.14327

Tortora, M., Cordelli, E., Sicilia, R., Nibid, L., Ippolito, E., Perrone, G., et al. (2023). Radiopathomics: Multimodal learning in non-small cell lung cancer for adaptive radiotherapy. *IEEE Access*. 11, 47563–47578. doi: 10.1109/ACCESS.2023.3275126

Tripathi, A., Waqas, A., Venkatesan, K., Yilmaz, Y., and Rasool, G. (2024a). Building flexible, scalable, and machine learning-ready multimodal oncology datasets. *Sensors* 24:1634. doi: 10.3390/s24051634

Tripathi, A., Waqas, A., Yilmaz, Y., and Rasool, G. (2024b). Honeybee: a scalable modular framework for creating multimodal oncology datasets with foundational embedding models. *arXiv* [preprint] arXiv:2405.07460.

Tripathi, A., Waqas, A., Yilmaz, Y., and Rasool, G. (2024c). Multimodal transformer model improves survival prediction in lung cancer compared to unimodal approaches. *Cancer Res.* 84:4905–4905. doi: 10.1158/1538-7445.AM2024-4905

Tripathi, S., Moyer, E. J., Augustin, A. I., Zavalny, A., Dheer, S., Sukumaran, R., et al. (2022). RadGenNets: Deep learning-based radiogenomics model for gene mutation prediction in lung cancer. *Inform. Med. Unlocked* 33:101062. doi: 10.1016/j.imu.2022.101062

Valsesia, D., Fracastoro, G., and Magli, E. (2021). RAN-GNNs: breaking the capacity limits of graph neural networks. *IEEE Trans. Neural Netw. Learn. Syst.* 34, 4610–4619. doi: 10.1002/9781119850830.ch3

Varlamova, E. V., Butakova, M. A., Semyonova, V. V., Soldatov, S. A., Poltavskiy, A. V., Kit, O. I., et al. (2024). Machine learning meets cancer. *Cancers* 16:1100. doi: 10.3390/cancers16061100

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). "Attention is all you need," in *Adv. Neural Inf. Process. Syst* (NeurIPS), 30.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., and Bengio, Y. (2017). Graph attention networks. *arXiv* [preprint] arXiv:1710.10903.

Vu, M., and Thai, M. T. (2020). PGM-explainer: probabilistic graphical model explanations for graph neural networks. *Adv. Neural Inf. Process. Syst.* 33, 12225–12235.

Waikhom, L., and Patgiri, R. (2022). A survey of graph neural networks in various learning paradigms: methods, applications, and challenges. *Artif. Intellig. Rev.* 56, 1–70. doi: 10.1007/s10462-022-10321-2

Wang, D., Su, J., and Yu, H. (2020a). Feature extraction and analysis of natural language processing for deep learning english language. *IEEE Access* 8:46335–46345. doi: 10.1109/ACCESS.2020.2974101

Wang, J., Chen, R. J., Lu, M. Y., Baras, A., and Mahmood, F. (2020b). "Weakly supervised prostate TMA classification via graph convolutional networks," in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)* (Iowa City, IA: IEEE), 239–243.

Wang, J., Ma, A., Chang, Y., Gong, J., Jiang, Y., Qi, R., et al. (2021). scGNN is a novel graph neural network framework for single-cell RNA-Seq analyses. *Nat. Commun.* 12:1882. doi: 10.1038/s41467-021-22197-x

Wang, S., Sun, S., Li, Z., Zhang, R., and Xu, J. (2017). Accurate de novo prediction of protein contact map by ultra-deep learning model. *PLoS Comput. Biol.* 13:e1005324. doi: 10.1371/journal.pcbi.1005324

Wang, T., Shao, W., Huang, Z., Tang, H., Zhang, J., Ding, Z., et al. (2021). MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification. *Nat. Commun.* 12:3445. doi: 10.1038/s41467-021-23774-w

Wang, Y., Wang, Y. G., Hu, C., Li, M., Fan, Y., Otter, N., et al. (2022). Cell graph neural networks enable the precise prediction of patient survival in gastric cancer. *NPJ Prec. Oncol.* 6:45. doi: 10.1038/s41698-022-00285-5

Waqas, A., Bui, M. M., Glassy, E. F., El Naqa, I., Borkowski, P., Borkowski, A. A., et al. (2023). Revolutionizing digital pathology with the power of generative artificial intelligence and foundation models. *Lab. Investigat*. 2023:100255. doi: 10.1016/j.labinv.2023.100255

Waqas, A., Dera, D., Rasool, G., Bouaynaya, N. C., and Fathallah-Shaykh, H. M. (2021). "Brain tumor segmentation and surveillance with deep artificial neural networks," in *Deep Learning for Biomedical Data Analysis* (Springer Nature), 311–350.

Waqas, A., Farooq, H., Bouaynaya, N. C., and Rasool, G. (2022). Exploring robust architectures for deep artificial neural networks. *Commun. Eng.* 1:46. doi: 10.1038/s44172-022-00043-2

Waqas, A., Tripathi, A., Ahmed, S., Mukund, A., Farooq, H., Schabath, M. B., et al. (2024a). SeNMo: a self-normalizing deep learning model for enhanced multi-omics data analysis in oncology. *arXiv preprint* arXiv:2405.08226. doi: 10.48550/arXiv.2405.08226

Waqas, A., Tripathi, A., Stewart, P., Naeini, M., and Rasool, G. (2024b). Embedding-based multimodal learning on pan-squamous cell carcinomas for improved survival outcomes. *arXiv* [preprint] arXiv:2406.08521. doi: 10.6004/jnccn.2023.7137

Wei, H., Feng, L., Chen, X., and An, B. (2020). "Combating noisy labels by agreement: A joint training method with co-regularization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13726–13735. doi: 10.1109/CVPR42600.2020.01374

Wei, Y., Wang, X., Nie, L., He, X., Hong, R., and Chua, T.-S. (2019). "MMGCN: Multi-modal graph convolution network for personalized recommendation of micro-video," in *Proceedings of the 27th ACM International Conference on Multimedia* (Nice: ACM), 1437–1445.

Wen, H., Ding, J., Jin, W., Wang, Y., Xie, Y., and Tang, J. (2022). "Graph neural networks for multimodal single-cell data integration," in *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (New York, NY: ACM), 4153–4163. doi: 10.1145/3534678.3539213

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., and Philip, S. Y. (2020). A comprehensive survey on graph neural networks. *IEEE Trans. Neural Netw. Learn. Syst*. 32:4–24. doi: 10.1109/TNNLS.2020.2978386

Xiao, Y., Codevilla, F., Gurram, A., Urfalioglu, O., and López, A. M. (2020). Multimodal end-to-end autonomous driving. *IEEE Trans. Intellig. Transp. Syst*. 23, 537–547. doi: 10.1109/TITS.2020.3013234

Xie, Y., Zhang, J., Shen, C., and Xia, Y. (2021). "Cotr: Efficiently bridging cnn and transformer for 3d medical image segmentation," in *Medical Image Computing and Computer Assisted Intervention-MICCAI 2021: 24th International Conference, Strasbourg, France, September 27-October 1, 2021, Proceedings, Part III 24* (Cham: Springer).

Xu, P., Zhu, X., and Clifton, D. A. (2023). Multimodal learning with transformers: a survey. *IEEE Trans. Pattern Anal. Mach. Intellig*. 45, 12113–12132. doi: 10.1109/TPAMI.2023.3275156

Xu, Y., Das, P., and McCord, R. P. (2022). SMILE: mutual information learning for integration of single-cell omics data. *Bioinformatics* 38, 476–486. doi: 10.1093/bioinformatics/btab706

Yang, K. D., Belyaeva, A., Venkatachalapathy, S., Damodaran, K., Katcoff, A., Radhakrishnan, A., et al. (2021). Multi-domain translation between single-cell imaging and sequencing data using autoencoders. *Nat. Commun*. 12, 31. doi: 10.1038/s41467-020-20249-2

Yang, L., Ng, T. L. J., Smyth, B., and Dong, R. (2020). "HTML: hierarchical transformer-based multi-task learning for volatility prediction," in *Proceedings of The Web Conference 2020, WWW '20* (New York, NY: Association for Computing Machinery), 441–451.

Yang, T., Hu, L., Shi, C., Ji, H., Li, X., and Nie, L. (2021). HGAT: Heterogeneous graph attention networks for semi-supervised short text classification. *ACM Trans. Inform. Syst*. (*TOIS)* 39, 1–29. doi: 10.1145/3450352

Yap, J., Yolland, W., and Tschandl, P. (2018). Multimodal skin lesion classification using deep learning. *Exp. Dermatol*. 27, 1261–1267. doi: 10.1111/exd.13777

Yi, H.-C., You, Z.-H., Huang, D.-S., and Kwoh, C. K. (2022). Graph representation learning in bioinformatics: trends, methods and applications. *Brief. Bioinform*. 23:bbab340. doi: 10.1093/bib/bbab340

Ying, Z., Bourgeois, D., You, J., Zitnik, M., and Leskovec, J. (2019). "GNNExplainer: generating explanations for graph neural networks," in *Adv. Neural Inf. Process. Syst*, 32.

Yogi, M. K., and Mundru, Y. (2024). Genomic data analysis with variant of secure multi-party computation technique. *J. Trends Comp. Sci. Smart Technol*. 5, 450–470. doi: 10.36548/jtcsst.2023.4.006

Yousefi, S., Amrollahi, F., Amgad, M., Dong, C., Lewis, J. E., Song, C., et al. (2017). Predicting clinical outcomes from large scale cancer genomic profiles with deep survival models. *Sci. Rep*. 7:1. doi: 10.1038/s41598-017-11817-6

Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., and Wu, Y. (2022). "Coca: Contrastive captioners are image-text foundation models," in *Transactions on Machine Learning Research* (New York, NY: JMLR).

Yuan, H., Yu, H., Wang, J., Li, K., and Ji, S. (2021). "On explainability of graph neural networks via subgraph explorations," in *International Conference on Machine Learning* (New York: PMLR), 12241–12252.

Zeng, Y., Zhou, X., Rao, J., Lu, Y., and Yang, Y. (2020). "Accurately clustering single-cell RNA-seq data by capturing structural relations between cells through graph convolutional network," in *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (Seoul: IEEE).

Zhang, H., Wu, B., Yuan, X., Pan, S., Tong, H., and Pei, J. (2022). Trustworthy graph neural networks: Aspects, methods and trends. *arXiv* [preprint] arXiv:2205.07424.

Zhang, N. (2020). Learning adversarial transformer for symbolic music generation. *IEEE Trans. Neural Netw. Learn. Syst*. 34, 1754–1763. doi: 10.1109/TNNLS.2020.2990746

Zhang, Y.-D., Satapathy, S. C., Guttery, D. S., Górriz, J. M., and Wang, S.-H. (2021). Improved breast cancer classification through combining graph convolutional network and convolutional neural network. *Inform. Proc. Manage*. 58(2):102439. doi: 10.1016/j.ipm.2020.102439

Zhang, Z., Yang, C., and Zhang, X. (2022). scDART: integrating unmatched scRNA-seq and scATAC-seq data and learning cross-modality relationship simultaneously. *Genome Biol*. 23:139. doi: 10.1186/s13059-022-02706-x

Zhao, B., Gong, M., and Li, X. (2022). Hierarchical multimodal transformer to summarize videos. *Neurocomputing* 468:360–369. doi: 10.1016/j.neucom.2021.10.039

Zhao, F., Zhang, C., and Geng, B. (2024). *Deep Multimodal Data Fusion*. Beijing: ACM Computing Surveys. doi: 10.1145/3649447

Zhao, M., Huang, X., Jiang, J., Mou, L., Yan, D.-M., and Ma, L. (2023). Accurate registration of cross-modality geometry via consistent clustering. *IEEE Trans. Visualizat. Comp. Graph*. 30, 4055–4067. doi: 10.1109/TVCG.2023.3247169

Zhong, Z., Schneider, D., Voit, M., Stiefelhagen, R., and Beyerer, J. (2023). "Anticipative feature fusion transformer for multi-modal action anticipation," in *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* (Vancouver, BC: IEEE/CVF), 6057–6066.

Zhu, H., Sun, X., Li, Y., Ma, K., Zhou, S. K., and Zheng, Y. (2022). *DFTR: Depth-supervised Fusion Transformer for Salient Object Detection*. arXiv preprint.

Zhuang, L., Wayne, L., Ya, S., and Jun, Z. (2021). "A robustly optimized BERT pre-training approach with post-training," in *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, S. Li, M. Sun, Y. Liu, H. Wu, K. Liu, W. Che (Huhhot, China: Chinese Information Processing Society of China), 1218–1227.

Zitnik, M., Agrawal, M., and Leskovec, J. (2018). Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics* 34, i457–i466. doi: 10.1093/bioinformatics/bty294