Artificial intelligence in multimodal data analysis for cancer survival prediction

Siti Saadah[a], Luis-Daniel Ibáñez[a], Rob M. Ewing[b,c], and Zehor Belkhatir[a,*]

[a]School of Electronics and Computer Science, University of Southampton, Southampton, United Kingdom
[b]Biological Sciences, Faculty of Environmental and Life Sciences, University of Southampton, Southampton, United Kingdom
[c]Institute for Life Sciences, University of Southampton, Southampton, United Kingdom
*Corresponding author. e-mail address: z.belkhatir@soton.ac.uk

Contents

Progress in Molecular Biology and Translational Science
ISSN 1877-1173, https://doi.org/10.1016/bs.pmbts.2026.03.001

Abstract

Artificial intelligence (AI) has been introduced to meet the demand for more precise predictions of cancer patient survival by simultaneously interpreting multiple types of input data. These data can originate from a broad spectrum of modalities, where a modality refers to one of several distinct forms in which data can be represented or observed. Using AI, researchers have investigated how to model the interactions among different data modalities as a promising approach to multimodal fusion in cancer prognosis. The goal is to build models capable of reliably handling and integrating these heterogeneous data sources to improve the accuracy and interpretability of patient outcome predictions. Nonetheless, current integration techniques encounter two major obstacles that must be overcome before survival prediction can be carried out: data alignment and data fusion. This chapter surveys existing alignment methods and fusion strategies within a pipeline designed to predict cancer survival. The chapter begins by describing the search strategy followed to identify the relevant literature to be analyzed here. Then, multimodal machine learning is introduced with the three key data types investigated in this study. we review 31 recent multimodal studies related to alignment and fusion using deep learning, covering the period from 2015 to 2025. Among these, 18 studies focus on alignment techniques followed by fusion strategies, while another 13 investigate fusion strategies independently of alignment. This chapter illustrates how synchronizing the data modalities before applying fusion can improve both the performance and the interpretability of the models. The synthesis of these studies uncovers methodological challenges and suggests future directions to develop more effective and clinically relevant AI-based survival prediction frameworks.

Glossary

BRCA	Breast Invasive Carcinoma
GBMLGG	Glioblastoma Lower Grade Glioma
GC	Gastric Cancer
BLCA	Bladder Urothelial Carcinoma
LUAD	Lung Adenocarcinoma
NSCLC	Non-Small Cell Lung Cancer
RCC	Renal Cell Carcinoma
COADREAD	Colon and Rectal Adenocarcinoma
STAD	Stomach Adenocarcinoma
HNSC	Head-neck Squamous Cell Carcinoma
GBM	Glioblastoma
LGG	Lower Grade Glioma
LUSC	Lung Squamous Cell Carcinoma
KIRC	Kidney Clear Cell Carcinoma
LIHC	Liver Hepatocellular Carcinoma
ESCA	Esophageal Carcinoma
CRC	Colon and Rectum Adenocarcinoma
KIRP	Kidney Renal Papillary Cell Carcinoma

1. Introduction

Survival analysis encompasses statistical and bioinformatic methods that model the time until a critical clinical event occurs, such as disease progression or death.[1] In cancer, accurate survival prediction is crucial for treatment planning, risk stratification, and efficient healthcare delivery.[2] It also enables the development of personalized therapies that balance treatment efficacy with potential side effects and toxicity. However, survival prediction remains challenging due to nonlinear time dependencies and the complexity of high-dimensional biomedical inputs.[3,4] These difficulties are amplified by integrating heterogeneous modalities, such as genomics and histopathology, that differ in scale, structure, and noise.[5] Artificial Intelligence (AI), including machine learning and deep learning, has emerged rapidly in the last decades to address these challenges and gaps. As a result, multimodal machine learning offers a robust framework for cross-modal modeling and enhances survival prediction.[6] Recent research has used machine/deep learning techniques for multimodal integration, such as encoder-based frameworks,[7–9] graph neural networks (GNNs),[10,11] and transformer models.[4,12,13] Many previous studies integrate data directly without initially addressing the need for alignment. Consequently, approaches such as the Kronecker product[14,15] and low-rank multimodal fusion[16] continue to face challenges in effectively managing diverse and heterogeneous multimodal medical information.

A typical comprehensive procedure for multimodal survival prediction includes steps for data acquisition, preprocessing, feature engineering, alignment, fusion, and model-based survival inference, as shown in Fig. 1. The challenges in efficiently analyzing various types of data are alignment and fusion,[17] which are the main topics studied in this survey paper. Alignment aims to create semantic linkage across modalities with varying levels of detail (e.g., patch-level histopathology and gene-level omics). Fusion aims to combine these aligned inputs into a unified predictive output. An expanding body of research uses alignment to bridge cross-modal variations in scale, structure, and distribution, enhancing the consistency and integration of representations.[7] In practice, this usually involves mapping the modalities into a shared representation space or otherwise rendering them sufficiently compatible to support effective fusion. Recent notable AI-based methods have employed optimal transport (OT),[3,4] contrastive learning,[13] hypergraph learning,[10] and hybrid methods[18] that combine these techniques. Motivated by this, we survey in this chapter the added value of performing

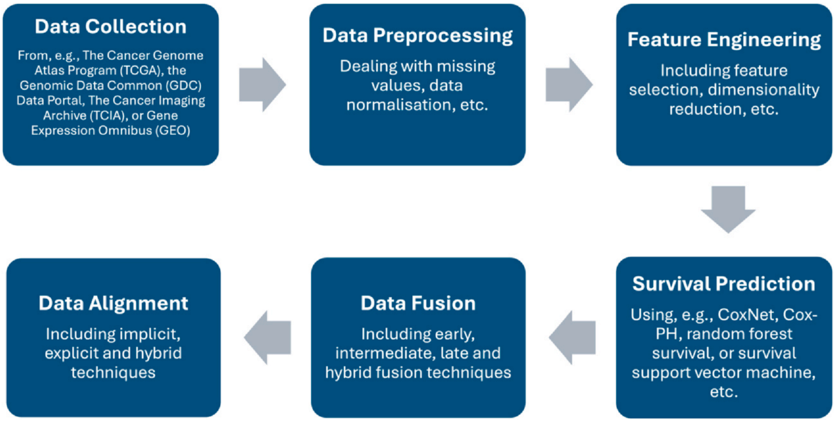
Data Collection
From, e.g., The Cancer Genome Atlas Program (TCGA), the Genomic Data Common (GDC) Data Portal, The Cancer Imaging Archive (TCIA), or Gene Expression Omnibus (GEO)
Data Preprocessing
Dealing with missing values, data normalisation, etc.
Feature Engineering
Including feature selection, dimensionality reduction, etc.
Survival Prediction
Using, e.g., CoxNet, Cox-PH, random forest survival, or survival support vector machine, etc.
Data Fusion
Including early, intermediate, late and hybrid fusion techniques
Data Alignment
Including implicit, explicit and hybrid techniques

Fig. 1 Schematic diagram of a typical flow for multimodal machine learning in cancer survival prediction.

alignment before fusion, with the goals of reducing inter-modal discrepancies, improving cross-modal interpretability, and increasing the predictive performance of survival analysis.

In the last 10 years, various studies have investigated multimodal machine learning (MML) strategies to predict survival outcomes; for example, Li et al.[19] focus on the role of multimodal deep learning-based fusion for cancer diagnosis highlighting different cancer-specific fusion models. Waqas et al.[20] focus on the analysis of applying deep neural networks for the integration of multimodal data in oncology, while Abbasi et al.[2] further investigate survival prediction. Systematic reviews or surveys have also been conducted, such as those by Farhadizadeh et al.,[21] exploring the challenges of handling multimodal data in medical settings, notably missing modality, interpretability, small data, dimensionality imbalance, and optimal fusion. In addition, Barua et al.[22] provided a comprehensive review of MML applications and their challenges. In 2017, Baltruaitis et al.[23] have elaborated more on the MML taxonomy. Their survey highlights key challenges, such as representation and co-learning in multimodal machine learning. However, it does not thoroughly categorize the field of alignment and fusion within the domain of survival prediction. Subsequently, in 2024 Li et al.[17] investigated multimodal alignment and fusion; however, they rarely discussed the impact of different alignment strategies on the design and improvement of fusion methods to predict cancer patient survival. To our knowledge, no comprehensive survey has examined how different alignment methods affect the selection and efficacy of various fusion architectures (early, intermediate, late, or hybrid) and their impact on survival cancer predictions.

This chapter aims to address this gap by exploring methods in AI, machine learning, and deep learning to predict cancer survival, with a particular emphasis on approaches that integrate pathology images and omics data. The focus in the role of alignment and fusion involves examining their advantages and disadvantages, organizing existing studies by the type of alignment employed (explicit, implicit, or hybrid), and examining how these choices shape subsequent fusion strategies (early, intermediate, late, or hybrid). We provide a review and taxonomy of alignment and fusion techniques in multimodal data analysis for cancer survival prediction, and summarize our main contributions as follows.

- Introduction of a taxonomy for alignment techniques used prior to fusion in MML for the prediction of survival cancer.
- Comparison of the benefits and limitations of alignment and fusion approaches for different datasets and cancer types.
- Discussion of remaining gaps and challenges, offering guidance for potential future multimodal integration frameworks that are computationally efficient and biologically informed.

The remainder of the chapter is organized as follows. Section 2 details the methodology for the search and selection of articles reviewed in this survey. Section 3 introduces multimodal machine learning along with key modalities of medical data in cancer research. Section 4 reviews multimodal alignment techniques and recent advances in this area. Section 5 investigates fusion methods, assessing those with and without alignment using standard performance metrics. Section 6 provides an overall discussion based on the alignment and fusion techniques reviewed. Section 7 highlights key current challenges and potential future extensions in multimodal data analysis for survival prediction. Section 8 concludes the paper.

2. Search strategy

This study adheres to the guidelines for conducting systematic literature reviews in software engineering as outlined by Kitchenham and Charters in.[24] These guidelines are applied to systematically identify, evaluate, and interpret prior research related to the research questions. (i) What key factors are essential to predict cancer survival? (ii) In what ways can alignment techniques be employed effectively before fusion to enhance prediction accuracy? (iii) What obstacles arise in designing multimodal data integration frameworks that incorporate alignment prior to fusion?

To address these questions, we used various keywords while searching through online databases. We examined four primary databases (PubMed, Science Direct Elsevier, Scopus, and Google Scholar) to ensure relevant results during the literature search. The next step involves developing procedures for locating scientific and technical articles within the four specified databases. This process consists of two segments: (1) selecting search terms such as “multimodal”, “multimodal machine learning”, “cancer survival prediction”, “alignment”, “fusion”, “genomics”, or “histopathology”; (2) using these search terms to create queries employing Boolean operators such as AND or OR. Our search for relevant articles for this study spanned 2015–2025, covering title, abstract, keywords, and introduction.

This survey utilized a specific set of inclusion and exclusion criteria to refine the collected search results, excluding studies that did not meet the inclusion criteria. These criteria are specified in Table 1. For the article selection process, this review followed the PRISMA guidelines.[25] The methodology comprised stages of identification, screening, eligibility, and sorting, as depicted in Fig. 2. Initially, 1926 articles were gathered from

Table 1 Screening criteria for study selection.

Inclusion criteria	Exclusion criteria
1. Studies should answer one or more research questions	1. Unavailable in database format
2. Studies involve the area of multimodal machine learning (MML)	2. Duplicate articles with similar results
3. Studies must be published in a journal and/or conference	3. Studies present a review or survey
4. Published between the year 2015–2025	4. Studies not related to MML to predict cancer survival
5. Focus on alignment and fusion of MML to predict cancer survival	5. Studies not written in English

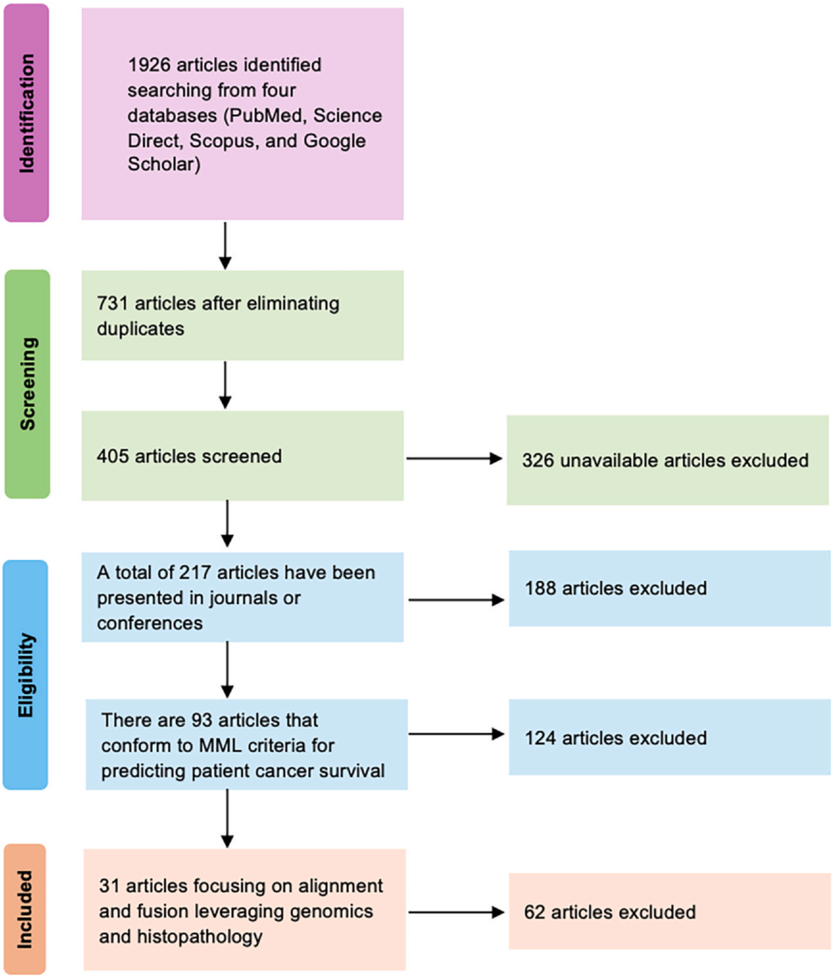
Identification
1926 articles identified searching from four databases (PubMed, Science Direct, Scopus, and Google Scholar)
Screening
731 articles after eliminating duplicates
405 articles screened
326 unavailable articles excluded
Eligibility
A total of 217 articles have been presented in journals or conferences
188 articles excluded
There are 93 articles that conform to MML criteria for predicting patient cancer survival
124 articles excluded
Included
31 articles focusing on alignment and fusion leveraging genomics and histopathology
62 articles excluded

Fig. 2 The process followed for selection of studies to be reviewed in this paper.

databases using keywords. After removing duplicate entries, the count reduced to 731 articles. Of these, 405 were selected and 326 were excluded due to inaccessibility. In the eligibility phase, 217 articles were initially considered, 188 excluded because they had not been published in a journal or conference.

Eventually, 93 scientific or technical articles were chosen, while 124 review or survey articles were excluded. Ultimately, 31 articles met the MML criteria for predicting patient cancer survival, excluding 62 additional articles that were not considered.

3. Multimodal machine learning

In the field of cancer research, multimodal machine learning (MML) has emerged as an interdisciplinary approach that integrates data of various types from various sources with machine learning (ML) techniques, with the aim of solving vital challenges such as prognosis, decision-making, and biomarker identification.[26] MML addresses five core technical challenges: representation, translation, alignment, fusion, and co-learning.[23] Representation *refers* to the representation of data by using information from different modalities and is classified as representation fusion, representation coordination, and representation fission.[27] *Translation* involves converting data from one mode to another. *Alignment* seeks to establish direct or continuous connections between (sub)elements in multiple modalities. *Fusion* entails the synthesis of information from various modalities to make predictions. *Co-learning* refers to the exchange of knowledge between modalities, their representations, and their predictive models.

In 2020, global cancer deaths were estimated to reach 10 million cases.[28] The production of reliable survival predictions depends on the accurate alignment and integration of various data sources.[12] Errors or inconsistencies introduced early in this process can result in faulty integration or loss of key information. Standard clinical management involves examinations such as microscopic pathological assessment and molecular genetic tests to track the state of the disease.[29] These modalities are widely accepted as reference standards for cancer diagnosis, treatment planning, and disease progression monitoring.[30] More information is presented in Fig. 3 and described in the following text.

- **Molecular data.**

  The investigation of molecular data modalities uncovers genetic alterations and other modifications present in cancer cells. Combining these molecular datasets has given rise to the field of multiomics research. This area primarily concentrates on leveraging omics data, which comprise extensive biological datasets generated by high-throughput methods.[31] These datasets contain information on a wide range of biological molecules—such as genes, proteins, and metabolites—each treated as a separate modality.[32] Molecular modalities clarify the genomic and proteomic alterations that drive tumor initiation and progression.[33]

  These modalities, derived through high-throughput technologies, include genomic sequences, transcriptomic profiles, proteomic markers,

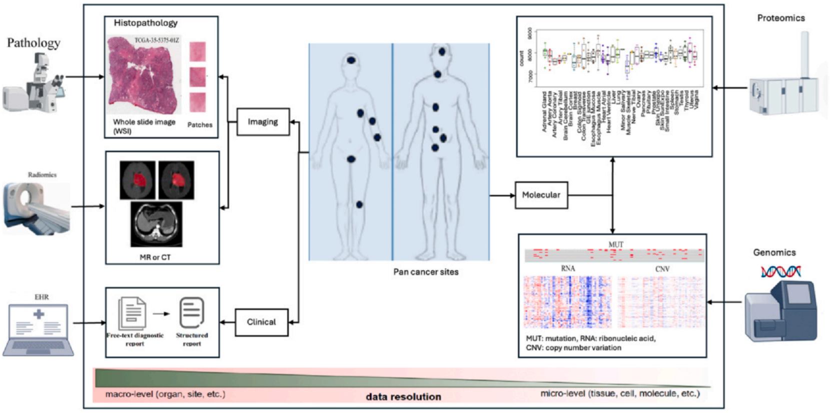

Fig. 3 Core data families in cancer research across macro- to micro-scale. Left: Starting from digital pathology via high-resolution Whole-Slide Imaging (WSI); radiomics extracted from MR or CT scans; and unstructured clinical text harvested from Electronic Health Record (EHR) systems.

and metabolomic signatures.[17,34] Combining these multiple omic layers supports a comprehensive system-level investigation of cancer biology.[20] The resulting datasets are generally high-dimensional, noisy, and complex; thus, require sophisticated computational techniques. Public repositories such as The Cancer Genome Atlas (TCGA)[35,36] and the Genome Data Commons,[34] together with tools such as UCSC Xena for data visualization and analysis,[37] provide a wide access to these heterogeneous data types.

- **Imaging data.**

  Imaging data offer rich visual information acquired through cameras that are seamlessly integrated into medical instruments such as endoscopes and dermatoscopes.[31] These devices can operate with different imaging modes, including white-light and narrow-band imaging. Such imaging methods are critical for identifying cancer, determining its stage, and tracking the response to therapy. In oncology research, imaging data primarily fall into two categories: radiological images and digitized pathology slides stored as whole slide images (WSI).[20]

  Radiological imaging is essential for obtaining detailed information about the location and spread of cancer throughout the body. In addition to determining where cancer is situated and how extensive it is, imaging helps assess tumor size and shape, track their progression over time, and evaluate how well treatments are working. Broadly, imaging can be divided into two main types: structural imaging, which produces images of the body's anatomy and morphology using modalities such as computed tomography (CT), magnetic resonance imaging (MRI), and mammography; and functional imaging, which visualizes the physiological activity of tissues and organs through techniques such as positron emission tomography (PET).[31]

In contrast, histopathological imaging focuses on analyzing tissue sections obtained through biopsy or surgical procedures.[38] Accurate pathology diagnosis through image analysis necessitates a detailed assessment of tissue organization and cellular features. Whole-slide imaging (WSI) offers comprehensive histopathological information, encompassing attributes such as morphology, texture, overall spatial arrangement, specific structural components, collagen configuration, and patterns of tumor-infiltrating lymphocytes (TIL).[29] Public databases like the Cancer Imaging Archive (TCIA) provide open access to such data.

- **Clinical report.**

  Clinical information for cancer patients is captured in detailed medical records that encompass medical histories, administered treatments, demographic data, and diagnostic and laboratory findings, among other elements.[39] When stored in a structured format, these records are organized systematically, including continuous variables such as age and tumor size, as well as discrete variables like race and metastasis status. A range of methods can be used to integrate continuous and discrete variables into a unified data representation. These data are stored in a centralized electronic health record (EHR) system within the clinic. In the EHR, time series data reflect clinical information collected at multiple time points, such as repeated blood tests, laboratory findings, or measurements of physical characteristics. Regularly updating these data supports ongoing monitoring of the patient's condition and assessment of cancer progression.

4. Multimodal alignment

Multimodal alignment focuses on establishing semantic relationships between different kinds of data. Its aim is to form coherent connections and maintain consistency across heterogeneous modalities. This alignment is achieved by measuring similarity between modalities and by handling possible long-range dependencies and ambiguities involved in mapping them to shared semantic representations. This step involves coordinating multiple modalities so that the information they convey is compatible and ready to be integrated.[17] The phrase "alignment before fusion" denotes the step in which connections and correspondences between elements from various modalities are identified.[22,23] Proper alignment of these modalities enables a more complete understanding and an accurate depiction of how they relate to each other. After alignment is achieved, the fusion stage leverages the aligned data to construct a more robust and comprehensive representation.

Alignment strategies are divided into three types: implicit, explicit, and hybrid. Explicit alignment generally uses similarity matrices to assess similarities directly. In contrast, implicit alignment involves methods applied as intermediaries, often in a latent fashion, throughout the execution of the principal task. In addition, alignment can be achieved by using explicit and implicit alignment simultaneously, a process known as hybrid alignment. This approach involves using one or more explicit alignments combined with implicit methods. This section provides a breakdown of alignment types (explicit, implicit, and hybrid), along with detailed explanations of the techniques and their respective advantages and disadvantages.

4.1 Explicit alignment

Explicit alignment denotes the process of aligning components by directly assessing their similarities through similarity matrices. This approach uncovers relationships between modalities by leveraging matrices or attention/retention mechanisms. It operates by matching sub-components across several modalities to form a unified overarching model. Numerous methods rely on the similarity between such sub-components from different modalities as a core building block. This type of alignment is generally divided into two categories: unsupervised and (weakly) supervised.[23]

The unsupervised approach functions without relying on explicit alignment labels between instances of different modalities. Instead, it uses methods such as dynamic time warping (DTW) or canonical correlation analysis (CCA). In contrast, the (weakly) supervised approach makes use of these alignment labels—even when they are incomplete or noisy—since it depends on labeled aligned instances to train similarity functions, which are essential for cross-modal alignment. Typical techniques in this setting include Gaussian mixture models (GMMs), self-supervised learning (SSL), and deep learning architectures such as LSTMs or CNNs. To investigate these ideas, this survey categorizes four selected papers into three main groups: CCA-based methods, optimal transport (OT)) or representation-based approaches. This subsection discusses these techniques in detail, examining their specific alignment mechanisms, and outlining their respective strengths and limitations Table 2.

4.1.1 Canonical correlation analysis (CCA)

Canonical Correlation Analysis (CCA) is a statistical technique designed to identify and quantify correlations between two sets of variables originating from different modalities. By discovering linear transformations of the input features that maximize the correlation between the transformed vectors, CCA uncovers latent relationships across modalities. This approach enables the extraction of shared underlying representations without requiring labeled data, making it particularly valuable in contexts where supervised annotations are scarce. In the context of cancer survival prediction, CCA has been leveraged to align histopathological imaging and genomic data by projecting them into a shared latent space where their associations are maximally correlated. This explicit alignment facilitates more meaningful fusion by capturing the complementary information embedded in each modality. A notable application, described by Subramanian et al.,[40] used penalized Canonical Correlation Analysis (pCCA) to integrate histological and genomic data to predict survival from breast cancer.

Within this framework, pCCA carries out prediction through a two-stage procedure. In the first stage, variants of CCA are applied to learn multidimensional joint embeddings of the imaging and genomic modalities

Table 2 Advantages and disadvantages of penalized Canonical Correlation Analysis (pCCA).

Framework	Advantages	Disadvantages
pCCA[12]	• Enables correlation discovery in high-dimensional, low-sample-size data by adding penalty constraints. • Incorporates prior domain knowledge (e.g., sparsity, graph structures) into the CCA framework.	• Computational complexity increases with penalty terms and optimisation becomes bi-convex (local minima issues). • Deflation schemes (e.g., Hotelling) may fail to ensure diversity of learned canonical weights without added orthogonality constraints.

in an unsupervised fashion. These embeddings capture correlated patterns across modalities by forming linear combinations of features that highlight shared variance. In the second stage, the resulting embeddings are then used to infer latent factors linked to patient survival outcomes. To improve the diversity and stability of these embeddings, the study proposes two new matrix deflation strategies that preserve orthogonality among canonical weight vectors across iterations. This iterative deflation guarantees that each embedding encodes a unique component of cross-modal correlation, thereby expanding the representation space with complementary information.

4.1.2 Optimal transport

A recent application of a framework called Optimal Transport (OT) in machine learning involves comparing and handling probability distributions.[41] The goal of OT is to identify the optimal transportation plan that determines how much of a discrete source distribution should be moved to a discrete target distribution for alignment.[42] This study will analyze two works that use the OT formulation for alignment, namely MOTCat and OTGL. The comparison between their advantages and disadvantages is described in Table 3.

Table 3 Advantages and disadvantages of MOTCat and OTGL frameworks.

Framework	Advantages	Disadvantages
MOTCat[3]	• Provides explicit alignment between histopathology and genomics via optimal transport, preserving global structural consistency. • Employs UMBOT to reduce computational complexity while approximating optimal transport solutions.	• Limited to two modalities (histopathology and genomics), lacking integration of additional data sources. • Computational cost remains high despite UMBOT approximation for very large WSI datasets.
OTGL[4]	• Introduces class tokens (global features) and instance tokens (local features) per modality along with aligns local tokens with UOT on micro-batches • Enhances interpretability through analysis of attention weights and attribution maps.	• Increased model complexity may challenge scalability and interpretability in clinical deployment. • Alignment accuracy may be sensitive to token initialization and transformer parameter choices.

4.1.3 MOTCat

The multimodal optimal transport-based co-attention transformer (MOTCat) studied by[3] introduces an explicit alignment mechanism that utilizes OT integrated within a co-attention transformer framework. This approach is designed to align histopathological and genomic data by establishing correspondences that preserve global structural consistency between the modalities. Specifically, OT is used to calculate an optimal matching solution that minimizes the total cost of aligning instances across the modalities, where the cost is derived from local pairwise dissimilarities between histopathology image patches and gene coexpression profiles.

Within this framework, those instances in a WSI whose structures most closely align with the corresponding genomic coexpression signals are selected to represent the WSI. To mitigate the substantial computational cost of solving OT problems on large WSI collections, MOTCat divides each WSI into smaller microbatches. These microbatches are handled separately, and their outcomes are then combined via an approximate method called UMBOT (Unbalanced Mini-batch Optimal Transport), which yields an efficient approximation of the global transport solution across all instances.

4.1.4 OTGL

The OTGL framework, also known as Optimal Transport with Global-Local Feature Fusion, was devised by[4] and takes inspiration from the alignment strategies of MOTCat.[3] The OTGL approach enhances multi-modal methodology by aligning three distinct data types (pathology, radiology, and genetic data) to provide a more comprehensive prediction of cancer survival. However, it does not use them simultaneously to predict survival. Furthermore, OTGL introduces novel advances using CMTA, drawing inspiration from the encoders studied by Zhou et al.[1]

OTGL uses OT to align the data by identifying which tissue regions, gene signatures, or imaging patterns correspond the most closely between patients. By applying OT at a fine-grained level, the framework identifies pathological patches and molecular signals that are most representative of each other, even when drawn from different sources such as radiology and gene expression. This alignment seeks to derive an OT solution that minimizes modality discrepancies by calculating instance-wise distances across histopathology, genomics, and radiology features. The resulting optimal transport matrix highlights regions within pathological images that receive higher alignment weights, effectively identifying the patches most representative of the biological signals captured in the other modalities.

4.1.5 Self-supervised representation learning

Multimodal self-supervised learning (SSL) offers a powerful approach to learning robust representations of oncological data without requiring extensive labeled datasets. Using intrinsic structures and correlations within and between modalities, SSL enables models to align and integrate heterogeneous biomedical data in an unsupervised or weakly supervised manner. Recent advances in multimodal SSL have demonstrated promising results in both natural and medical imaging domains, facilitating explicit modality alignment while preserving complementary information. In this section, we examine one representative approach, namely Multimodal PathologIcal self-supposed representation learning through alignment of mOdality and Retention (MIRROR).[7] Table 4 summarizes its main advantages and disadvantages.

MIRROR demonstrates explicit alignment through an unconventional SSL framework that is purpose-built to integrate histopathology and transcriptomic data by incorporating a modality retention module. This explicit alignment strategy preserves the unique characteristics of each modality, while enabling the alignment module to concentrate on common, cross-modal representations. In doing so, it tackles key issues in multimodal cancer analysis, such as the scarcity of paired data annotations, the mitigation of redundancy between histopathology and transcriptomic signals, and the danger of discarding modality-specific information during the alignment process.[43] MIRROR technically employs a specialized transformer architecture, drawing inspiration from,[44] and incorporates attention mechanisms to capture interactions both within and between modalities. The model undergoes self-supervised pretraining to autonomously align features across different modalities, achieving this alignment without direct supervision.

Table 4 Advantages and disadvantages of MIRROR framework.

Framework	Advantages	Disadvantages
MIRROR[7]	• Simultaneously preserves both shared and modality-specific information by balancing modality alignment and retention. • Operates effectively in in data-scarce settings via self-supervised learning without extensive labelled data.	• Pretraining objectives require precise tuning to avoid under- or over-alignment. • Computational complexity increases with large datasets and high-dimensional inputs.

4.2 Implicit alignment

Implicit alignment denotes alignment approaches that avoid explicitly forcing a direct mapping between modalities. Instead, they obtain alignment by learning a shared latent representation space.[17] Serving as an intermediate (often latent) stage within another task, implicit alignment does not depend on supervised alignment pairs; rather, it acquires latent alignment of the data throughout model training.[23] These methods aim to bridge heterogeneity between data types by enabling cross-modal interactions within learned feature spaces, rather than performing explicit instance-level matching. In survival prediction, implicit alignment is typically achieved using neural network-based techniques, cross-modal translation, contrastive learning, or graph-based methods. Below, we explain each of these approaches and discuss their advantages and limitations.

4.2.1 Neural network

This subsection highlights implicit alignment methods that are based on neural network frameworks. In particular, we explore two studies, namely the Mutually Guided Cross-Modality Transformer (MGCT) and the Dual Stream Cross-Modal Alignment Network for Survival Prediction (DSCASurv), which employ neural network layers and design inspired by transformers to implicitly align histopathology with genomic data to predict survival outcomes. A summary of their respective strengths and weaknesses is presented in Table 5.

4.2.2 MGCT

This approach introduced by[45] addresses the task of integrating histopathology and genomic information using dual sequence streams within the transformer layer. Initially, each modality undergoes a separate feature extraction: Histopathological WSIs are handled by a convolutional neural network (CNN), while genomic data is modeled by a self-normalizing neural network (SNN). The key to integration is mutual guided cross-modality attention (MGCA) which simultaneously processes genomic feature embeddings and extracts features from WSIs. To highlight pertinent signals, a gated attention pooling (GAP) mechanism is used during this stage. This cycle of MGCA is then iterated twice.

4.2.3 DSCASurv

This technique proposed in[12] features a convolutional mamba token mixer (CM-Mixer) that inherently aligns the WSI bags with gene expression data.

Table 5 Comparison of advantages and disadvantages of MGCT and DSCASurv frameworks.

Framework	Advantages	Disadvantages
MGCT[45]	• Employs Mutual-Guided Cross-Modality Attention (MGCA) to capture salient interactions between histopathology and genomics. • Implements iterative fusion and attention pooling to progressively refine cross-modal representations.	• Requires complex recursive fusion and attention mechanisms, increasing computational overhead. • Lacks explicit spatial correspondence between gigapixel histopathology and genomic data, limiting interpretability.
DSCASurv[12]	• Combines convolutional and state-space models to integrate local and global features across modalities. • Dual-stream architecture enables parallel processing of histopathology and genomics for balanced alignment.	• Cross-modal alignment depends on hyperparameter tuning and balanced feature weighting to avoid overfitting. • Increased model complexity may challenge scalability in large-scale datasets.

Within the framework of state space sequence model (S4) and mamba, these parallel mixers are structured based on the encoder-decoder model to extract cross-modal public information. The initial mixer integrates WSI embeddings with genomic pathway summaries. By employing input-dependent depthwise convolution (IDConv) alongside Bi-Mamba, the outcome of this process feeds into self-attention. Subsequently, these components are combined and the squeezed token enhancer (STE) is used to enhance the distance between them; they are stored within the softmax matrix. The final CM-Mixer determines which groups are stratified as low-risk or high-risk.

4.2.4 Cross-modal translation

Cross-modal translation techniques facilitate interaction and information exchange across different modalities by enabling one modality to inform or predict features of another. In the context of multimodal cancer survival prediction, this approach leverages attention-based mechanisms to integrate complementary information while fostering cross-modal alignment. An integral part of numerous cross-modal translation models is the cross-modal attention module. This module enables features from one modality to focus on pertinent signals in another modality, thereby improving the learning of joint representations. This section emphasizes three key methods: Cross-Modal Translation and Alignment (CMTA), Cross-Aligned Multimodal Representation (CAMR), and PathoGen-X; which utilize cross-modal translation within their integration and alignment frameworks. Table 6 presents a summary of their specific strengths and weaknesses.

4.2.5 CMTA

CMTA, proposed by,[1] does not enforce an explicit geometric or distribution-level alignment between modalities. Rather, alignment arises implicitly through a cross-modal translation framework. Each modality is first encoded separately, preserving its internal structure. Cross-modal

Table 6 Advantages and disadvantages of CMTA, CAMR, and PathoGen-X frameworks.

Framework	Advantages	Disadvantages
CMTA[1]	• Explicitly aligns modalities via parallel encoder-decoder translation with cross-modal attention. • Enforces representation consistency with an L_1 alignment constraint, improving biological interpretability.	• Requires substantial paired multimodal training data, limiting applicability in low-resource settings. • Computational demands increase with gigapixel whole-slide image processing.
CAMR[8]	• Utilises adversarial alignment to learn modality-invariant representations across heterogeneous data types. • Preserves modality-specific features via orthogonality constraints and a reconstruction module.	• Sensitive to adversarial training instability; requires careful balance of discriminator and encoder losses. • Increased model complexity may affect interpretability and training convergence.
PathoGen-X[6]	• The alignment and translation methods can be adapted for other downstream tasks. • Reduces paired-data dependency by translating pathology features into genomic space.	• Translation accuracy is highly dependent on the quality of histopathology-derived features. • Introduces some redundancy during pre-training that can hinder its effectiveness.

attention is then applied to translate information from one modality into the representational space of the other, resulting in modality-specific cross-modal embeddings. Alignment is achieved indirectly by encouraging these translated embeddings to remain close to their respective intra-modal embeddings via a similarity-based loss. Consequently, CMTA focuses on aligning modalities in terms of semantic consistency rather than strict feature-level matching, enabling the exchange of complementary information without collapsing them into a single shared latent space.

4.2.6 CAMR

Proposed by,[8] this work presents another novel implicit alignment strategy that leverages an alignment learning network for cross-modality representation. The method maps heterogeneous modality representations into a shared subspace, where discrepancies in their distribution can be substantially reduced through an adversarial learning process. This approach is based on the work of,[46] where a cross-modality representation alignment learning network formulates an adversarial interaction between a shared encoder and two discriminators. The shared encoder is primarily responsible for learning modality-invariant representations within a common subspace while simultaneously attempting to mislead the discriminators. In contrast, the two discriminators aim to tell one modality apart from the others, thus guiding the encoder toward improved learning of modality-invariant representation. The shared encoder produces modality-invariant representations that decrease the variability between representations of different modalities, thus reducing the gaps between them.[47]

4.2.7 PathoGen-X

PathoGen-X, introduced by,[6] proposes a cross-modal genomic feature translation and alignment framework designed to improve survival prediction from pathology images. Rather than projecting both modalities into a shared latent space, PathoGen-X implicitly translates pathology-derived features into the genomic feature space using transformer-based models. This translation allows the model to approximate genomic signals from imaging data, thus enriching weaker pathological signals with complementary genomic information. A key advantage of PathoGen-X lies in its ability to train using paired pathology and genomic data while enabling inference solely from histopathology images during testing. This design improves real-world clinical applicability by reducing the dependence on multimodal data in deployment.

To achieve alignment, PathoGen-X incorporates a composite loss function that combines latent loss (L_l) and translation loss (L_t), ensuring that pathology-derived features align accurately with genomic representations. The latent loss enforces the similarity between pathology and genomic latent embeddings by minimizing a weighted combination of Kullback-Leibler divergence and Euclidean distance. Meanwhile, translation loss ensures that the projected genomic features of the decoder align closely with the original genomic embeddings. This loss ensures that the representation of pathology and genomic data is consistent and robust.

4.2.8 Contrastive learning

Contrastive learning is a powerful self-supervised learning paradigm that trains models by optimizing their ability to distinguish between similar (positive) and dissimilar (negative) sample pairs.[48] Positive pairs typically represent related or semantically similar instances (e.g., data from the same patient), while negative pairs denote unrelated or dissimilar instances. During training, the model is encouraged to minimize the distance between positive pairs while maximizing the distance between negative pairs in the embedding space.[49] The objective of this enables the model to learn robust and discriminative representations even in the absence of explicit supervision. In the context of cancer survival prediction, contrastive learning offers a flexible and effective alignment strategy by implicitly organizing multimodal data in a shared latent space. By aligning samples across modalities based on their semantic similarity, contrastive learning accommodates uncertainty inherent in censored survival data and facilitates the discovery of meaningful cross modal patterns.

Given its versatility, contrastive learning-based alignment has gained increasing attention in recent multimodal survival cancer studies. This subsection highlights three representative methods: HySurvPred (Multimodal hyperbolic embedding with angle-aware hierarchical contrastive learning and uncertainty constraint for survival prediction), CPathomic (Contrastive learning-based multimodal fusion approach), and MoSaRe (Mixture of Experts, Symmetric Alignment, and Reconstruction). A comparison of their key advantages and disadvantages is presented in Table 7.

4.2.9 HySurvPred

The authors of[48] proposed HySurvPred, a novel framework to predict cancer survival by using contrastive learning within a hyperbolic embedding space. The implicit alignment of this approach aims to improve feature

Table 7 Advantages and disadvantages of HySurvPred, CPathomic, and MoSaRe frameworks.

Framework	Advantages	Disadvantages
HySurvPred[48]	• Utilizes hyperbolic embedding to capture hierarchical structures across histopathology and genomics. • Incorporates angle-aware contrastive loss and censored uncertainty constraint for improved survival prediction with censored data.	• Hyperbolic space modeling introduces greater mathematical and computational complexity. • Requires large and balanced datasets to fully exploit the benefits of hyperbolic geometry.
CPathomic[13]	• Cross-modal contrastive learning aligns heterogeneous features while preserving modality-specific signals. • Cross-modal attention selectively emphasizes complementary interactions, enhancing interpretability.	• Relies on complete paired data for training, reducing robustness under missing modality conditions. • Sensitive to feature extraction quality; variability in WSI or genomic data may affect alignment accuracy.
MoSaRe[50]	• Adaptive mixture-of-experts fusion handles missing modalities and incomplete data effectively. • Symmetric and multi-prototype contrastive learning preserves biological heterogeneity in the embedding space.	• Computational demands increase due to multiple alignment and reconstruction objectives. • Requires careful tuning of multi-stage losses and fusion mechanisms for optimal performance.

representation in hyperbolic mapping by proposing the Angle-Aware Ranking-Based Contrastive Loss (ARCL) module and using ranking-based contrastive learning to preserve the ordinal nature of survival time with the Censor-Conditioned Uncertainty Constraint (CUC).

HySurvPred uses a multimodal hyperbolic mapping module (MHM) to capture local and global hierarchical relationships between histopathology and genomic data. By embedding samples into hyperbolic space, the model leverages geometric properties that naturally encode hierarchical relationships, aligning patient representations in a biologically meaningful manner.

4.2.10 CPathomic

The CPathomic approach[13] presents an innovative multimodal framework that combines histopathological images of the entire slide and genomic data to predict cancer survival. A key innovation of CPathomic lies in its cross-modal contrastive learning module, inspired by the CLIP framework. This module aligns histopathological and genomic representations in a shared latent space by pulling embeddings from the same patient closer together (positive pairs) and pushing embeddings from different patients apart (negative pairs).

This contrastive alignment facilitates semantic consistency across modalities while preserving modality-specific information. Beyond alignment, CPathomic employs a cross-modal attention mechanism to enable dynamic interaction between histopathology and genomic features. This mutually guided attention selectively emphasizes complementary signals, capturing subtle cross-modal relationships, such as morphological patterns linked to gene expression changes, thus enriching the fused representation for survival prediction.

4.2.11 MoSaRe

This technique, introduced in,[50] offers a flexible framework for multimodal survival prediction that is capable of handling missing or incomplete modalities. The architecture comprises five integrated modules that span feature extraction, adaptive fusion, and deep alignment. For alignment, MoSaRe implements symmetric contrast learning (SymCL) and multi-prototype contrast learning to enforce sample-level consistency while preserving biological heterogeneity in the embedding space. These strategies align both global and local representations without collapsing them into trivial solutions, maintaining biologically meaningful distinctions between patient profiles. The model jointly optimizes a composite objective that combines symmetric contrast loss, multiprototype contrast loss, reconstruction loss, and classification loss, achieving robust and interpretable alignment for survival prediction.

4.2.12 Graph-based

Implicit alignment across modalities using graph-based approaches offers a powerful yet complex strategy for integrating heterogeneous biomedical data. By modeling multimodal data as graphs where data elements are nodes and their relationships are edges, these methods capture intricate structural and relational dependencies that are not readily accessible through conventional matrix or vector representations. Such graph structures provide a flexible framework to encode both intra-modal and inter-modal interactions, making them well-suited for aligning heterogeneous signals without requiring explicit one-to-one correspondence.

Despite their advantages, graph-based approaches face challenges due to the irregular, sparse, and dynamic nature of graph topologies, which complicates optimization and increases computational demands.[19] Memory constraints, scalability, and convergence difficulties are common issues when operating on large, high-dimensional graph representations. However, the interpretability offered by graph nodes and edges linking biological entities, molecular features, or clinical variables makes this approach particularly valuable for biomedical applications. In this section, we review three representative graph-based alignment methods, namely The Hybrid Graph Convolutional Network (HGCN), Dropping and fearure Alignment based Graph (DAGraph), and Counterfactual Biderctional Co-Attention Transformer (CFBCT). A comparison of their key advantages and disadvantages is summarized in Table 8.

4.2.13 HGCN

Proposed in,[11] HGCN tackles the issue of missing modalities in real-world clinical datasets by combining a graph-based framework with a masked autoencoder architecture. It offers an implicitly aligned multimodal fusion scheme designed for cancer survival prediction when patient data are incomplete. In this framework, each patient's information is first represented as modality-specific graphs. Pathology images are converted into graph-structured features via convolutional neural networks, clinical variables are transformed into vectors and encoded using one-hot connectivity, and genomic data are mapped into embeddings via gene set enrichment methods. This graph-centric encoding maintains the intrinsic structure of each modality, allowing interpretable intra-modal message passing through separate graph-convolutional networks (GCNs).

4.2.14 DAGraph

Zhang et al.[51] introduce an inter-modal implicit alignment objective that matches the global representations of a pathology graph with those of a genomics graph. The underlying intuition is that gene expression influences cell proliferation and growth, and that these biological processes can be reflected as morphological patterns in WSIs. For the pathology modality, WSIs are partitioned into patches that act as nodes in a slide-level graph;

Table 8 Advantages and disadvantages of HGCN, DAG, and CFBCT frameworks.

Framework	Advantages	Disadvantages
HGCN[11]	• Robust to missing modality data via integrated Online Masked Autoencoder during model training. • Combines graph and hypergraph networks for structured, process-aligned fusion across modalities.	• Does not explicitly correct modality-induced biases or confounding effects. • May underutilize causal relationships and interdependencies across modalities.
DAG[51]	• Extract robust and general feature via mininmising the feature discrepancy between original and modified graph by introducing intra-modal cosistency loss. • Reduce the feature heterogeneity by adopting an inter-modal alignment loss.	• Control the stability the model based on random node dropping in patch and pathway level. • Not fully capture the variations caused by diverse factors in clinical scenarios.
CFBCT[52]	• Explicitly reduces spurious modality biases through counterfactual alignment and causal inference. • Enhances multimodal integration via bidirectional co-attention at both instance and distribution levels.	• Requires complete modality data for optimal performance; less robust under missing data conditions. • More complex training pipeline with increased computational and modeling requirements.

edges are constructed based on local spatial neighborhoods (e.g., 8-neighbor connectivity), and node embeddings are updated using graph attention. For the genomics modality, gene expression profiles are aggregated into pathway-level nodes, with edges defined by shared genes; these pathway nodes are likewise updated via graph attention. Each modality is then summarized by pooling the attention-refined features, yielding global representations at the slide level and at the pathway level.

4.2.15 CFBCT

Implicit alignment in CFBCT[52] emerges from its formulation of causal graphs rather than from an explicit matching module. Although CFBCT does not incorporate a dedicated alignment component in the conventional sense, correspondence between histological and genomic modalities is implicitly achieved through the causal graph structure and its bidirectional co-attention mechanism. Instead of imposing fixed pairings between WSI instances and genomic variables, CFBCT reframes multimodal fusion as a causal process in which cross-modal dependencies surface as mediated effects.

From this perspective, CFBCT can be seen as performing an implicit, causality-guided alignment. Instead of explicitly representing feature-level structural correspondences in the graph, it regulates how multimodal information is combined by specifying which cross-modal interactions are causally important for survival prediction. Consequently, alignment is not an isolated architectural component, but a task-dependent, learned property that emerges within the model's causal reasoning process.

4.3 Hybrid alignment

Hybrid alignment integrates both explicit and implicit alignment methods to better manage the complexity of multimodal data fusion.[17] This approach is particularly critical in multimodal learning (MML) for cancer survival prediction, where data modalities vary greatly in their structure and scale. Histopathology whole-slide images are often analyzed with multiple instance learning (MIL) frameworks,[53] yielding thousands of patch-level representations per patient. In contrast, transcriptomic profiles are usually modeled with feedforward neural networks or organized into biologically meaningful structures, such as gene families and signaling pathways.[54] Because these modalities differ in dimensionality, statistical properties, and underlying relational organization, achieving direct alignment between them is inherently challenging.

The challenge becomes more pronounced when large token sets from both modalities are fed into transformer-based models. While attention and cross-attention modules enable flexible, implicit alignment by learning interactions between modalities, their computational cost scales sharply with the number of tokens. Techniques such as token subsampling or optimal transport (OT)-based matching can partially reduce this load by enforcing more structured correspondences, but they introduce additional optimization complexity and still do not fully address the scalability problem. As a result, hybrid alignment has emerged as a compromise: it combines the adaptability of learned attention with structured constraints that help steer cross-modal matching.

In this survey, recent hybrid alignment methods can be grouped into three broad directions: (1) integrating OT-based cross-alignment with transformer attention within unified architectures; (2) combining cross-modal attention with graph-based modeling to incorporate structural or biological priors; and (3) coupling encoder-based representation learning with contrastive objectives to encourage global cross-modal consistency. The following section organizes these approaches into three categories, outlining their methodological foundations and comparing their strengths and limitations, as summarized in Table 9.

4.3.1 Optimal transport and cross-modal attention

In this survey, Multimodal Prototyping (MMP)[18] will be explored as a hybrid alignment framework that condenses WSI into morphological

Table 9 Advantages and disadvantages of MMP, PGHG, and mSTAR frameworks.

Framework	Advantages	Disadvantages
MMP[18]	• Integrates optimal transport and transformer attention for cross-modal alignment at prototype level. • Enables unsupervised learning of morphological and pathway-level prototypes without labeled data.	• Computationally demanding due to high token dimensionality and transformer complexity. • Requires careful design of prototype aggregation to avoid information loss.
PGHG[55]	• Models histopathology and genomics as interconnected graphs, enabling interpretable relational structures. • Combines graph attention and contrastive learning for robust intra-modal and inter-modal alignment.	• High computational complexity due to heterogeneous graph construction and exhaustive connectivity. • Dependent on quality of gene set variation analysis to guide feature extraction.
mSTAR[43]	• Integrates microscopic, macroscopic, and molecular modalities to mirror clinical diagnostic workflows. • Dual-stage pretraining enhances both slide-level and patch-level representations with clinical context.	• Increased training complexity and resource requirements due to multi-stage pretraining. • May require large multimodal datasets to achieve generalizable pretraining benefits.

prototypes while summarizing transcriptomic profiles through prototypes of biological pathways. This unsupervised approach integrates information at the pathway and morphology levels, aligning them via optimal transport.

In MMP, optimal transport facilitates cross-modal alignment by matching morphological and pathway prototypes, enabling the model to bridge histology and transcriptomic representations. This alignment operates alongside dense intra-modal and cross-modal interactions modeled by transformer attention mechanisms, reflecting conceptual parallels between attention maps and transport plans.

4.3.2 Cross-modal attention and graph learning

A critical goal of implementing this approach is to develop intra- and inter-modal correlations that align with known biological relationships, enabling effective multimodal fusion. Hybrid alignment methods have taken advantage of cross-modal attention to aggregate information to biologically meaningful pathway nodes while constructing global representations of pathway subgraphs. As an example, in this survey the study by Zhang et al.[55] will be explored as this type of hybrid alignment. It proposes the Pathology-Genome Heterogeneous Graph (PGHG), a biologically grounded framework that links deep representation learning with heterogeneous graph modeling.

In PGHG, histological image patches and genomic pathways are represented as nodes within separate but interconnected subgraphs, with edges encoding spatial relations among image patches and gene-sharing relationships among pathways. Full inter-modal connectivity is established by linking every heterogeneous pair of nodes, facilitating extensive cross-domain information exchange. The model representation learning process is guided by Gene Set Variation Analysis (GSVA), directing the network to prioritize genome-relevant histopathological features while suppressing irrelevant variability. An attention-based pooling mechanism aggregates pathway-level features into global representations, which are then projected into a shared embedding space. Cross-modal alignment is enforced using a contrastive loss objective that encourages similarity between global pathology and genomic representations. Additionally, a graph attention mechanism allows the model to progressively integrate intramodal and inter-modal features across layers. The overall training objective combines survival prediction loss, embedding alignment loss, GSVA supervision loss, RNA expression reconstruction loss, and adjacency matrix reconstruction loss, providing a comprehensive and biologically interpretable alignment strategy.

4.3.3 Encoder and contrastive learning

Hybrid alignment approaches have also emerged from the integration of encoder with contrastive learning, addressing the underutilization of multimodal data during model initialization.[56] Clinical diagnostic workflows often involve multimodal reasoning, yet conventional computational pathology models remain limited by isolated patch-level characteristics and lack a broader clinical context. Recent work has introduced pre-training strategies that align histopathology and genomics at both the slide and patch levels, enabling richer and more integrated multimodal representations. This approach is denoted mSTAR for self-taught multimodal pretraining.[43]

It is a foundation model that integrates microscopic (slides), macroscopic (text reports), and molecular (gene expression) data. Inspired by clinical diagnostic workflows, mSTAR aims to capture cross-scale cancer pathology representations. mSTAR implements a two-stage pre-training process. In the first stage, slide-level contrastive learning aligns pathological images, reports, and gene profiles, enabling holistic multimodal understanding. In the second stage, self-taught patch-level pre-training propagates aligned slide-level knowledge into fine grained patch embeddings, overcoming limitations of isolated patch-based extraction. This dual stage strategy equips mSTAR with both global and local context, enhancing interpretability and predictive performance.

5. Multimodal fusion

Multimodal fusion is the process of integrating data from several modalities to produce a single, consistent prediction. Rather than handling each modality independently within a model, this approach merges information from diverse sources, which has been shown to enhance predictive accuracy.[57] By aggregating and combining complementary contextual information across modalities, multimodal fusion yields more unified predictions and supports better decision-making. This strategy is particularly crucial in cancer prognosis, where the interpretation of results can change substantially depending on how findings from different modalities are combined. For example, the mutation status or histological profile of IDH1 alone cannot adequately explain the variations in patient outcomes. In contrast, their combination has been widely used to revise the WHO classification for diffuse glioma.[58]

The fusion of multimodal data, fundamentally rooted in information theory, consists of integrating diverse data sources to construct an information state that exploits their complementary strengths.[59] Data integration can be performed at multiple stages: early (feature level), intermediate (model level), late (decision level), or via a hybrid strategy that combines two or more of these stages. Each fusion stage presents specific benefits and limitations, determined by the nature of the data and the requirements of the task. In this section, the 18 alignment-based studies from the previous chapter are compared with the 13 works that employ fusion strategies without alignment. A detailed examination of these studies, including an assessment of their strengths and weaknesses for survival analysis, is provided in the next section.

5.1 Early fusion

Early fusion is the most straightforward strategy for integrating information from multiple modalities. In this approach, features from different data sources are combined at the input stage, before model training begins. All modalities are merged into a single feature vector, which is then fed into one unified model. Because fusion occurs at the feature level, early fusion is often referred to as feature level fusion. Its effectiveness depends on whether the model can process heterogeneous inputs jointly and whether the modalities provide complementary rather than redundant information.

The input features can take various forms, including raw data, handcrafted descriptors, or deep features extracted from pretrained networks. To construct a joint representation, common operations include simple vector concatenation, element-wise addition, and multiplicative interactions such as Hadamard or Kronecker products. Although conceptually simple, early fusion requires careful consideration of feature compatibility. Differences in scale, dimensionality, or statistical distribution between modalities can negatively affect learning if not properly normalized or aligned.

In this survey, we review several representative early fusion approaches, including Pathomic Fusion, Multimodal Disentangled Representation Learning (MMDR), and the multimodal multi-instance learning framework for cancer survival prediction (MMsurv). We compare their design choices, advantages, and limitations in Table 10. Additionally, methods such as pCCA, MIRROR, and MMP—previously discussed in the Alignment section—are revisited here with specific attention to how they implement early fusion at the feature level.

Table 10 Comparison of early fusion techniques.

Authors and year	Advantages	Disadvantages
MMDR[9]	Uses a shared encoder plus private encoders and a disentangle contrastive loss to make cross-modal shared representations similar while keeping them distinct from private ones.	The study uses pre-extracted features and reduces each modality to 80 features for experiments, so it does not model whole-slide tiles via MIL or similar patch-level mechanisms.
Pathomic Fusion[60]	Introduced Pathomic Fusion, which enables interpretable end-to-end multimodal integration of histopathology and genomics using a gating-based attention mechanism and Kronecker product for interaction modeling.	Requires training of separate unimodal networks and may be computationally intensive due to Kronecker-based fusion.
MMSurv[61]	Combines multi-instance and multimodality learning; captures global and local information; scalable for heterogeneous inputs.	Model complexity challenges interpretability; training stability may require careful tuning.

5.1.1 Pathomic fusion

The study by[60] proposes a novel framework for the multimodal fusion of histology and genomic characteristics by implementing an early strategy called pathomic fusion. The fusion occurred whilst pairwise feature interactions across modalities by taking the Kroonecker product of gated featured representations, and controls the expressiveness of each representation using a gated-based attention mechanism. Pathomic Fusion employs a deep feature-level fusion strategy, often categorized as early-stage integration since modalities are combined prior to survival prediction. Pathomic Fusion is a deep learning framework designed to integrate histopathological imaging and genomic data for a more accurate cancer prognosis. Recognizing the complementary value of tissue morphology and molecular alterations, the model combines convolutional neural networks (CNNs) to extract features from whole-slide images, graph convolutional networks (GCNs) to model cellular architecture, and self-normalizing networks (SNNs) to learn genomic profiles.

The framework employs an early-fusion strategy in which the features of each modality are first processed independently, allowing the modality-specific representations to develop fully before being combined. These processed embeddings are then fused using a multimodal tensor product, allowing the model to capture complex interactions across modalities, including unimodal, bimodal, and trimodal relationships. Applied to data sets such as TCGA glioma and renal carcinoma, Pathomic Fusion showed superior performance in survival prediction and patient risk stratification compared to unimodal and conventional approaches by obtaining CI results 0.826 ± 0.009.

5.1.2 MMSurv

Proposed in,[61] MMSurv is an early fusion prediction approach for cancer survival that integrates pathological images, clinical records, and genomic sequencing data. It adopts a two-stage multi-instance learning (MIL) model that identifies prognostically relevant tissue regions within whole slide images, while simultaneously incorporating structured clinical variables and high-dimensional gene expression data. Clinical features are encoded using word embedding techniques borrowed from natural language processing to better capture semantic relationships.

To fuse these heterogeneous inputs, the model introduces a compact bilinear pooling mechanism integrated with a transformer-based module (or called as MMF-CBPT), allowing nuanced interaction among modalities. MMSurv improves interpretability by using tile-level attention scores and analyzing cellular composition through nucleus segmentation. This method achieves an overall concordance index (CI) of 0.7283, with specific CIs for six types of cancer being: BRCA (0.7643), COAD (0.782), ESCA (0.7803), LIHC (0.6864), LUAD (0.6927) and STAD (0.6641).

5.1.3 MMDR

Proposed in,[9] MMDR offers an interpretable strategy to integrate histopathological images, gene expression, and copy number alteration (CNA) data to improve the prediction of cancer survival. Unlike traditional fusion models that combine all modalities into a single representation, MMDR explicitly performs the early fusion by disassembling the information into two biologically significant components: (1) modality-shared representations that capture the consensus or overlap information between different data types and (2) modality-specific representations that preserve the unique and complementary characteristics of each modality. This approach reflects the clinical intuition that while different data types can converge on shared signals of disease severity (e.g., molecular markers visible in both tissue morphology and gene profiles), each also retains distinct diagnostic insights.

MMDR employs dedicated encoders for each component and applies a disentanglement loss to promote a clear separation between shared and specific signals. Furthermore, reconstruction loss ensures that key unimodal information is not lost during the learning process, aligning with the clinical principle of data traceability and preservation of source integrity. The final representations are then concatenated and passed to a survival prediction model using the Cox proportional hazards framework. Employing this method, MMDR achieves a CI of 0.838, surpassing the performance of cutting-edge techniques like Pathomic Fusion and DeepSurv.

5.1.4 pCCA

This method utilizes early fusion, since it integrates RNA sequences and tissue-level histopathology images right from the raw data stage. The two types of data are combined into a latent variable and processed sequentially in two stages. The initial stage involves estimating the model parameters through CCA before the prediction. In the second stage, CCA canonical weights are directly utilized to produce two-modality embeddings for input into the predictors.

This final stage is then expanded to incorporate pCCA embeddings to predict latent data by introducing two innovative matrix deflation techniques. pCCA creates a biologically coherent basis for predicting patient outcomes. This method is especially effective in scenarios where sample sizes are limited, even though it still needs a vast array of features. The precision of the method was assessed using orthogonalized projected deflation (OPD), achieving an accuracy rate of about 94.32 with a standard deviation of approximately 1.08.

5.1.5 MIRROR

This method uses a model-driven approach for early integration, which takes place immediately after the modality encoders complete their feature extraction. The system comprises three components: alignment, retention, and style clustering. Integration is performed during the construction of the self-supervised learning framework. MIRROR is designed to perform efficiently in data-scarce settings, enhancing its applicability in both clinical practice and research. employing a ResNet-50 architecture and analyzing datasets from NSCLC, BRCA, COADREAD, and RCC, MIRROR achieved accuracies (ACC) of 0.992 ± 0.011, 0.958 ± 0.017, 0.902 ± 0.039, 0.938 ± 0.022 and 0.998 ± 0.003, respectively.

5.1.6 MMP

Described as early fusion, MMP first matches the dimensions of tokens from each modality by utilizing linear projection for histology and multilayer perceptron (MLP) and self-normalizing neural network (SNN) for pathway prototypes. Subsequently, interactions between dense intra-modal and cross-modalities will be learned by harnessing transformer attention and optimal transport cross-alignment. This tokenization framework, which is based on prototypes, yields CI values for various cancers as follows: BLCA: 0.753 ± 0.069, BRCA: 0.635 ± 0.051, LUAD: 0.643 ± 0.013, STAD: 0.598 ± 0.051, CRC: 0.636 ± 0.120, and KIRC: 0.748 ± 0.099.

5.2 Late fusion

In late fusion, called decision-level fusion, distinct models are developed for each modality, with their predictions combined to produce the final result.[62] Each modality-specific model generates an independent output that contributes to decision making in the late fusion stage.[26] This aggregation process might involve majority voting, associating weights with each vote, or employing different machine learning algorithms.[38] Here, all independent decisions are synthesized to reach a conclusive decision. This approach is suitable when the modalities provide complementary information, but are not necessarily independent of each other. In cases of missing or incomplete data, late fusion retains the ability to make predictions, since each model is trained separately. The primary difficulty with late fusion lies in choosing a suitable approach for merging the separate predictions. To illustrate the research on late fusion, we will examine various study cases, including MSEN and MIFAPS. A comparative analysis of their advantages and disadvantages will also be included in Table 11. Furthermore, we will introduce DAGraph, for which its alignment technique was previously introduced, primarily to emphasize its late integration, survival prediction strategy, and obtained results.

5.2.1 MSEN

A recent study introduced a new late fusion technique called the Multimodal Survival Ensemble Network (MSEN),[63] which employs a weakly supervised

Table 11 Comparison of late fusion techniques.

Authors and year	Advantages	Disadvantages
MSEN[63]	Proposed MSEN model that integrates histopathology and genomic data with transformer-based architectures and ensemble learning, demonstrating improved generalizability in survival prediction across diverse cohorts.	Model complexity due to ensemble and transformer mechanisms may pose challenges for interpretability and reproducibility.
MIFAPS[64]	Developed MIFAPS, a fully automated and multimodal system incorporating MRI, WSI, and clinical data to predict pathological complete response (pCR), achieving high performance in internal and prospective validations.	May have limitations in generalizability beyond breast cancer due to domain-specific data and clinical focus.

approach to effectively integrate genomic and histopathological data. This model enhances the combination of features in the output layer to create an innovative encoder system that specializes in detailed genomic feature extraction. MSEN incorporates extensive histopathological image patches into the multimodal framework. This framework achieved the highest CI scores for five datasets: BLCA (0.662 ± 0.042), BRCA (0.648 ± 0.045), UCEC (0.661 ± 0.063), GBMLGG (0.827 ± 0.018) and LUAD (0.631 ± 0.080).

5.2.2 MIFAPS

The fully automated multimodal integrated pipeline system (MIFAPS)[64] is a deep learning-based framework designed to predict the complete pathological response (pCR) to neoadjuvant chemotherapy in patients with locally advanced breast cancer. The system integrates information from three key sources: magnetic resonance imaging (MRI), whole-slide histopathology images (WSI), and clinical risk factors. Each modality contributes unique information on tumor characteristics and their combination aims to provide a more holistic understanding of treatment response. MIFAPS adopts a late fusion strategy in which modality-specific characteristics are extracted independently using convolutional neural networks for imaging data and logistic regression for clinical variables. These features are then combined at the decision level, allowing each modality to preserve its representation before integration. This approach facilitates robust multimodal learning by gaining the AUC 0.994.

5.2.3 DAGraph

Using transformer-based fusion, DAGraph essentially achieves late fusion.[51] Fusion proceeds after graph updates at the representation level. Node embeddings from the pathology and genomics graphs are concatenated and passed through a transformer fusion layer that models long-range, cross-modal interactions among nodes. To improve robustness within each modality, DAGraph performs node drop augmentation by randomly removing a portion of graph nodes and penalizing discrepancies between the original and augmented global representations. This consistency objective is applied independently to pathology and genomics, encouraging stability to patch sampling, staining variability, and expression noise. Beyond intra-modal robustness, the method adds an inter-modal alignment loss that directly pulls the global pathology and global genomics embeddings into correspondence. This "drop-and-align" combination reduces heterogeneity across data types while preserving modality-specific information that remains predictive of survival. The fused representation is then fed to a discrete-time survival head trained with a negative log-likelihood loss; the total objective combines survival, consistency, and alignment terms with tunable weights. This method achieves an CI of LGG data of around 0.78 in its ablation study and improves to 0.814 compared to other baselines.

5.3 Intermediate fusion

Intermediate fusion, alternatively called representation-level or model-level fusion, is a strategy where the multimodal model's loss is fed back to the feature extraction layer of each modality. This process iteratively improves feature representations within the multimodal framework.[26] This method incorporates fusion directly into the model training process and makes decisions on fusion in a way to optimize the objective, as well as to combine individual modalities at different levels of abstraction. This method contrasts with early fusion, as it leverages its benefits by preserving modality-specific structures until the interaction phase; it differs from late fusion by allowing the model to identify cross-modal interactions prior to the prediction phase.[65] Integration can be achieved using methods such as graph neural networks (GNNs), transformers, and attention mechanisms.[31]

Table 12 Comparison of intermediate fusion.

Authors and year	Advantages	Disadvantages
APL[66]	Promotes robustness via class-level prototypes; suitable for imbalanced data; supports interpretable decision making.	Prototype definition can be sensitive to noisy data; may underperform with weak inter-modality correlation.
DIMAF[67]	Introduces an interpretable attention-based fusion framework (DIMAF) that disentangles modality-specific and shared representations for explainability and robustness.	Higher computational cost due to modular design and attention disentanglement; needs high-quality multimodal data to achieve optimal results.
MIF[68]	Introduces a multi-shot interactive fusion to collaborate with various affinity-based interactive modules	Relying solely on a single affinity matrix may not have power to model the intricate modality-specific interactions for cancer prognosis.
PONET[69]	Utilizes adaptive prototype learning to dynamically adjust feature representation from pathology and genomics for survival tasks; improves robustness to modality variation.	Model may underperform if prototype clustering is poorly initialized or the modality quality varies significantly.

This survey will illustrate the use of intermediate fusion through multiple study cases, including PONET, DIMAF, APL, and MIF. Furthermore, Table 12 will present a comparison of their advantages and disadvantages. Moreover, MOTCat, OTGL, DSCRASurv, Pathogen-X, CPAthomic, MoSARe, and CFBCT studies will be discussed, specifically to emphasize their intermediate fusion techniques following the description of their alignment approaches presented earlier in this chapter.

5.3.1 PONET

The authors of[69] introduces PONET, a deep learning model designed to combine pathological images with genomic data - specifically gene expression, copy number variation (CNV) and mutation (MUT) – to predict cancer survival. What makes PONET stand out is its hierarchical intermediate fusion strategy, which includes three levels of fusion: unimodal, bimodal, and trimodal. This allows the model to learn meaningful interactions both within and across different data types. PONET has four key strengths. First, it builds a deep integration framework that uses the knowledge of the biological pathway to guide the learning process. Second, it captures complex relationships between the modalities using a factorized bilinear model, which is more efficient and less prone to overfitting than traditional fusion methods such as Kronecker products. Third, it improves interpretability by linking its predictions to specific genes and pathways, making it useful for discovering potential biomarkers. Lastly, it performs well on several TCGA cancer datasets, outperforming previous models.

In the unimodal fusion step, each data type is processed separately using multimodal Factorized Bilinear pooling (MFB) to reduce complexity and highlight important features. Unlike other methods such as ARGF, PONET uses a weighted sum of features, which helps preserve important information. In bimodal fusion, pairs of modalities are combined, and their contributions are weighted based on attention scores. Finally, in trimodal fusion, all three data sources are merged to capture deeper interactions. PONET effectively demonstrates intermediate fusion in multimodal learning by achieving the following confidence intervals: BLCA (0.643 ± 0.037), KIRC (0.726 ± 0.056), KIRP (0.829 ± 0.054), LUAD (0.646 ± 0.047), LUSC (0.567 ± 0.066), and PAAD (0.639 ± 0.080).

5.3.2 DIMAF

The work in[67] presented an intermediate fusion technique, denoted disentangled and interpretable multimodal attention fusion (DIMAF). It offered a structured and disentangled perspective of multimodal information, improving interpretability and removing the necessity for post-attention feedforward networks (FNNs). Moreover, it evaluated and ensured the disentanglement between latent representations without relying on ground-truth data.

Within this framework, DIMAF uniquely models inter-modal and intra-modal interactions by employing two self-attention layers to encode the modality-specific details of transcriptomics and WSI data. Once the representations are acquired, Layer Normalization (LN) is implemented, followed by an averaging process across pathways or Gaussian Mixture Model (GMM) components for each representation. Through the use of different learnable query, key and value weight matrices tailored to various types and modalities of connection, the model is driven to concentrate on diverse data characteristics, enhancing effective disentanglement.

To further encourage disentanglement among modality-specific representations, distance correlation (DC) has been utilized. DIMAF not only advances the current state-of-the-art performance and disentanglement abilities across four publicly available cancer datasets but also offers a robust basis for enhanced interpretability studies, with the aim of providing a deeper understanding of multimodal cancer biology. Furthermore, employing SHAP both pre- and post-fusion allows for capturing feature importance alongside multimodal interactions. DIMAF provides the CI for BRCA (0.759 ± 0.067), BLCA (0.679 ± 0.043), LUAD (0.669 ± 0.062), and KIRC (0.752 ± 0.092).

5.3.3 APL

Adaptive prototype learning (APL) performs intermediate fusion by turning the high-dimensional features of each modality (e.g., WSI patches and pathway-level transcriptomics) into a small set of learnable prototypes before prediction.[66] For each modality, query vectors attend to the raw tokens to distill compact, task-relevant prototypes that retain salient morphology/pathway signals while suppressing noise and redundancy. The prototype sets are then concatenated and passed through a block of mixed self-attention, allowing dense cross-modal interactions at the representation stage rather than at the decision stage. This prototype-first design balances modalities fewer, comparable tokens, improves efficiency, and yields interpretable artifacts (e.g., prototype-to-patch or prototype-to-pathway attributions). APL shows consistent gains in five cohorts of TCGA (BRCA-CI 0.794, BLCA-CI 0.677, HNSC-CI 0.653, COADREAD-CI 0.812, STAD-CI 0.686). It achieves the best average CI among strong unimodal, multimodal and prototype-based baselines (avg. 0.724).

5.3.4 MIF

The research work in[68] illustrates an enhanced type of intermediate fusion achieved through multiple instances of interactive fusion. Known as Multishot Interactive Fusion (MIF), this model aims to predict cancer survival by combining pathological imagery with genomic datasets. This study utilizes multi-shot fusion to tackle the insufficient examination of intricate high-order interactions among different modalities by efficiently creating and integrating unimodal, bimodal, and trimodal representations. The intricacy of intermediate mechanisms emerges when designing various affinity-based interactive modules to address the capability-efficiency conflict. Specifically, affinity-guided interactive (AGI) modules have been utilized in unimodal contexts, while co-guided interactive (ACGI) modules are poised for deployment in bimodal and trimodal settings. This novel multi-shot framework achieves a mean CI for LGG of 0.853 with a standard deviation of 0.013, and for BRCA, it reaches a mean CI of 0.788 with a standard deviation of 0.023.

5.3.5 MOTCat

Using optimal transport to maintain the global coherence of the potential structure between WSI and genomics, MOTCat adopted an intermediate fusion approach.[3] This fusion took place as the results from the OT-based co-attention were integrated from the transformer of the modalities. These concatenated results are then fed into a loss function to optimize the model to predict survival time. Following alignment, the transformer module integrates bag-level representations from both histopathology and genomics by concatenating the aligned embeddings. During model training, fused representations are used to estimate patient risk scores via a negative log-likelihood (NLL) loss function, facilitating end-to-end optimization for survival prediction. Through this design, MOTCat achieves explicit modality alignment while preserving biological relationships, thereby enabling more coherent and interpretable multimodal integration. This method enables researchers to pinpoint the concordance index (CI) as approximately 0.683, 0.673, and 0.670 for the BLCA, BRCA, and LUAD datasets, respectively, while achieving a CI of 0.849 for GBMLGG.

5.3.6 OTGL

Following alignment of modalities using optimal transport (OT) through intermediate fusion, OTGL extracts the features at the bag-level.[4] These bag-level representations subsequently employ a path transformer to identify long-distance interactions among internal instances of pathology and omics. Once processed through the Transformer layer, they are allocated within a weighting module using a co-attention mechanism. Subsequently, to merge these bags, a two-layer structure consisting of a nonlinear network with an Exponential Linear Unit (ELU) and a Fully Connected (FC) layer is utilized. OTGL incorporates interpretability mechanisms by analyzing intermediate-layer attention weights to assess the contributions of each modality to survival prediction. This interpretability analysis enables the generation of attribution maps, illustrating how genomic, histopathology, and radiological characteristics influence model predictions. By merging class tokens with instance tokens from each modality within the transformer, OTGL constructs a unified embedding space that supports survival prediction from both global and local perspectives by obtaining CI results for GBMLGG 0.849, LUAD 0.631, and BLCA 0.651.

5.3.7 DSCASurv

The fusion process in DSCASurv[12] follows an intermediate fusion approach. Instead of concatenating raw input or combining separate prediction outputs, DSCASurv integrates learned latent representations from each modality in an intermediate layer within the network. These modality-specific embeddings are processed through a cross-modal representation, which ensures a meaningful interaction between histological and molecular features. This attention-guided fusion enables the model to highlight biologically relevant information and mitigate modality imbalance. This mechanism supports the survival prediction task by generating a joint representation using both cross-modal and intra-modal representations with the overall CI results obtained around five datasets (COADREAD, BRCA, BLCA, HNSC, STAD) 0.721.

5.3.8 MGCT

The fusion strategy in MGCT[45] can be interpreted as intermediate fusion, where the characteristics of both modalities are processed independently and subsequently integrated at a mid-level through attention-based interactions. This design balances modality-specific learning with integrative reasoning, allowing the model to retain rich semantic information while leveraging complementary insights for improved survival prediction. Following these guided attention steps, the modality-specific embeddings are fused into a joint representation.

Recognizing the challenge posed by the lack of direct spatial correspondence between gigapixel WSIs and high-dimensional genomic data, MGCT implements a recursive fusion strategy. Specifically, the output of the initial fusion phase is reintroduced as input for subsequent alignment and fusion iterations. This iterative design aims to progressively refine cross-modal interactions and strengthen the joint representation learned by the model to achieve the overall CI results from study cases of BLCA, BRCA, LUAD, GBMLGG and UCEC is 0.663.

5.3.9 PathoGen-X

PathoGen-X introduces a new cross-modal translation and alignment strategy for multimodal survival prediction, utilizing both histopathological imaging and genomic data during training.[6] The fusion mechanism is performed through a three-stage process: (1) a pathology encoder extracts latent features from whole-slide images, (2) a genomic decoder reconstructs genomic features from these pathology-derived embeddings, and (3) a survival prediction module uses translated features to predict risk scores through a Cox loss function. The model is guided by two key losses, they are latent loss and translation loss to ensure accurate alignment and translation of the features of the image into the genomic space. PathoGen-X effectively predicts survival outcomes, even with scarce paired multimodal data. In particular, it shows strong predictive performance with a CI of 0.067 ± 0.020 for BRCA, 0.062 ± 0.008 for LUAD, and 0.81 ± 0.0023 for GBM.

5.3.10 CPathomic

CPathomic presents innovative cross-modal alignment and contrastive learning techniques aimed at improving cancer survival prediction.[13] Fusion is carried out through a cross-modal attention mechanism that emphasizes the most informative cross-modal interactions. This attention-guided fusion allows the model to learn synergistic relationships across modalities while maintaining modality-specific features. Based on this design, where alignment is implicitly enforced and fusion occurs after initial feature encoding and interaction modeling, CPathomic is best categorized as using an intermediate fusion strategy. This computational framework for the management of intricate medical data achieves CI values of BLCA: 0.678 ± 0.025; BRCA: 0.677 ± 0.021; UCEC: 0.677 ± 0.043; GBMLGG: 0.842 ± 0.024; and LUAD: 0.666 ± 0.038.

5.3.11 MoSARe

MoSARe employs an intermediate fusion strategy by integrating WSI patches, the RNA sequence, and the clinical report through a global and local representation-level fusion mechanism.[50] The model separately extracts modality-specific features using two distinct encoders: a Vision Transformer (ViT) for histopathological patches, a multilayer perceptron (MLP) for RNA-seq gene expression data, and a Gaussian mixture model (GMM) to deal with clinical data. These features are projected into a shared representation space and combined using a weighted attention-based fusion module. The fused representation is then passed to a survival prediction head based on the DeepSurv architecture. MoSARe enhances the learning of modality-specific semantics while effectively leveraging cross-modal interactions at the representation level, making it a clear example of intermediate fusion. MoSARe effectively handles missing modalities without relying on strong imputation assumptions. This pipeline provides the Area Under Curve (AUC) measurements for BRCA: 98.53 ± 0.9, RCC: 99.53 ± 0.8, and NSCLC: 98.66 ± 0.8.

5.3.12 CFBCT

CFBCT is an intermediate fusion that aggregates features in co-attention,[52] where (1) fine-grained information leverages the genomic bag guided by the WSI bag feature, (2) coarse-grained information harnesses the genomic bag guided by histology to a specific pattern, and (3) histological-specific patterns guide the genomic bag feature. Subsequently, a transformer encoder for genomics with global attention pooling (GAP) and for histology a similar aggregation function were developed yielding bag-level representation. It means that CFBCT primarily adopts an intermediate fusion through bidirectional co-attention and transformee-based multimodal representation learning, while incorporating an auxiliary decision-level combination for counterfactual debiasing rather than for core fusion. This framework was validated in eight diverse TCGA cancer benchmark datasets to directly address and reduce bias, obtaining the CI results for BLCA 0.697 ± 0.033; BRCA 0.701 ± 0.037; LUAD 0.691 ± 0.033; UCEC 0.753 ± 0.031; LGG 0.800 ± 0.093; COADREAD 0.707 ± 0.055; HNSC 0.623 ± 0.029; and STAD 0.661 ± 0.043.

5.4 Hybrid fusion

This fusion involves combining elements that are in an early, late, or intermediate stage. Using this form of fusion, the limitations of different fusion methods can be mitigated. Its purpose is to enhance effectiveness by balancing the strengths among the various techniques during the fusion process. This approach is suitable when dealing with complex needs in integrating modalities. This survey presents hybrid fusion studies: HEALNET, FORESEE, M2EF-NN, and MultiDeepSurv, highlighting their benefits and drawbacks, as detailed in Table 13. Moreover, the proposed fusion frameworks for CMTA, HySurvPred,CAMR, mSTAR, HGCN, and PGHG are described in addition to their alignment pipelines explained in Section 4.

5.4.1 HEALNet

The authors of[70] introduced the initial phase of a hybrid method known as the Hybrid Early-fusion Attention Learning Network. This approach is crucial for maintaining structural information while facilitating cross-modal

Table 13 Comparison of key advantages and disadvantages of hybrid fusion.

Authors and year	Advantages	Disadvantages
HEALNet[70]	Hierarchical fusion over structured/unstructured data; attention improves interpretability	Multi-stage fusion adds compute; may require large training sets
MultiDeepSurv[71]	Combines multiple early and late fusion schemes with the hybrid network	Performance depends on appropriate tuning and feature balancing across modalities; limited generalizability without diverse datasets.
FORESEE[72]	Combines multi-view, multi-modal inputs; contrastive learning improves robustness and generalisation	Contrastive setup needs careful negative design; may reduce interpretability
M2EFNN[73]	Multi-instance, multimodal fusion; evidential fusion yields calibrated uncertainty	Fusion logic is complex; depends on quality of instance-level evidence aggregation

interaction during multi-modal integration. Early fusion methods integrate raw data at an initial stage, requiring strategies to prevent dimension explosion when handling multiple modalities. Thus, this deep learning technique with multimodal capabilities utilizes an iterative training process to extract essential cross-modal data. In other words, the hybrid fusion here is based on the integration of both early and intermediate fusion methods across two stages. HEALTNet's primary concept involves utilizing both a shared and a modality-specific parameter domain concurrently within a recursive attention framework. A shared latent bottleneck array navigates through the network, undergoing iterative updates that capture common information and facilitate implicit interactions between data modalities. Simultaneously, attention weights are developed for each modality and propagated across layers to acquire modality-specific structural insights. The HEALNet framework achieves CI values for BRCA (0.638 ± 0.073), BLCA (0.668 ± 0.036), KIRP (0.812 ± 0.055), and UCEC (0.626 ± 0.037).

5.4.2 FORESEE

This approach integrates various types of biomedical data, such as WSI and molecular profiles, into a cohesive predictive framework.[72] The fusion process operates on three essential dimensions: scale, modality, and completeness. Initially, in the area of histopathological imaging, FORESEE employs cross-scale fusion via its Cross-Fusion Transformer (CFT). By examining image patches at varying magnification levels – cellular, tissue, and tumor – and combining them using graph neural networks along with transformers, the model merges localized and contextual information from tissues. This fusion grants an improved understanding of tumor heterogeneity. Secondly, in terms of molecular data, the hybrid attention encoder (HAE) achieves multilevel molecular fusion through the integration of both local and global feature attentions. This approach captures detailed gene-level information and overarching molecular patterns while employing wavelet denoising to diminish noise and improve the clarity of molecular signals. Third, FORESEE tackles the issue of missing data using the Triplet Masked Autoencoder (TriMAE), which facilitates fusion through reconstruction. This framework achieves a CI of 0.686 ± 0.008 for BLCA, 0.697 ± 0.013 for BRCA, 0.672 ± 0.013 for LUAD, and 0.730 ± 0.002 for UCEC.

5.4.3 M2EF-NN

A study presented by Luo et al. [73] introduces M2EF-NN, a neural network model that uses multiinstance evidence and integration to predict cancer survival outcomes. This framework comprises three key components: multi-instance multi-modal feature extraction, a network for multi-modal feature integration, and the application of Dempster-Shafer theory (DST) for reliable survival prediction. This approach used a pre-trained ViT model on ImageNet to derive feature embeddings from histology images, followed by two fully connected layers to extract genomic features. The hybrid fusion approach utilizes late fusion by employing genomic embeddings as queries, which results in co-attention mapping between genomic traits and histopathological images to promote early interactions. Following this, global attention pooling gathers multi-instance features into attributes at the image level. Moreover, early fusion occurs when the features from various modalities are integrated via concatenation to derive the final fused feature vector. employing data from cancer types BLCA and GBMLGG, this novel achieved an overall CI score of 0.736.

5.4.4 MultiDeepSurv

The study by Mao et al.[71] introduced MultiDeepSurv, a hybrid multimodal survival prediction framework adapted for gastric cancer, which integrates histopathological images, gene expression profiles and clinical data. A distinctive component of their approach is the proposed SFusion module, a self-attention-based fusion mechanism designed to emphasize cross-modal relationships and enhance predictive precision. The model initially extracts local and global features from histopathology slides using a dual-branch GLFUnet, which combines ConvNeXt and Swin Transformer encoders. In parallel, graph convolutional networks (GCNs) are employed to encode genomic and clinical characteristics, forming a uniform 512-dimensional latent space for each modality. SFusion then facilitates the integration of these modalities in two stages: (1) a Correlation Extraction (CE) module that transforms modality-specific vectors into tokens and applies self-attention to uncover inter-modality interactions and (2) a Modality Attention (MA) module that assigns learned weights to each modality, yielding a fused representation that reflects their relative contributions. Compared to baseline models such as DeepCorrSurv, HFBSurv, and classical Cox regression variants, MultiDeepSurv achieved superior performance with a CI of 0.806 and an AUC of 0.842.

5.4.5 CMTA

CMTA proposed in[1] is classified as a hybrid fusion technique because it couples an early bidirectional cross-modal interaction with a downstream feature-level fusion step. Concretely, it first learns modality-specific embeddings for pathology (p) and genomics (g) via parallel encoders, then uses a cross-modal attention bridge to extract genomics-related cues from pathology and pathology-related cues from genomics; these are translated by decoders into cross-modal representations ($\hat{p}$, $\hat{g}$) that are explicitly aligned to their intra-modal counterparts with a unidirectional $\boldsymbol{L}_1$ constraint (detaching p, g to prevent collapse to shared noise), i.e., early alignment/interaction rather than naÏve stacking. Finally, CMTA performs late fusion by concatenating the recalibrated characteristics $\left(\frac{p+\hat{p}}{2}\right) \oplus \left(\frac{g+\hat{g}}{2}\right)$, where $\oplus$ denotes the concatenation operation, before survival prediction, integrating complementary signals only after the cross-modal calibration has enriched each stream. Presenting this integrated lead CMTA, deriving the CI across five cancer types, namely: BLCA: 0.691 ± 0.0426; BRCA: 0.6679 ± 0.0434; UCEC: 0.6975 ± 0.0409; GBMLGG: 0.8531 ± 0.0116; and LUAD: 0.6864 ± 0.0359.

5.4.6 HySurvPred

HySurvPred couples early intra- and inter-modal structuring/alignment with a downstream joint fusion and risk modeling stage.[48] Concretely, its Multimodal Hyperbolic Mapping (MHM) first organizes each modality (histopathology and genomics) in hyperbolic space $\mathbb{H}^n$, explicitly capturing hierarchical within-modality relations (e.g., tissue → cell clusters → cells; pathways → genes) rather than merely stacking features: a representation step *early* that prepares modalities for interaction and fusion. Its Angle-aware Ranking-based Contrastive Loss (ARCL) then imposes inter-modal geometric and ordinal constraints (risk order preservation) to align the modalities in $\mathbb{H}^n$, effecting *early cross-modal* interaction rather than postponing integration to the classifier. Only after these calibrations are the recalibrated embeddings fused in hyperbolic space for survival prediction, i.e. a *late* integration at the predictor level. Finally, the Censor-Conditioned Uncertainty Constraint (CUC) operates during risk modeling to modulate contributions of censored/uncensored cases using hyperbolic uncertainty (distance to the origin), reinforcing the late-stage integration without undoing the earlier alignment. Using five publicly accessible cancer datasets: BLCA, BRCA, UCEC, LUAD, and GBMLGG, HySurvPred achieves mean CI values of 0.711, 0.757, 0.726, 0.709, and 0.859, respectively.

5.4.7 CAMR

The CAMR framework[8] employs a hybrid fusion strategy to predict cancer survival using multimodal data. CAMR begins with early fusion, aligning data across modalities alongside modality-specific learning. Following this, CAMR utilizes the CMFM (Cross-Modality Fusion Module) for late fusion. This method combines characteristics from histopathological images and gene expression data by independently encoding each modality with specialized neural encoders, followed by employing a cross-modal attention mechanism to facilitate modal interaction. The attention module manages both local and global cross-modal interactions, while the memory retention unit safeguards unique modality representations, maintaining a balance between shared and individual information. As a result, CAMR achieves a CI of 0.780 ± 0.048 for BRCA, 0.841 ± 0.020 for LGG, and 0.650 ± 0.037 for LUSC in three datasets.

5.4.8 mSTAR

mSTAR is a multimodal foundation model in computational pathology that applies an innovative hybrid fusion approach to synergize WSI, pathological reports, and gene expression data.[43] The model employs a dual-stage pretraining methodology: First, it transfers multimodal insights to a patch-level encoder using self-supervised learning (early fusion); then, it engages in slide-level contrastive learning to align comprehensive representations across different modalities (late fusion). By integrating early and late fusion advantages, mSTAR adeptly captures intricate local features and overarching clinical insights, facilitating strong downstream task performance in areas like diagnosis, prognosis, survival prediction, and report creation. This fusion strategy improves the contextual analysis of pathological information, setting a new standard for multimodal foundation models in digital pathology, with CI results: BRCA: 0.7076, UCEC: 0.6975 ± 0.8092, GBMLGG: 0.7923, LUAD: 0.6864 ± 0.6329, CRC: 0.6895, LUSC: 0.6323, and KIRC: 0.7027.

5.4.9 HGCN

HGCN is a hybrid fusion technique that enables interactions within-modality and between-modalities in multimodal graphs.[11] Within-modal communication was performed through three identical graph-convolutional network (GCN) layers applied to pathological slides, clinical records, and genomic profiles. Furthermore, the hyperedge mix employed two MLP layers to promote extensive cross-modality interaction. For fusion of information between modalities, the HGCN integrates a Hypergraph Convolutional Network (HCN) to establish higher-order associations among features using hyperedge mixing. To address missing-modality issues, HGCN employs an on-line masked auto-encoder (MAE) that operates alongside model training. This approach effectively uncovers intrinsic relationships between modalities, achieving a CI of 0.747 ± 0.007 for KIRC, 0.693 ± 0.010 for LIHC, 0.634 ± 0.015 for ESCA, 0.598 ± 0.012 for LUSC, 0.651 ± 0.008 for LUAD, and 0.747 ± 0.017 for UCEC.

5.4.10 PGHG

PGHG is classified as a hybrid fusion because it combines (i) early, knowledge-driven cross-modal alignment, (ii) intermediate interaction in the course of representation learning, and (iii) late predictive integration.[55] First, the module guided by biological knowledge shapes each stream before any final merger: WSIs and pathways are regularized through pathway/graph reconstructions (L_{RNA}, L_{Adj}), pathology is supervised by the activity of the GSVA pathway, and global embeddings are explicitly aligned across modalities (L_{align})—all of which constitute early cross modal calibration rather than previous stacking. Second, PGHG constructs a pathology genome heterogeneous graph and applies attention-based message passing that aggregates both intramodal and inter-modal neighbors, so representation learning itself occurs under multimodal context (classic intermediate fusion). Third, only after these calibrations are unimodal and global descriptors of modalities (f_{pp}, f_{pg}, f_{gp}, f_{gg}) gated and concatenated for survival risk modeling (discrete-time NLL), which is a deliberate late integration step. With these techniques, PGHG achieved a CI of 0.823 ± 0.026 for TCGA LGG and 0.773 ± 0.109 for GBM. Furthermore, using data from the hospital in Zhengzhou associated with FAHZU, the CI was 0.685 ± 0.011 for LGG and 0.600 ± 0.032 for GBM.

6. Discussions

This section integrates the findings of the previous sections, shifting from a descriptive comparison of models to a conceptual examination of how alignment and fusion jointly influence multimodal cancer survival prediction. Although earlier sections considered alignment mechanisms and fusion paradigms separately, the discussion here reframes them as mutually dependent design choices whose interaction determines robustness, interpretability, and generalization in clinical applications. Furthermore, many recent methods struggle to fuse without an alignment process because they cannot be replaced and remain interdependent.

Building on the empirical results summarized in Tables 14 and 15 and the conceptual overview in Fig. 4, the discussion is organized into three complementary points of view. First, Section 6.1 approaches alignment and fusion from a co-design perspective, arguing that the success of multimodal integration hinges on when and how modality correspondence is enforced during fusion. Second, Section 6.2 examines current alignment practices beyond performance scores, underscoring the limitations of the concordance index (CI) as a representative of alignment quality and highlighting the need for explicitly aligned evaluation protocols. Third, Section 6.3 re-examines fusion paradigms in a comparative manner, showing that differences in reported performance frequently stem from disease-specific factors and cohort characteristics rather than from architectural choices alone.

In general, these perspectives underpin the main conclusion of this chapter, which is stated as follows: Advancement of AI for multimodal cancer survival analysis requires redirecting focus from isolated architectural

Table 14 Summary of multimodal survival prediction models by fusion strategy with alignment.

Method	Data (modalities)	Cancer types and Cohort size	Align type	Alignment method	Metric evaluation (CI mean ± std)	Avg. CI	References	Fusion type
pCCA	WSI, RNA-seq	BRCA (n = 974)	**Explicit**	Canonical correlation analysis	OPD* accuracy 94.32 ± 1.08	—	Subramanian et al. (2021)	Early
OTGL	WSI, MRI and CT images, RNA-seq	PandR: GBMLGG (n = 167), GC (n = 121); PandG: GBMLGG (n = 573), BLCA (n = 373), LUAD (n = 453)	**Explicit**	Optimal Transport	PandR: GBMLGG 0.72 ± 0.042, GC 0.601 ± 0.056; PandG: BLCA 0.651 ± 0.027, GBMLGG 0.849 ± 0.016, LUAD 0.631 ± 0.013	0.7103	Sun et al. (2025)	Intermediate
MOTCat	WSI, RNA-seq, CNV	BLCA (n = 373), BRCA (n = 956), UCEC (n = 480), GBMLGG (n = 569), LUAD (n = 453)	**Explicit**	Optimal Transport	BLCA 0.683 ± 0.026, BRCA 0.673 ± 0.006, UCEC 0.675 ± 0.040, GBMLGG 0.849 ± 0.028, LUAD 0.670 ± 0.038	0.71	Xu and Chen (2023)	Intermediate

(continued)

Table 14 Summary of multimodal survival prediction models by fusion strategy with alignment. (*cont'd*)

Method	Data (modalities)	Cancer types and Cohort size	Align type	Alignment method	Metric evaluation (CI mean ± std)	Avg. CI	References	Fusion type
MIRROR	WSI, transcriptomics	NSCLC (n = 1053), BRCA (n = 955), RCC (n = 943), COADREAD (n = 623)	**Explicit**	Self-supervised representation (contrastive)	NSCLC 0.613 ± 0.043, BRCA 0.665 ± 0.082, RCC 0.803 ± 0.043, COADREAD 0.721 ± 0.033	0.7005	Wang et al. (2025)	Early
MGCT	WSI, RNA-seq, CNV	TCGA BRCA (n = 437), BLCA (n = 1023), UCEC (n = 539), GBMLGG (n = 1024), LUAD (n = 516)	**Implicit**	Neural network (cross-modality attention)	BLCA 0.640 ± 0.039, BRCA 0.608 ± 0.026, UCEC 0.645 ± 0.039, GBMLGG 0.827 ± 0.024, LUAD 0.596 ± 0.078	0.71	Liu et al. (2023)	Intermediate
DSCASurv	WSI, gene expression	BRCA (n = 869), STAD (n = 317), BLCA (n = 359), HNSC (n = 392), COADREAD (n = 296)	**Implicit**	Neural network (dual-stream cross-attention)	BLCA 0.646 ± 0.034, BRCA 0.765 ± 0.061, COADREAD 0.832 ± 0.095, HNSC 0.666 ± 0.071, STAD 0.698 ± 0.093	0.721	Song et al. (2025)	Intermediate

CMTA	WSI, RNA, CNV, SNV	BLCA (n = 373), BRCA (n = 956), UCEC (n = 480), GBMLGG (n = 569), LUAD (n = 453)	**Implicit**	Cross-modal translation	BLCA 0.691 ± 0.0426, BRCA 0.6679 ± 0.0434, UCEC 0.6975 ± 0.0409, GBMLGG 0.8531 ± 0.0116, LUAD 0.6864 ± 0.0359	0.719	Zhou and Chen (2023)	Hybrid
PathoGen–X	WSI, RNA-seq	BRCA (n = 987), GBM (n = 576), LUAD (n = 509)	**Implicit**	Cross-modal translation	BRCA 0.67 ± 0.020, GBM 0.81 ± 0.0023, LUAD 0.62 ± 0.008	0.70	Krishna et al. (2024)	Intermediate
CAMR	WSI, gene expression, CNA	BRCA (n = 987), LGG (n = 576), LUSC (n = 509)	**Implicit**	Adversarial cross-modal translation	BRCA 0.780 ± 0.048, LGG 0.841 ± 0.020, LUSC 0.650 ± 0.037	0.757	Wu et al. (2023)	Intermediate
HySurvPred	WSI, gene expression	BLCA (n = 373), BRCA (n = 956), UCEC (n = 480), GBMLGG (n = 987), LUAD (n = 987)	**Implicit**	Contrastive learning (hyperbolic map)	BLCA 0.711, BRCA 0.757, UCEC 0.726, GBMLGG 0.859, LUAD 0.709	0.752	Yang et al. (2025)	Hybrid

(continued)

Table 14 Summary of multimodal survival prediction models by fusion strategy with alignment. (*cont'd*)

Method	Data (modalities)	Cancer types and Cohort size	Align type	Alignment method	Metric evaluation (CI mean ± std)	Avg. CI	References	Fusion type
CPathomic	WSI, genomics	BLCA (n = 373), BRCA (n = 957), UCEC (n = 480), GBMLGG (n = 573), LUAD (n = 453)	**Implicit**	Contrastive learning (cross-modal contrast + attention)	BLCA 0.678 ± 0.025, BRCA 0.677 ± 0.021, UCEC 0.677 ± 0.043, GBMLGG 0.842 ± 0.024, LUAD 0.666 ± 0.038	0.708	Li et al. (2025)	Intermediate
MoSaRe	WSI, RNA, clinical text	BRCA (n = 955), RCC (n = 943), NSCLC (n = 1053)	**Implicit**	Contrastive learning	AUC**: BRCA 98.53 ± 0.9, RCC 99.53 ± 0.8, NSCLC 98.66 ± 0.8	98.90	Moradinasab et al. (2025)	Intermediate
HGCN	WSI, clinical, genomic	KIRC (n = 385), LIHC (n = 287), ESCA (n = 153), LUSC (n = 438), LUAD (n = 452), UCEC (n = 387)	**Implicit**	Hypergraph + masked autoencoding	KIRC 0.747 ± 0.007, LIHC 0.693 ± 0.010, ESCA 0.634 ± 0.015, LUSC 0.598 ± 0.012, LUAD 0.651 ± 0.008, UCEC 0.747 ± 0.017	0.6611	Hou et al. (2023)	Hybrid

CFBCT	WSI, gene expression, CNV	BLCA (n = 373), BRCA (n = 947), LUAD (n = 445), UCEC (n = 478), LGG (n = 390), COADREAD (n = 294), HNSC (n = 327), STAD (n = 392)	**Implicit**	Graph-based (counterfactual bidirectional co-attention)	BLCA 0.697 ± 0.033, BRCA 0.701 ± 0.037, LUAD 0.691 ± 0.033, UCEC 0.753 ± 0.031, LGG 0.800 ± 0.093, COADREAD 0.707 ± 0.055, HNSC 0.623 ± 0.029, STAD 0.661 ± 0.043	0.704	Ji et al. (2025)	Intermediate
MMP	WSI, transcriptomics	BLCA (n = 359), BRCA (n = 868), LUAD (n = 412), STAD (n = 318), CRC (n = 296), KIRC (n = 340)	**Hybrid**	OT + transformer prototypes (token reduction)	BLCA 0.753 ± 0.069, BRCA 0.635 ± 0.051, LUAD 0.643 ± 0.013, STAD 0.598 ± 0.051, CRC 0.636 ± 0.120, KIRC 0.748 ± 0.099	0.668	Song et al. (2024)	Early

(continued)

Table 14 Summary of multimodal survival prediction models by fusion strategy with alignment. (*cont'd*)

Method	Data (modalities)	Cancer types and Cohort size	Align type	Alignment method	Metric evaluation (CI mean ± std)	Avg. CI	References	Fusion type
PGHG	WSI, RNA-seq	LGG (n = 466), GBM (n = 256)	**Hybrid**	Cross-modal + graph learning (GSVA pathways)	LGG 0.823 ± 0.026, GBM 0.773 ± 0.109; External (FAHZU): LGG 0.685 ± 0.011, GBM 0.600 ± 0.032	0.798	Zhang et al. (2024)	Intermediate
mSTAR	WSI, gene expression, clinical reports	BRCA (n = 1023), UCEC (n = 495), GBMLGG (n = 830), LUAD (n = 455), CRC (n = 579), LUSC (n = 452), KIRC (n = 498), HNSC (n = 441), SKCM (n = 415)	**Hybrid**	Encoder + contrastive learning (dual-stage)	BRCA 0.7076 ± 0.001, UCEC 0.6975 ± 0.01, GBMLGG 0.7923 ± 0.05, LUAD 0.6864 ± 0.6329, CRC 0.6895 ± 0.05, LUSC 0.6323 ± 0.001, KIRC 0.7027 ± 0.005, HNSC 0.6604 ± 0.001, SKCM 0.6281 ± 0.05	0.701	Xu et al. (2025)	Hybrid

Abbreviations: OPD = Orthogonalized Projected Deflation; UOT = Unbalanced Optimal Transport; WSI = whole-slide image; CNV = copy-number variation; SNV = single-nucleotide variant; P = Pathology; G = Genomics; R = Radiology; AUC = Area Under ROC Curve.

Table 15 Summary of multimodal survival prediction models by fusion strategy without alignment.

Information	Early fusion			Intermediate fusion				Late fusion		Hybrid Fusion			
	MMDR	**Pathomic fusion**	**MMSurv**	**PONET**	**DIMAF**	**APL**	**MIF**	**MSEN**	**MIFAPS**	**HEALNet**	**FORESEE**	**M2EFNN**	**MultiDeep-Surv**
CI (mean)	**0.838**	**0.7730**	**0.7283**	**0.675**	**0.715**	**0.724**	**0.8205**	**0.6858**	**(AUC) 0.882**	**0.686**	**0.6963**	**0.736**	**0.806**
CI per cancer type	–	GBMLGG (0.826) KIRC (0.720)	BRCA (0.7643) COAD (0.782) ESCA (0.7803) LIHC (0.6864) LUAD (0.6927) STAD (0.6641)	BLCA (0.643 ± 0.037) KIRC (0.726 ± 0.056) KIRP (0.829 ± 0.054) LUAD (0.646 ± 0.047) LUSC (0.567 ± 0.066) PAAD (0.639 ± 0.080)	BRCA (0.759 ± 0.067) BLCA (0.679 ± 0.043) LUAD (0.669 ± 0.062) KIRC (0.752 ± 0.092)	BRCA (0.794 ± 0.062) BLCA (0.677 ± 0.060) COAD-READ (0.812 ± 0.115) HNSC (0.653 ± 0.045) STAD (0.686 ± 0.053)	LGG (0.853 ± 0.013) BRCA (0.794 ± 0.062)	LGG (0.853 ± 0.013) BRCA (0.788 ± 0.023)	BLCA (0.662 ± 0.042) BRCA (0.648 ± 0.045) UCEC (0.661 ± 0.063) GBML-GG (0.827 ± 0.018) LUAD (0.631 ± 0.080)	BRCA (0.638 ± 0.073) BLCA (0.666 ± 0.036) KIRP (0.812 ± 0.055) UCEC (0.626 ± 0.037)	BLCA (0.686 ± 0.008) BRCA (0.697 ± 0.013) LUAD (0.672 ± 0.013) UCEC (0.730 ± 0.002)	BLCA (0.651 ± 0.021) GBML-GG (0.821 ± 0.034)	–
Std of CI mean	0.0022	0.0185	–	0.056	0.066	0.067	0.018	0.0496	–	0.0502	0.009	0.0275	0.008

(continued)

Table 15 Summary of multimodal survival prediction models by fusion strategy without alignment. (*cont'd*)

Information	Early fusion			Intermediate fusion			Late fusion			Hybrid Fusion			
	MMDR	**Pathomic fusion**	**MMSurv**	**PONET**	**DIMAF**	**APL**	**MIF**	**MSEN**	**MIFAPS**	**HEALNet**	**FORESEE**	**M2EFNN**	**MultiDeep-Surv**
CI (mean)	**0.838**	**0.7730**	**0.7283**	**0.675**	**0.715**	**0.724**	**0.8205**	**0.6858**	**(AUC) 0.882**	**0.686**	**0.6963**	**0.736**	**0.806**
Interpretability	No	Yes (Grad-CAM and Heatmap)	Yes (Hover-Net)	Yes (Integrated Gradients attribution)	Yes (SHAP)	Yes (Cross-Attention Maps and biological pathways)	No	No	Yes (Grad-CAM)	No	No	Yes (Attention-based genomic heatmaps with patches)	Yes (Attention mechanisms and graph-based modeling)
Log-rank p-value	5.88E-45	6.95E-13	<0.001	6.60e-7	–	–	6.21E-31	5.70E-42	<0.05	–	BLCA 2.07e-4 BRCA 5.76e-9 LUAD 9.99e-3 UCEC 1.96e-5	BLCA 2.13e-6 GBML-GG 3.59e-31	0.00024

-Cohort size	LGG 629	GBMLGG 769 CCRCC 417	BRCA 801 COAD 487 ESCA 110 LIHC 331 LUAD 410 STAD 381	BLCA 437 KIRC 350 KIRP 284 LUAD 515 LUSC 484 PAAD 180	BRCA 868 BLCA 359 LUAD 412 KIRC 340	BRCA 869 BLCA 359 COAD-READ 296 HNSC 392 STAD 317	LGG 629 BRCA 1015	BLCA 373 BRCA 897 UCEC 478 GBMLGG 492 LUAD 430	BRCA 1004	BRCA 1021 BLCA 436 KIRP 285 UCEC 538	2479	BLCA 373 GBMLGG 567	GC 443
Dataset Source	TCGA	TCGA/ cBioPortal	TCGA	TCGA	XENA DB	TCGA	TCGA	TCGA	Multiple Medical Centre	TCGA	TCGA	TCGA	TCGA
Modalities	Gene expr. CNA Pathology	mRNA CNV Mutation WSI	RNA-seq H&E WSI	Gene expr. CNV Mutation Pathology	Transcriptomic RNA-seq WSI	Gene expr. WSI	Gene expr. CNA Pathology	Bulk RNA-seq CNV Mutation WSI	Genomic & Transcriptomic WSI + MRI	RNA-seq CNV Mutation WSI	Genomic profile WSI	mRNA expr. Pathology	Pathological image & Gene espression Clinical data

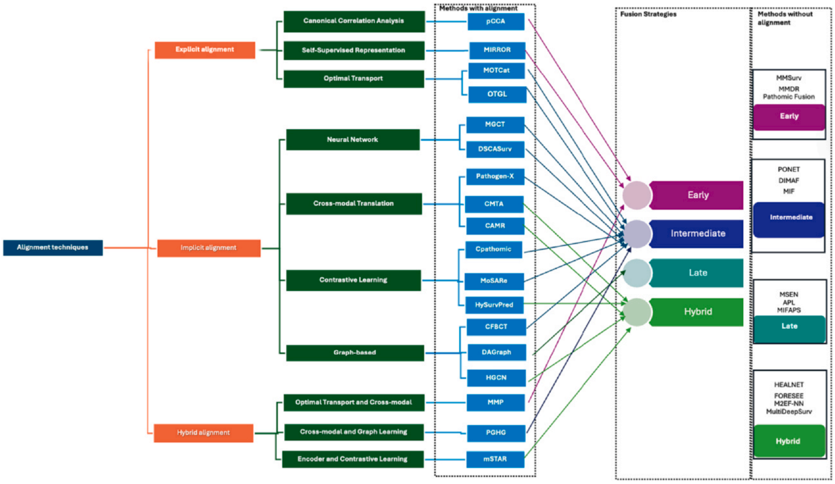

Fig. 4 Alignment and fusion strategies in multimodal survival prediction. The diagram organizes models by alignment paradigm (explicit, implicit, hybrid), delineates the underlying alignment techniques, and links these choices to the corresponding fusion strategies (early, intermediate, late, hybrid), while distinguishing approaches that incorporate a dedicated alignment stage from those that fuse modalities without prior alignment.

innovations toward systematic design, evaluation, and interpretation of alignment-fusion interactions. By articulating common insights across alignment methods and fusion strategies, this discussion establishes a coherent link between the surveyed literature and the broader challenges and future directions developed in the following sections.

6.1 Co-design principles for alignment and fusion in multimodal survival models

One key message of this survey is that alignment and fusion should not be viewed as separate design choices in multimodal survival modeling. Alignment specifies how heterogeneous modalities are mapped into mutual correspondence, whereas fusion specifies where and in what manner these aligned representations are combined to make predictions. As shown in Fig. 4, explicit, implicit, and hybrid alignment strategies are systematically linked to particular fusion stages, highlighting that these two aspects are fundamentally interdependent rather than freely interchangeable.

Early and intermediate fusion approaches typically depend on explicit alignment procedures to define cross-modal correspondences before combining features. This design is largely driven by the need for a biological perspective, where clear and traceable links between modalities are crucial, such as molecular pathways and histopathological features. By contrast, implicit alignment is more commonly embedded in intermediate and hybrid fusion schemes, in which alignment and fusion are learned jointly within shared representations. These approaches are frequently implemented using cross modal translation, contrastive learning, or graph-based models with the objectives that operate in common latent spaces and capture structural relationships across modalities. Hybrid alignment methods, such as MMP, PGHG, and mSTAR, are generally paired with early, intermediate, or hybrid fusion, allowing complementary information to be merged once a certain level of cross-modal coordination has been achieved.

Importantly, Fig. 4 serves not only as a taxonomy but also as a design blueprint that shapes how alignment strategies should influence fusion choices. When the goal emphasizes the mechanism of showing the representation to be interpreted, for example, uncovering biologically meaningful relationships between genomic alterations and morphological phenotypes, explicit alignment in combination with early fusion is typically the most justifiable option. This setup enables straightforward examination of cross-modal relationships, but it also requires rigorous calibration and sensitivity analyzes, since incorrectly specified alignments can introduce errors that propagate into subsequent survival modeling. In contrast, when scalability and robustness to noisy or weakly matched data are the main priorities, implicit alignment incorporated into intermediate or hybrid fusion schemes becomes more suitable. Under these conditions, alignment arises through shared representation learning, supporting flexible accommodation of heterogeneous data, although this comes with an elevated risk of overfitting if regularization is insufficient.

In more heterogeneous clinical settings, where the strength and nature of modality coupling vary between subgroups of patients or cancer subtypes, hybrid alignment strategies combined with early or intermediate fusion offer a pragmatic compromise. By integrating structured correspondence with adaptive learning, these models can capture both stable biological signals and cohort-specific variations. However, this added flexibility introduces architectural complexity that must be rigorously justified. Each alignment and fusion component should demonstrate its necessity through ablation analyzes and external validation, ensuring that complexity contributes to genuine predictive and interpretive gains rather than superficial performance improvements.

Overall, this analysis underscored that successful multimodal survival modeling is not driven by the adoption of increasingly sophisticated architectural components, but by principled co-design of alignment and fusion. The critical question is not which alignment or fusion method is most advanced, but whether their combination yields measurable improvements over alignment-free or fusion-only baselines. This perspective reinforces the larger conclusion of the chapter that robust and clinically meaningful survival prediction depends on coherent design choices that balance the need for the nature of the data cancer itself with aspect interpretability, flexibility, and generalization.

6.2 Evaluating alignment beyond performance metrics

Table 14 presents a general summary of the multimodal survival models surveyed that directly incorporate alignment into their fusion pipelines. A key takeaway from this comparison is that alignment quality cannot be fully evaluated using performance metrics alone. Although the reported concordance indices (CIs) are informative, they provide only a limited perspective on how alignment strategies affect robustness, generalization, and interpretability in survival prediction.

Explicit alignment strategies make this limitation apparent. Approaches like pCCA provide highly interpretable projections, as evidenced by strong correspondence accuracy, but their dependence on linear models constrains both scalability and representational power. Optimal transport-based frameworks, such as MOTCat and OTGL, enforce structural consistency between whole slide images and omics data, resulting in modest gains in CI. However, their susceptibility to choices in cost functions, entropic regularization, and token resolution reveals a fragility that CI alone cannot capture. When transport plans are mis-specified, they can cause over- or under-alignment, introducing geometric distortions that carry over into subsequent fusion and survival prediction.

Implicit alignment techniques, as an indirect method of alignment with intermediate step for tasks, further highlight the discrepancy between performance metrics and the robustness of alignment. From performance measurement, the use of attention-based models and cross-modal translation frameworks shows that learned correspondences can strengthen interactions between modalities, as reflected in their strong CI scores. Meanwhile, the robustness considerations that emerge from several aspect for example tightly coupled to sampling procedures, typically require paired modalities at inference time, and are still susceptible to domain shifts that disproportionately degrade histopathology-derived features.

Hybrid alignment frameworks integrate explicit constraints with implicit learning mechanisms and often report state-of-the-art concordance indices. Methods like PGHG and mSTAR exemplify this pattern: they achieve high performance on their own training cohorts but experience marked performance declines when tested on external datasets (for example, PGHG evaluated on FAHZU data, and mSTAR assessed on five distinct, non-public datasets from different hospitals). This suggests that elevated CI scores may, in part, reflect alignment specific to the original cohort rather than truly robust and generalizable mappings. Similarly, prototype-based reductions, as used in MMP, enhance computational efficiency but may mask infrequent, yet prognostically important signals. Together, these points highlight that alignment quality must be assessed not only by predictive accuracy but also by its stability, transferability, and consistency with biological knowledge.

Collectively, these results motivate evaluation practices that explicitly account for alignment. First, no single alignment framework performs best across all datasets or cancer types, suggesting that many reported performance improvements are attributable to disease-specific morphology-molecular relationships rather than to inherently superior architectures. Second, existing benchmarks largely equate alignment quality with CI, an assumption that is not strongly justified, especially in the presence of censored survival data. Third, alignment is better understood as a coherent, consistent interface constrained by biological priors across modalities, in which semantic relationship estimates are propagated through fusion and risk prediction rather than being ignored or collapsed.

Thus, future evaluation protocols should extend beyond reporting performance in isolation. It is necessary to conduct controlled comparisons of explicit, implicit, and hybrid alignment strategies in the same patient cohorts, complemented by ablation studies that measure the added value of alignment over alignment-free baselines. Furthermore, robustness assessments under missing modalities, center-specific staining differences, and broader domain changes should become standard practice. As illustrated by the benchmarking in Table 14, alignment should not be treated as a static preprocessing step, but rather as a flexible design choice that influences the entire survival pipeline of fusion and prediction. Establishing standardized criteria and procedures for evaluating alignment is therefore crucial for the development of clinically reliable and interpretable multimodal survival models.

6.3 Comparative insights into fusion paradigms for survival prediction

Table 15 provides an overview of various multimodal survival models, grouped by fusion strategy (early, intermediate, late, and hybrid), all under the common assumption that no alignment mechanism is used. These results should be read with caution, as the underlying studies differ markedly in cohort composition, sample sizes, and evaluation protocols, for example, some models are tested in a single TCGA cohort such as GBMLGG, while others are applied in multiple types of cancer. Moreover, the reported performance metrics are not fully consistent, with MIFAPS, for example, reporting AUC instead of CI. Consequently, the table is more appropriate for highlighting broad patterns and design trade-offs among fusion approaches than for drawing strict, quantitative performance comparisons between specific models.

Early fusion approaches (e.g., MMDR, Pathomic Fusion, MMSurv) indicate that feature-level integration can perform competitively when the pipeline is carefully standardized using one, two, or six cancer data types. Within this category, MMDR has the highest mean CI (0.838). These results indicate that, although measurement and modality are shared, using a single cancer type yields a high CI. However, when cancer-specific metrics are reported, performance varies between cohorts (e.g. Pathomic Fusion achieves a higher CI in GBMLGG than in KIRC). This is similar to the notion that architectures that rely heavily on feature concatenation tend to accentuate cohort-specific statistical patterns—advantageous when the cohort is relatively homogeneous but less reliable in the presence of pronounced disease heterogeneity and modality-specific noise. Meanwhile, MMSurv indicated that using not only compact bilinear pooling and a transformer but also word-embedding techniques on clinical data effectively fused diverse cancer types.

Intermediate fusion seeks to model cross-modal interactions via shared latent spaces instead of simple feature concatenation. As shown in the table, methods in this category can achieve strong performance, but not consistently. For example, DIMAF achieves a moderate mean CI (0.715) with substantial variability (standard deviation of the mean CI, 0.066), indicating sensitivity to the specific dataset and/or training configuration. PONET's average CI is lower (0.675) using six types of cancer types. These results underscore that intermediate fusion is not intrinsically more "robust" than early fusion; it can still break down when the shared representation is poorly constrained by limited data, class imbalance, or missing modalities in several types of cancer. By contrast, MIF is a clear positive example in this group, with a high mean CI (0.8205) and strong performance in LGG and BRCA.

Late fusion methods prioritize modularity and robustness to missing data modalities, traits that are especially important in real-world clinical environments. Ensemble-style techniques such as MIFAPS and MSEN deliver strong aggregate performance, benefiting from the robustness that arises when modality-specific predictors are decoupled. Yet this modularity has trade-offs. In the absence of explicit cross-modal alignment mechanisms, late fusion can miss complementary interactions between modalities, which may manifest as marked performance discrepancies across cancer types. The contrast between results on GBMLGG and LUAD cohorts illustrates that late fusion outcomes are still heavily influenced by disease-specific characteristics, rather than by model architecture alone.

Hybrid fusion approaches combine modalities at several stages in the model, aiming to trade off between expressivity (capturing cross-modal interactions) and robustness (coping with variability across cohorts). In the table, MultiDeepSurv achieves a high mean CI (0.806) using only one type of cancer data, while M2EFNN achieves a mean CI of 0.736 and exhibits pronounced cohort-specific behavior in GBMLGG and BLCA. However, the results shown do not justify a general statement that hybrid fusion is consistently effective across various types of cancer for all hybrid architectures, since not every hybrid method reports performance by cancer type and the sets of cohorts differ. A more nuanced conclusion is that hybrid architectures can deliver competitive results and, in some cases, enhanced stability, but this benefit is coupled with increased model complexity and a higher risk of confounding factors (e.g., more hyperparameters, multiple fusion stages, and stronger dependence on the training protocol).

Taken together, these comparisons support another key conclusion of this chapter: the choice of fusion strategy should be governed by the properties of the data and the practical clinical setting, rather than by superficial performance rankings. The CI alone is inadequate for substantive comparison, as variability between cancer types and cohorts often exceeds performance differences attributable to the fusion approach itself. Although statistical significance, for example via log-rank tests, demonstrates prognostic separation, it does not capture generalizability, explainability, or stability. Consequently, future benchmarking should employ disease-aware evaluation schemes that report calibration, time-varying performance, and robustness to missingness and domain shifts in addition to standard metrics. In the absence of such standardized protocols, higher concordance scores may be over-interpreted as universally transferable, masking the subtle trade-offs that characterize multimodal survival modeling.

7. Challenges and future works

7.1 Challenges

Challenge 1 – Alignment-fusion co-design remains under-specified.

As discussed in Section 6.1, alignment and fusion constitute a coupled design decision rather than two independent modeling steps. However, much of the existing literature still treats alignment as an auxiliary module appended to an otherwise fixed fusion strategy. Explicit alignment methods are often applied without assessing whether the chosen fusion stage can effectively exploit the induced correspondence structure, whereas implicit alignment mechanisms are frequently paired with aggressive fusion that may dilute biologically meaningful relationships. This lack of principled co-design complicates both interpretability and robustness, particularly in heterogeneous clinical settings where modality coupling varies across patient subgroups.

Challenge 2 - Alignment quality is rarely evaluated beyond prediction performance metrics.

As highlighted in Section 6.2, most studies continue to rely on CI as the primary criterion for model comparison. Although CI captures ranking performance, it does not reflect alignment fidelity, stability, or biological plausibility. Models achieving high CI may still rely on fragile or cohort-specific alignments that fail under distribution shifts. The absence of standardized metrics for assessing alignment uncertainty and robustness poses a significant barrier to fair comparison and limits the interpretability of reported gains.

Challenge 3 - Generalization under real-world clinical conditions.

The comparative analysis in Section 6.3 demonstrates that fusion performance is strongly influenced by cancer type, cohort composition, and modality availability. Many multimodal models experience substantial performance degradation when evaluated in external cohorts or under missing-view conditions. Overly rigid alignment can lead to error propagation, while insufficient alignment may prevent meaningful cross-modal interaction. Obtaining robust generalization across institutions, staining protocols, and molecular distributions remains an open challenge.

Challenge 4 – Interpretability beyond post-hoc analysis.

High predictive performance does not necessarily translate into clinical utility. Late-stage fusion may obscure cross-modal reasoning, while implicit alignment risks exploiting spurious correlations. In many cases, interpretability is introduced post hoc rather than being embedded into the alignment and fusion design. This limits the ability to trace prognostic decisions back to biologically meaningful entities such as pathways, cellular regions, or morphological patterns.

7.2 Future works

7.2.1 Alignment-aware benchmarking frameworks

Future studies should adopt benchmarking protocols that explicitly disentangle the contribution of alignment from that of fusion. Retraining early, intermediate, late, and hybrid fusion models on identical cohort splits would enable a more reliable comparison and clarify whether performance gains arise from architectural design or dataset-specific effects.

7.2.2 Probabilistic and biologically constrained alignment

Alignment should be reframed as a probabilistic interface between modalities rather than a deterministic mapping. Incorporating biological priors, such as gene pathways, spatial tissue organization, or cellular hierarchies, can constrain alignment to plausible interactions while allowing uncertainty to propagate into downstream fusion and survival estimation. Approaches such as unbalanced optimal transport, uncertainty-aware contrastive learning, and causally informed alignment offer promising directions.

7.2.3 Adaptive fusion strategies

Rather than committing to a fixed fusion stage, future models may benefit from adaptive fusion policies that select early, intermediate, or late integration based on alignment confidence, modality completeness, or clinical context. Such strategies directly operationalize the co-design principles discussed in Section 6 and offer a pathway toward more flexible and reliable multimodal systems.

7.2.4 From performance optimization to clinical reliability

Beyond improving CI, future research should prioritize reliability, interpretability, and failure analysis. Reporting calibration, time-dependent performance, robustness to missing data, and clearly documented failure cases will be essential for translating multimodal survival models into clinical practice.

8. Conclusions

This chapter has explored how AI contributes to the multimodal analysis of cancer survival, emphasizing the ways in which alignment and fusion strategies together influence model performance, interpretability, and robustness. Through a systematic review of work published between 2015 and 2025, we argued that alignment should not be viewed as a minor preprocessing step, but as a core design element that governs how diverse data modalities interact within survival prediction models.

As discussed in Section 6, no single alignment or fusion strategy emerges as universally superior. Explicit alignment yields clear cross-modal correspondences, but depends heavily on modeling assumptions and can therefore be hard to satisfy. Implicit alignment is more flexible, yet it risks encoding unstable or biologically questionable relationships. Hybrid strategies are promising, but their additional complexity requires careful management of uncertainty and strict evaluation. Overall, differences in performance across fusion strategies are more strongly influenced by disease-specific factors and cohort composition than by model architectural complexity alone.

The primary contribution of this chapter is to propose an integrated viewpoint that connects alignment, fusion, and evaluation under a unified framework. By highlighting co-design principles of alignment and fusion, alignment-based evaluation practices, and disease-specific benchmarking, this work aims to inform the development of multimodal survival models that are accurate, interpretable, robust, and clinically relevant. In the long term, progress in AI for cancer survival analysis depends on moving beyond incremental performance gains to the reliable integration of heterogeneous biomedical data sources.

Declaration

The authors declare that the material in this chapter, including the scientific content, analyzes, and interpretations & conclusions, was entirely developed by the authors without the use of generative AI tools. Editorial tools, Overleaf and Grammarly, were used solely for scientific writing, managing references, and polishing the language.

References

1. Zhou F, Chen H. *Cross-modal translation and alignment for survival analysis. 2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2023; 2023:21428–21437.
2. Abbasi AF, Asim MN, Ahmed S, Vollmer S, Dengel A. Survival prediction landscape: an in-depth systematic literature review on activities, methods, tools, diseases, and databases. *Front. Artif. Intell.* 2024;7.

3. Xu Y, Chen H. Multimodal optimal transport-based co-attention transformer with global structure consistency for survival prediction. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2023; 2023:21184–21194.
4. Sun B, Peng Y, Ge Y. Multimodal survival analysis using optimal transport matching and global-local feature fusion. *Digit. Signal Process*. 2025;161:105119.
5. Gao F, Ding J, Gai B, et al. Interpretable multimodal fusion model for bridged histology and genomics survival prediction in pan-cancer. *Adv Sci*. 2025;12.
6. Krishna A, Kurian NC, Patil A, Parulekar A, Sethi A, Pathogen-x: a cross-modal genomic feature trans-align network for enhanced survival prediction from histopathology images, ArXiv abs/2411.00749, 2024.
7. Wang T, Fan J, Zhang D, et al., Mirror: multi-modal pathological self-supervised representation learning via modality alignment and retention, 2025.
8. Wu X, Shi Y, Wang M, Li A. Camr: cross-aligned multimodal representation learning for cancer survival prediction. *Bioinformatics*. 2023;39.
9. Xu Y, Liu H, Shi Y, Li A, Wang M. Cancer survival prediction by multimodal disentangled representation learning. *Proceedings of the 2023 6th International Conference on Machine Learning and Natural Language Processing*. 2023; 2023.
10. Qu M, Yang G, Di D, et al. Multimodal cancer survival analysis via hypergraph learning with cross-modality rebalance, ArXiv abs/2505.11997, 2025.
11. Hou W, Lin C, Yu L, Qin J, Yu R, Wang L. Hybrid graph convolutional network with online masked autoencoder for robust multimodal cancer survival prediction. *IEEE Trans Med Imaging*. 2023;42:2462–2473.
12. Song J, Hao Y, Zhao S, et al. Dual-stream cross-modal fusion alignment network for survival analysis. *Brief Bioinform*. 2025;26.
13. Li T, Zhou X, Xue J, et al. Cross-modal alignment and contrastive learning for enhanced cancer survival prediction. *Comput Methods Programs Biomed*. 2025;263:108633.
14. Braman N, Gordon J, Goossens ET, Willis CS, Stumpe MC, Venkataraman J. Deep orthogonal fusion: multimodal prognostic biomarker discovery integrating radiology, pathology, genomic, and clinical data. *International Conference on Medical Image Computing and Computer-Assisted Intervention*. 2021; 2021.
15. Boehm KM, Khosravi P, Vanguri RS, Gao J, Shah SP. Harnessing multimodal data integration to advance precision oncology. *Nat Rev Cancer*. 2021;22:114–126.
16. An Y, Lan R, Lin H, et al. Multimodal fusion framework based on low-rank interaction for tumor prognostic prediction. *IEEE Transactions on Computational Biology and Bioinformatics*. 2025.
17. Li S, Tang H, Multimodal alignment and fusion: a survey, ArXiv abs/2411.17040, 2024.
18. Song AH, Chen RJ, Jaume G, Vaidya A, Baras AS, Mahmood F, Multimodal prototyping for cancer survival prediction, ArXiv abs/2407.00224, 2024.
19. Li Y, Pan L, Peng Y, et al. Application of deep learning-based multimodal fusion technology in cancer diagnosis: a survey. *Eng Appl Artif Intell*. 2025;143:109972.
20. Waqas A, Tripathi A, Ramachandran R, Stewart P, Rasool G. Multimodal data integration for oncology in the era of deep neural networks: a review. *Front Artif Intell*. 2023;7.
21. Farhadizadeh M, Weymann M, Blass M, et al. A systematic review of challenges and proposed solutions in modeling multimodal data, ArXiv abs/2505.06945, 2025.
22. Barua A, Ahmed MU, Begum S. A systematic literature review on multimodal machine learning: applications, challenges, gaps and future directions. *IEEE Access*. 2023;11:14804–14831.
23. Baltrušaitis T, Ahuja C, Morency L-P. Multimodal machine learning: a survey and taxonomy. *IEEE Trans Pattern Anal Mach Intell*. 2017;41:423–443.

24. Kitchenham B, Charters S, Guidelines for performing systematic literature reviews in software engineering, 2, 2007.
25. Moher D, Liberati A, Tetzlaff JM, Altman DG. Preferred reporting items for systematic reviews and meta-analyses: the prisma statement. *PLoS Med*. 2009;6.
26. LipkovÁ J, Chen RJ, Chen B, et al. Artificial intelligence for multimodal data integration in oncology. *Cancer Cell*. 2022;40(10):1095–1110.
27. Liang PP, Zadeh A, Morency L-P. Foundations & trends in multimodal machine learning: principles, challenges, and open questions. *ACM Comput Surv*. 2022;56:1–42.
28. Ferlay J, Colombet M, Soerjomataram I, et al. Cancer statistics for the year 2020: an overview. *Int J Cancer*. 2021;149:778–789.
29. Feng X, Shu W, Li M, Li J, Xu J, He M. Pathogenomics for accurate diagnosis, treatment, prognosis of oncology: a cutting edge overview. *J Transl Med*. 2024;22.
30. Hu Y, Li X, Yi Y, Huang Y, Wang G, Wang D. Deep learning-driven survival prediction in pan-cancer studies by integrating multimodal histology-genomic data. *Brief Bioinform*. 2025;26.
31. Zhou H, Zhou F, Zhao C, Xu Y, Luo L, Chen H. Multimodal data integration for precision oncology: challenges and future directions, ArXiv abs/2406.19611, 2024.
32. Zheng X, Tang C, Wan Z, Hu C, Zhang W. Multi-level confidence learning for trustworthy multimodal classification. *AAAI Conf Artif Intell*. 2023; 2023.
33. Liu T, Huang J, Liao T, Pu R, Liu S, Peng Y. A hybrid deep learning model for predicting molecular subtypes of human breast cancer using multimodal data. *IRBM*. 2021.
34. Thangudu RR, Rudnick PA, Holck M, et al. Abstract lb-242: proteomic data commons: a resource for proteogenomic analysis. *Bioinform, Converg Sci, Syst Biol*. 2020.
35. Tomczak K, Czerwińska P, Wiznerowicz M. The cancer genome atlas (tcga): an immeasurable source of knowledge. *Contemp Oncol*. 2015;19:A68–A77.
36. Liu J, Lichtenberg TM, Hoadley KA, et al. An integrated tcga pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell*. 2018;173:400–416.e11.
37. Goldman MJ, Craft B, Hastie M, et al. Visualizing and interpreting cancer genomics data via the xena platform. *Nat Biotechnol*. 2020;38:675–678.
38. Waqas A, Bui MM, Glassy EF, et al. Revolutionizing digital pathology with the power of generative artificial intelligence and foundation models. *Lab Invest; J Tech Methods Pathol*. 2023:100255.
39. Quinn M, Forman JH, Harrod M, et al. Electronic health records, communication, and data sharing: challenges and opportunities for improving the diagnostic process. *Diagnosis*. 2018;6:241–248.
40. Subramanian V, Syeda-Mahmood TF, Do MN, Multi-modality fusion using canonical correlation analysis methods: application in breast cancer survival prediction from histology and genomics, ArXiv abs/2111.13987, 2021.
41. Montesuma EF, Mboula FN, Souloumiac A. Recent advances in optimal transport for machine learning. *IEEE Trans Pattern Anal Mach Intell*. 2023;47:1161–1180.
42. Imfeld M, Graldi J, Giordano M, Hofmann T, Anagnostidis S, Singh SP. Transformer fusion with optimal transport, ArXiv abs/2310.05719, 2023.
43. Xu Y, Wang Y, Zhou F, et al. A multimodal knowledge-enhanced whole-slide pathology foundation model, ArXiv abs/2407.15362, 2024.
44. Vaswani A, Shazeer NM, Parmar N, et al. Attention is all you need. *Neural Inf Process Syst*. 2017; 2017.
45. Liu M, Liu Y, Cui H, Li C. Mutual-guided cross-modality transformer for survival outcome prediction using integrative histopathology-genomic features. *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. 2023; 2023:1306–1312.
46. Mai S, Hu H, Xing S. Modality to modality translation: an adversarial representation learning and graph fusion network for multimodal fusion, ArXiv abs/1911.07848, 2019.

47. Hazarika D, Zimmermann R, Poria S. Misa: modality-invariant and -specific representations for multimodal sentiment analysis. *Proceedings of the 28th ACM International Conference on Multimedia*. 2020; 2020.
48. Yang J, Chen W, Xing X, et al. Hysurvpred: multimodal hyperbolic embedding with angle-aware hierarchical contrastive learning and uncertainty constraints for survival prediction, 2025.
49. Zhang Y, Jiang H, Miura Y, Manning CD, Langlotz C. Contrastive learning of medical visual representations from paired images and text. *Machine Learning in Health Care*. 2020; 2020.
50. Moradinasab N, Sengupta S, Liu J, Syed S, Brown DE. Towards robust multimodal representation: a unified approach with adaptive experts and alignment, 2025.
51. Zhang Z, Zhao Y, Duan J, Liu Y, Liang D, Li Z-C. Drop and align: fusing pathology and genomics via graph learning for cancer survival prediction. *2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI)*. 2025; 2025:1–5.
52. Ji Z, Ge Y, Chukwudi CC, et al. Counterfactual bidirectional co-attention transformer for integrative histology-genomic cancer risk stratification. *IEEE J Biomed Health Inform*. 2025.
53. Lu MY, Williamson DFK, Chen TY, Chen RJ, Barbieri M, Mahmood F. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nat Biomed Eng*. 2020;5:555–570.
54. Jaume G, Vaidya A, Chen RJ, Williamson DFK, Liang PP, Mahmood F. Modeling dense multimodal interactions between biological pathways and histology for survival prediction. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023; 2023:11579–11590.
55. Zhang Z, Zhao Y, Duan J, et al. Pathol1ogy-genomic fusion via biologically informed cross-modality graph learning for survival analysis, ArXiv abs/2404.08023, 2024.
56. Xu H, Usuyama N, Bagga J, et al. A whole-slide foundation model for digital pathology from real-world data. *Nature*. 2024;630:181–188.
57. Ghulam M, Alshehri F, Karray F, Saddik AE, Alsulaiman M, Falk TH. A comprehensive survey on multimodal medical signals fusion for smart healthcare systems. *Inf. Fusion*. 2021;76:355–375.
58. Louis DN, Perry A, Reifenberger G, et al. The 2016 world health organization classification of tumors of the central nervous system: a summary. *Acta Neuropathol*. 2016;131:803–820.
59. Kline AS, Wang H, Li Y, et al. Multimodal machine learning in precision health: a scoping review. *NPJ Digit Med*. 2022;5.
60. Chen RJ, Lu MY, Wang J, et al. Pathomic fusion: an integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis. *IEEE Trans Med Imaging*. 2019;41:757–770.
61. Yang H, Wang J, Wang W, et al. Mmsurv: a multimodal multi-instance multi-cancer survival prediction model integrating pathological images, clinical information, and sequencing data. *Brief Bioinform*. 2025;26.
62. Gaw N, Yousefi S, Gahrooei MR. Multimodal data fusion for systems improvement: a review. *IISE Trans*. 2021;54:1098–1116.
63. Zhou C, Wang H, Li X, Liu W, Liu Z. *Multimodal Survival Ensemble Network: Integrating Genomic and Histopathological Insights for Enhanced Cancer Prognosis*. IEEE; 2024:2330–2334.
64. Mao N, Dai Y, Zhou H, et al. A multimodal and fully automated system for prediction of pathological complete response to neoadjuvant chemotherapy in breast cancer. *Sci Adv*. 2025;11.
65. Stahlschmidt SR, Ulfenborg B, Synnergren J. Multimodal deep learning for biomedical data fusion: a review. *Brief Bioinform*. 2022;23.

66. Liu H, Yang H, Eduati F, Pluim JPW, Veta M. Adaptive prototype learning for multimodal cancer survival analysis, 2025.
67. Eijpe A, Lakbir S, Cesur ME, Oliveira SP, Abeln S, Silva W. Disentangled and interpretable multimodal attention fusion for cancer survival prediction, ArXiv abs/2503.16069, 2025.
68. Shi Y, Wang M, Liu H, Zhao F, Li A, Chen X. Mif: multi-shot interactive fusion model for cancer survival prediction using pathological image and genomic data. *IEEE J Biomed Health Inform*. 2024.
69. Qiu L, Khormali A, Liu K. Deep biological pathway informed pathology-genomic multimodal survival prediction, ArXiv abs/2301.02383, 2023.
70. Hemker K, Simidjievski N, Jamnik M. Healnet: multimodal fusion for heterogeneous biomedical data. *Neural Inf Process Syst*. 2023; 2023.
71. Mao S, Liu J. Mulitdeepsurv: survival analysis of gastric cancer based on deep learning multimodal fusion models, biomedical. *Opt Express*. 2024;16:126–141.
72. Pan L, Peng Y, Li Y, et al. Foresee: multimodal and multi-view representation learning for robust prediction of cancer survival, ArXiv abs/2405.07702, 2024.
73. Luo H, Huang J, Ju H, Zhou T, Ding W. Multimodal multi-instance evidence fusion neural networks for cancer survival prediction. *Sci Rep*. 2025;15.