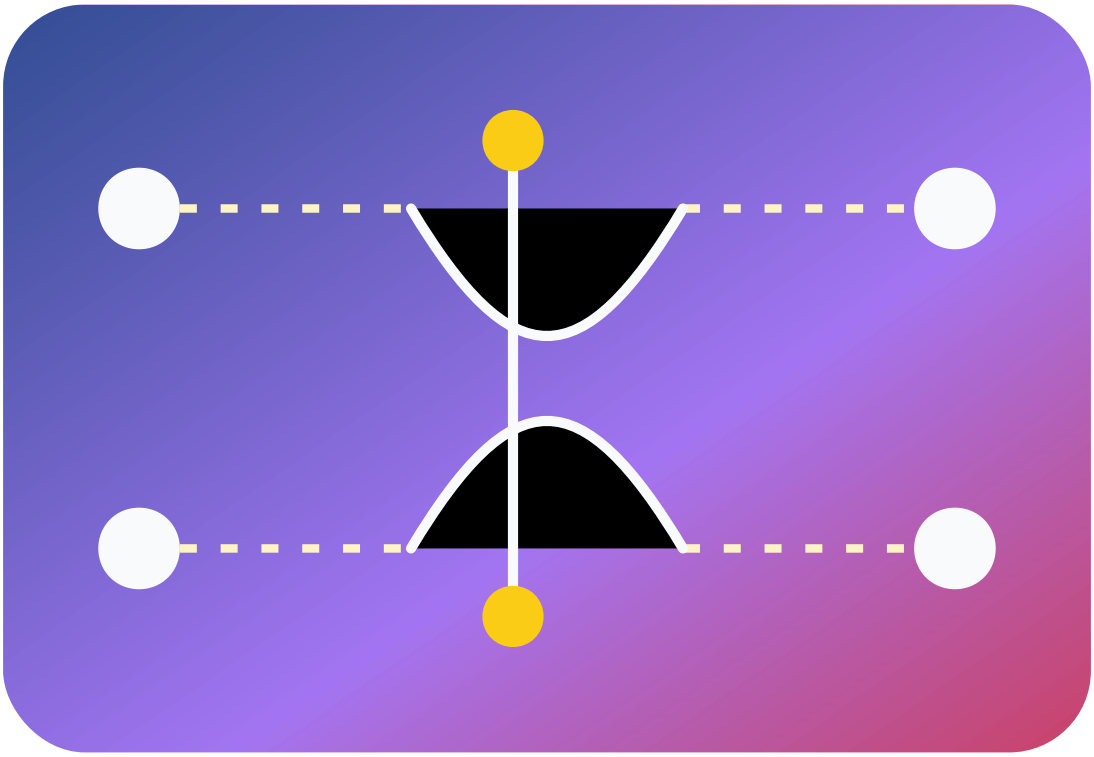
Multimodal Data Integration for Oncology in the Era of Deep Neural Networks: A Review



<h3>Multimodal Data Integration for Oncology in the Era of Deep Neural Networks: A Review · Concept Sketch</h3>
Late-fusion inspired view showing coordinated yet modality-specific streams.

Multimodal Data Integration for Oncology in the Era of Deep Neural Networks: A Review

Authors: Asim Waqas, Aakash Tripathi, Ravi P. Ramachandran, Paul A. Stewart, Ghulam Rasool
Year: 2024
Venue: Frontiers in Artificial Intelligence

---

1. Classification

- **Domain Category:**
  - **Multimodal / Integration** (specifically applied to oncology/cancer research). This is a review paper surveying multimodal deep learning methods—especially Graph Neural Networks (GNNs) and Transformers—for integrating diverse cancer-related data modalities (imaging, omics, clinical records).
- **FM Usage Type:**
  - **Review of multimodal FM and integration methods**. The paper does not propose a new foundation model but systematically reviews how modern deep learning architectures (GNNs, Transformers) are being applied to multimodal oncology data fusion, including references to foundation models like CLIP, GPT-4, and FLAVA.
- **Key Modalities:**
  - Radiological imaging (CT, MRI, PET scans)
  - Digitized histopathology (whole-slide images, H&E stains)
  - Multi-omics (genomics, transcriptomics, proteomics, metabolomics)
  - Electronic health records (EHR) and clinical data
  - Hybrid/derived modalities (radiomics, pathomics features)

---

2. Executive Summary

This review paper surveys the landscape of multimodal data integration in oncology, focusing on how deep neural networks—particularly Graph Neural Networks (GNNs) and Transformers—are being used to fuse diverse cancer data types for improved diagnosis, prognosis, and treatment prediction. Cancer research generates heterogeneous data across multiple scales and modalities: from imaging (radiology, pathology) to molecular profiles (genomics, transcriptomics, proteomics) to clinical records. Traditional unimodal analyses fail to capture the complex, interconnected nature of cancer biology. The authors present a comprehensive taxonomy of multimodal learning approaches, covering fusion strategies (early, intermediate, late), neural architectures (CNNs, RNNs, GNNs, Transformers), and domain-specific applications in oncology. They highlight how GNNs naturally model relationships among heterogeneous entities (patients, genes, images) and how Transformers, through self-attention, can integrate sequences of multimodal tokens. Key studies are reviewed across tasks like tumor classification, survival prediction,

treatment response, and biomarker discovery. The paper also discusses major challenges—data heterogeneity, missing modalities, alignment across scales, interpretability, and the need for large labeled datasets—and points to promising directions, including foundation models and self-supervised pretraining. For researchers and clinicians, this review provides both a conceptual framework for understanding multimodal fusion and a practical roadmap for applying state-of-the-art deep learning to personalized cancer care.

---

3. Problem Setup and Motivation

Scientific / practical problem:

- Cancer is inherently multimodal: its biology spans genomic mutations, protein expression, tissue morphology, organ-level imaging, and clinical phenotypes.
- **Prediction and personalization goals:**
  - Accurate early diagnosis and cancer subtype classification
  - Prognosis (survival, recurrence risk)
  - Treatment response prediction and therapy selection
  - Discovery of prognostic and predictive biomarkers
- Traditional approaches analyze modalities in isolation (e.g., only imaging or only genomics), missing synergistic information that could improve accuracy and clinical utility.

Why this is hard:

- **Data heterogeneity:**
  - Different modalities have different dimensionalities, scales, and noise characteristics (e.g., high-resolution images vs sparse genomic variants vs tabular clinical data).
  - Modalities are collected with different protocols, scanners, and sequencing platforms, leading to batch effects and site-specific biases.
- **Integration complexity:**
  - No universal representation: images are spatial grids, omics are vectors or graphs, clinical data are tabular.
  - Determining *when* and *how* to fuse (early vs late) requires domain knowledge and empirical tuning.
- **Missing data:**
  - Not all patients have all modalities (e.g., some lack genomic profiling, others lack certain imaging studies).
  - Models must handle partial observations gracefully.
- **Label scarcity and class imbalance:**
  - High-quality multi-modal datasets with expert annotations are rare and expensive.
  - Many cancer subtypes are rare, leading to imbalanced training sets.
- **Interpretability and clinical trust:**
  - “Black-box” deep learning predictions are hard to explain, yet clinicians need to understand *why* a model predicts a certain outcome.
  - Regulatory and ethical considerations demand transparency.

---

4. Data and Modalities

Oncology data modalities covered in the review:

- **Radiological imaging:**
  - CT, MRI, PET scans
  - Used for tumor detection, staging, and monitoring response
- **Digitized histopathology:**
  - Whole-slide images (WSI) of H&E-stained tissue
  - Immunohistochemistry (IHC) and other stains
  - Pathomics: quantitative features extracted from slides
- **Multi-omics:**
  - **Genomics:** DNA mutations, copy number variations, single nucleotide polymorphisms
  - **Transcriptomics:** RNA-seq, gene expression profiles
  - **Proteomics:** Protein abundance and post-translational modifications
  - **Metabolomics:** Small molecule profiles
  - **Epigenomics:** DNA methylation, histone modifications
- **Electronic Health Records (EHR) and clinical data:**
  - Demographics, clinical notes, laboratory results, treatment history, survival outcomes
- **Radiomics:**
  - Hand-crafted or learned features from imaging (texture, shape, intensity)

Major datasets mentioned:

- **The Cancer Genome Atlas (TCGA):** Pan-cancer multi-omics and clinical data
- **Genomic Data Commons (GDC):** Centralized repository for TCGA and other NCI programs
- **UK Biobank, All of Us:** Large-scale cohorts with imaging and genomics
- **TCIA (The Cancer Imaging Archive):** Radiological imaging datasets
- Various cancer-specific cohorts (e.g., NSCLC, breast cancer, glioblastoma)

Preprocessing / representation:

- Images: patches or whole-image embeddings from CNNs
- Omics: normalized vectors, sometimes projected into lower dimensions
- Clinical: tabular features, often encoded or embedded
- Graphs: patients, genes, images as nodes; relationships (co-expression, similarity) as edges

---

5. Model / Foundation Model

Model Types:

The review covers a range of architectures for multimodal fusion:

Architecture	Role in Multimodal Oncology
CNNs	Feature extraction from images (radiology, pathology)
RNNs/LSTMs	Sequential clinical data, temporal progression

Architecture	Role in Multimodal Oncology
Autoencoders/VAEs	Dimensionality reduction, unsupervised feature learning
GANs	Data augmentation, synthetic image generation
Graph Neural Networks (GNNs)	Model relationships among patients, genes, and multi-modal entities; handle heterogeneous graphs
Transformers	Self-attention over multimodal token sequences; pre-trained vision-language models adapted to oncology

Focus on GNNs:

- **Graph representation:**
  - Nodes: patients, genes, images, or feature vectors from different modalities
  - Edges: similarity (clinical, genomic), co-occurrence, known biological interactions
- **GNN architectures reviewed:**
  - Graph Convolutional Networks (GCN)
  - Graph Attention Networks (GAT)
  - GraphSAGE
  - Message Passing Neural Networks (MPNN)
- **Applications:**
  - Patient similarity networks for survival prediction
  - Gene regulatory networks combined with patient omics
  - Pathology graphs (cells/patches as nodes) integrated with omics

Focus on Transformers:

- **Vanilla Transformers:** Self-attention to integrate sequences of multimodal embeddings.
- **Vision Transformers (ViT):** Patches of histopathology or radiology images as tokens.
- **Multimodal Transformers:**
  - Cross-modal attention between image and text/omics modalities
  - CLIP-like contrastive learning adapted to radiology-pathology or image-genomics pairs
- **Foundation models mentioned:** CLIP, GPT-4, FLAVA, and domain-specific models like MedCLIP.

Training setup (general patterns):

- **Pretraining:** Often on large unimodal datasets (e.g., ImageNet for images, public omics for gene embeddings), sometimes with self-supervised objectives.
- **Fine-tuning / transfer learning:** Adapt pretrained encoders to oncology tasks with smaller labeled datasets.
- **Fusion stages:**
  - **Early fusion:** Concatenate raw features before modeling.
  - **Intermediate (joint) fusion:** Learn shared representations mid-network.
  - **Late fusion:** Train separate modality-specific models, combine predictions at the end.
- **Scale:** Varies widely; some studies use hundreds of patients, others leverage TCGA's thousands of samples. Model sizes range from small task-specific networks to large Transformer-based foundation models.

6. Multimodal / Integration Aspects (If Applicable)

This is fundamentally a **multimodal integration review**, so this section is central.

Which modalities are integrated:

- Common pairs/triplets:
  - **Radiology + pathology:** CT/MRI + histopathology WSI
  - **Imaging + omics:** Radiology or pathology + genomics/transcriptomics
  - **Omics + clinical:** Gene expression + EHR/treatment history
  - **Triple integration:** Imaging + omics + clinical (less common, more challenging)

How they are integrated:

- **Early fusion:**
  - Concatenate features from all modalities into a single vector and feed into a downstream classifier.
  - *Pros:* Simple, allows the model to learn joint patterns from the start.
  - *Cons:* Can be dominated by the highest-dimensional modality; requires careful normalization; struggles with missing data.
- **Late fusion:**
  - Train separate models for each modality, then combine predictions (e.g., averaging, voting, stacking).
  - *Pros:* Preserves modality-specific signals; robust to missing modalities; easier to interpret.
  - *Cons:* May miss complex cross-modal interactions.
- **Intermediate (joint) fusion:**
  - Modality-specific encoders produce embeddings that are fused at a middle layer (e.g., via concatenation, attention, or graph pooling) before final prediction.
  - *Pros:* Balances flexibility and integration.
  - *Cons:* Requires architectural design choices; harder to optimize.
- **GNN-based fusion:**
  - Construct a heterogeneous graph with nodes from different modalities (patient omics, image features, clinical variables).
  - GNN message passing aggregates cross-modal information.
  - *Example:* A patient node connected to its gene expression profile node and its pathology image embedding node; GNN learns to propagate and combine information.
- **Transformer-based fusion:**
  - Represent each modality as a sequence of tokens (e.g., image patches, genomic regions, clinical features).
  - Self-attention and cross-attention layers integrate across modalities.
  - *Example:* Multimodal Transformer taking pathology image patches and omics embeddings as separate token sets, with attention heads learning cross-modal dependencies.

Why this integration is useful:

- **Complementary information:** Imaging reveals spatial tumor characteristics, omics show molecular drivers, clinical data provide context (age, stage, treatment).

- **Improved prediction:** Studies show multimodal models often outperform unimodal baselines on survival, classification, and treatment response tasks.
- **Biological insight:** Cross-modal associations (e.g., imaging phenotypes correlated with gene expression) can reveal biomarkers and mechanisms.
- **Personalization:** Comprehensive profiles enable tailored treatment recommendations.

Relation to the integration baseline plan:

- **Late fusion first under heterogeneous semantics:**
  - The review aligns with this principle: many successful oncology studies use late fusion or ensemble methods, preserving modality-specific encoders.
  - GNNs and Transformers can implement late fusion naturally (separate encoding + graph/attention-based aggregation).
- **Robustness and evaluation discipline:**
  - The review emphasizes the need for rigorous cross-validation, proper train/test splits (especially important given small sample sizes), and metrics like AUROC, AUPRC, C-index for survival.
  - Challenges of missing data and distribution shift are highlighted, aligning with the baseline plan's focus on residualization, covariate adjustment, and bootstrap confidence intervals.
- **CCA and permutation testing:**
  - Not explicitly covered in the review, but the principle of exploring modality correlations before heavy fusion is implicit in studies that perform feature selection or canonical correlation analysis on omics and imaging features.
- **Modality sequencing:**
  - The review suggests starting with well-characterized modalities (e.g., standard imaging + omics) and progressively adding more complex ones (e.g., pathology WSI, radiomics), consistent with the plan's incremental approach.

---

7. Experiments and Results

Tasks / benchmarks reviewed:

The paper surveys a wide range of studies across multiple oncology tasks:

- **Tumor classification and subtyping:**
  - GNN and Transformer models classify cancer types (e.g., glioma grades, breast cancer molecular subtypes) using combined imaging and omics.
  - Studies report accuracy improvements of 3-10% over unimodal baselines.
- **Survival prediction and prognosis:**
  - Multimodal Cox regression models, GNN-based survival networks, and attention-based models integrate clinical, omics, and imaging data.
  - C-index improvements of ~0.05-0.15 compared to clinical-only or omics-only models.
- **Treatment response prediction:**
  - Predicting response to chemotherapy, immunotherapy, or targeted therapies.
  - Multimodal approaches combining radiology (baseline tumor imaging) with genomics (mutation profiles) show better discrimination (AUC gains of 0.05-0.10).

- **Biomarker discovery:**
  - GNNs identify gene modules and image features associated with outcomes.
  - Transformers' attention weights highlight cross-modal associations (e.g., specific image regions correlating with gene expression patterns).

Baselines:

- Unimodal models (imaging-only, omics-only, clinical-only)
- Traditional ML methods (logistic regression, random forests on concatenated features)
- Early fusion baselines (simple concatenation + MLP)

Key findings (trends and insights):

- **Multimodal consistently beats unimodal:** Across most studies, integrating multiple data types improves predictive performance, often significantly.
- **Fusion strategy matters:** Late fusion and intermediate fusion tend to outperform early fusion, especially when modalities have very different characteristics.
- **GNNs excel at relational data:** When patient or gene relationships are explicitly modeled, GNNs capture network effects that simpler models miss.
- **Transformers scale well:** Transformer-based models benefit from larger datasets and can leverage pretrained vision-language models (transfer learning from CLIP-like architectures).
- **Interpretability via attention:** Attention weights in Transformers and GNN message passing provide some interpretability, highlighting which modality or feature drives predictions.
- **Challenges remain:** Performance gains are modest in some cases; missing data and small sample sizes limit generalization; computational cost is high for large-scale models.

---

8. Strengths, Limitations, and Open Questions

Strengths:

- **Comprehensive survey:** Covers a wide range of architectures (CNNs, RNNs, GANs, GNNs, Transformers) and applications in oncology.
- **Taxonomy and framework:** Provides a clear taxonomy of fusion strategies and multimodal learning paradigms, useful for researchers entering the field.
- **Focus on emerging methods:** Highlights GNNs and Transformers, which are underexplored in oncology compared to computer vision and NLP.
- **Identifies data resources:** Lists major multimodal oncology datasets (TCGA, GDC, TCIA, UK Biobank), facilitating reproducible research.
- **Balances technical depth and accessibility:** Suitable for both ML researchers new to oncology and oncology researchers new to advanced deep learning.

Limitations:

- **Limited discussion of causal inference:** The review focuses on predictive modeling; less attention to causal relationships or confounding (e.g., how to distinguish direct biological effects from correlations).

- **Sparse coverage of robustness and fairness:** Issues like model bias across demographics, generalization to external cohorts, and adversarial robustness are mentioned but not deeply explored.
- **Lack of standardized benchmarks:** The field lacks common evaluation protocols and public leaderboards, making it hard to compare methods across studies.
- **Interpretability still nascent:** While attention weights and GNN message passing offer some transparency, true mechanistic interpretability (linking predictions to biological pathways) remains an open challenge.
- **Computational and data barriers:** Many proposed methods require large computational resources and extensive labeled data, limiting accessibility for smaller research groups and clinical settings.

Open Questions and Future Directions:

- **Foundation models for oncology multimodal data:**
  - Can we pretrain large multimodal Transformers on diverse cancer datasets (like CLIP for vision-language) to create a general-purpose oncology FM that transfers to many downstream tasks?
- **Handling missing modalities robustly:**
  - Develop architectures that gracefully handle partial observations (e.g., modality dropout during training, imputation via cross-modal generation).
- **Causal multimodal modeling:**
  - Move beyond association to causal discovery: which modality changes *drive* outcomes? How to design experiments or observational studies to infer causality?
- **Fairness and generalization:**
  - Ensure multimodal models perform equitably across different patient demographics, cancer subtypes, and institutions (multi-site validation, fairness-aware training).
- **Integration with clinical workflows:**
  - Design models that output actionable, interpretable predictions usable by oncologists in real-time decision-making.
- **Self-supervised and few-shot learning:**
  - Leverage unlabeled multimodal data (vast amounts available) via self-supervised pretraining; adapt models to rare cancers with few labeled examples using few-shot learning.
- **Explainability and biological insight:**
  - Develop methods to extract mechanistic understanding from multimodal models (e.g., which genes and image features co-vary and why?).

---

9. Context and Broader Impact

Position in the FM and multimodal learning landscape:

- This review sits at the intersection of **multimodal machine learning** and **biomedical AI**, specifically focused on cancer.
- It relates to broader trends in foundation models:
  - **CLIP, GPT-4, FLAVA** demonstrate that large-scale pretraining on multimodal data (vision + language) yields versatile representations.

    - Oncology is following suit: researchers are exploring whether similar pretraining strategies (e.g., on large radiology-pathology-omics datasets) can yield "cancer foundation models."
- **Analogy to well-known ideas:**
    - GNNs for multimodal oncology are like "knowledge graphs for cancer," where nodes and edges capture heterogeneous entities and relationships.
    - Transformers for multimodal oncology are like "BERT/GPT but for diverse cancer data tokens," using self-attention to integrate across modalities.

Relation to the integration baseline plan:

- The review's taxonomy (early vs late fusion, GNN vs Transformer architectures) directly informs the baseline plan's integration strategy recommendations.
- **Late fusion** (modality-specific encoders + final aggregation) is a recurring theme in successful studies, consistent with the plan's preference for preserving modality-specific signals.
- The emphasis on **evaluation rigor** (cross-validation, proper metrics, missing data handling) aligns with the plan's robustness and evaluation discipline.
- The review highlights **GNNs and Transformers** as promising architectures for escalation beyond simple concatenation-based fusion, matching the plan's suggestion to explore two-tower contrastive or hub-token architectures if late fusion proves valuable.

Why this paper is a useful reference:

- **Educational value:** For a new grad student, this review provides a structured entry point into multimodal oncology, with clear definitions, examples, and a roadmap of key papers.
- **Design patterns:** The taxonomy of fusion strategies and architectures serves as a design template for building new multimodal systems in cancer or other biomedical domains.
- **Data resources:** The compilation of datasets accelerates research by pointing to readily available multimodal cohorts.
- **Future directions:** The open questions guide thesis topics and grant proposals, highlighting high-impact areas for methodological development.

---

10. Key Takeaways (Bullet Summary)

Problem:

- Cancer data is inherently multimodal (imaging, omics, clinical records), but traditional analyses treat modalities in isolation.
- Integrating diverse data types promises improved diagnosis, prognosis, and treatment prediction, but poses significant technical and domain challenges.

Method / Model:

- The review surveys **deep neural network architectures** for multimodal fusion in oncology:
    - **CNNs and RNNs** for imaging and sequential data
    - **Graph Neural Networks (GNNs)** for modeling relationships among patients, genes, and features across modalities

    - **Transformers** for self-attention-based integration of multimodal token sequences
- **Fusion strategies** covered:
    - **Early fusion:** Concatenate features before modeling
    - **Late fusion:** Separate modality-specific models, combine predictions
    - **Intermediate fusion:** Joint representations learned mid-network, often via attention or graph pooling
- The review emphasizes **foundation model** concepts (pretraining, transfer learning) and references CLIP, GPT-4, FLAVA as inspiration for oncology-specific multimodal FMs.

Results / Insights:

- Multimodal models generally **outperform unimodal baselines** on classification, survival, and treatment response tasks, with typical improvements of 3-15% in accuracy/AUC and 0.05-0.15 in C-index.
- **Late and intermediate fusion** strategies often yield the best performance, preserving modality-specific information while capturing cross-modal interactions.
- **GNNs** are particularly effective when relational structure (patient networks, gene-gene interactions) is explicitly modeled.
- **Transformers** scale well and benefit from pretrained vision-language models, showing promise for large-scale oncology FMs.
- **Interpretability** is improved via attention weights and GNN message passing, but mechanistic understanding remains limited.

Why it matters:

- This review provides a **comprehensive, structured overview** of multimodal deep learning in oncology, filling a gap in the literature by focusing on GNNs and Transformers.
- It offers a **taxonomy and design framework** that researchers can use to build and evaluate multimodal systems.
- By highlighting **data resources, challenges, and future directions**, it accelerates progress toward personalized cancer care powered by integrated, interpretable AI.
- For the **integration baseline plan**, this review validates key principles (late fusion under heterogeneity, robustness discipline, modality sequencing) and points to GNNs/Transformers as architectural choices for escalation beyond simple baselines.

Generated via custom pipeline · 2025-11-18