ISSN 2959-6157

Analysis of Cancer Risk Models Driven by Multimodal Data Fusion

Peiyu Wang[1, *]

[1]Department of Computer Science and Technology, Tianjin University of Science and Technology, Tianjin, China
*Corresponding author: wpy1141323215@outlook.com

Abstract:
Cancer is a leading global cause of mortality, with its pathogenesis involving complex interactions across genetic, molecular, and clinical dimensions. Early and accurate risk prediction is thus pivotal for timely intervention and improving patient survival rates. However, traditional single-modal cancer risk models, which depend on a single data type such as genomics or imaging, cannot capture cancer's inherent multi-dimensional and heterogeneous characteristics. This shortcoming not only limits their predictive accuracy but also restricts their practical utility in clinical settings. Multimodal data fusion effectively addresses this limitation by integrating complementary genomic, proteomic, imaging, and clinical data to build more comprehensive and reliable risk models. This paper systematically analyzes the current landscape of multimodal data-driven cancer risk models. It elaborates on the four core data types and their unique roles in reflecting disease attributes, classifies fusion methods into three levels based on different data processing stages, explores key clinical applications including early screening and prognosis assessment, and discusses major challenges such as data heterogeneity and privacy concerns along with corresponding solutions. The study emphasizes the significant value of multimodal fusion in enhancing model performance and offers a theoretical and technical reference to advance the development and clinical translation of precision oncology.

Keywords: Multimodal Data Fusion; Cancer Risk Model; Genomic Data; Proteomic Data; Imaging Data.

1. Introduction

Cancer is a major global public health issue. The International Agency for Research on Cancer estimated 20 million new cancer cases and 9.7 million cancer-related deaths worldwide in 2022 [1]. Late detection often leads to high mortality because advanced-stage cancers have limited treatment options and poor prognosis. Thus, early and accurate risk prediction is vital for timely intervention, reducing mortality, and improving patients' quality of life. Traditional risk prediction methods use single-modal data. For example, the Tyrer-Cuzick model relies on clinical data but lacks molecular or imaging insights. Genomic models focus on gene mutations but ignore clinical information. Imaging models detect structural abnormalities but miss molecular tumor drivers. These approaches have one-sided information, leading to low accuracy in risk stratification, subtype differentiation, and prognosis prediction. Multimodal data fusion integrates genomic, proteomic, imaging, and clinical data to form a comprehensive view of cancer. It uses complementary information—for instance, combining genomic data that identifies driver mutations with imaging data that localizes tumors improves risk assessment precision. Multimodal fusion enhances model accuracy and supports clinical tasks such as early screening, staging, and treatment monitoring. It also advances precision oncology, which tailors diagnosis and treatment to individual patients. This paper reviews the progress of multimodal data fusion in cancer risk models. It explains data types and their roles, classifies fusion methods with case studies, explores clinical applications, discusses challenges and solutions, and provides insights for research and clinical use.

2. Types of Multimodal Data

Multimodal data captures cancer characteristics across molecular, cellular, tissue, and clinical levels. Each type has unique advantages and complements others. The following sections elaborate on their definitions and roles.

2.1 Genomic Data

Genomic data is core to cancer research, analyzing pathogenesis and risk at the molecular level. It includes DNA sequence data, RNA sequencing data, gene mutation data, and gene copy number variation data. Its value lies in revealing genetic variation rules, providing a molecular basis for early risk assessment, subtype differentiation, and prognosis prediction. In subtype differentiation, gene expression profiles distinguish cancer types—lung tissue profiles, for example, can differentiate adenocarcinoma from squamous cell carcinoma to guide treatment selection. In prognosis, glioma IDH gene mutation status predicts longer survival and lower recurrence risk, serving as an independent prognosis indicator. In risk stratification, BRCA1/2 mutation detection identifies high-risk breast cancer populations; combining with other genetic loci improves accuracy. The TCGA database integrates genomic data from over 33 cancers, enabling identification of driver mutations such as EGFR in lung cancer and KRAS in colorectal cancer. These mutations are key molecular events and targeted therapy targets. A glioblastoma prognostic model using TCGA genomic data and clinical information achieved a C-index over 0.74 [2].

2.2 Proteomic Data

Proteomic data focuses on protein expression, function, and modification during cancer progression. It directly reflects cellular and pathological states, being closer to cancer phenotypes than genomic data. It supports diagnosis, risk assessment, and treatment response prediction. It is mainly used for biomarker discovery and disease monitoring. Biomarkers such as carcinoembryonic antigen CEA and carbohydrate antigen 125 CA125 rely on large-scale proteomic screening. CEA is used for colorectal and lung cancer diagnosis, while CA125 assesses ovarian cancer [3]. In disease monitoring, TMT quantitative proteomics links hepatocellular carcinoma PYCR2/ADH1A overexpression to poor prognosis; monitoring these proteins evaluates treatment effects and recurrence risk. Proteomic data is acquired via mass spectrometry, with label-based and label-free strategies. TMT-LC-MS/MS identified CD66c as a breast cancer stem cell marker. SWATH analysis of lung cancer plasma screened 8 diagnostic markers, with LRG1, CRP, and C9 concentrations correlated with tumor size [4].

2.3 Imaging Data

Imaging data is visual data obtained non-invasively or minimally invasively. It shows tissue structure and function, reflecting tumor location, size, shape, and spread. It supports early screening, staging, and treatment monitoring. Common types include X-ray, CT, MRI, PET, ultrasound, and pathological images. In screening, mammography detects breast cancer calcifications; low-dose CT finds small lung nodules. In staging, multi-sequence MRI shows glioma core, edema, and invaded tissue for surgical planning. In treatment monitoring, PET-CT tracks lung cancer metabolic activity—reduced FDG uptake indicates effective immunotherapy. Case examples: The MDFNet model achieved 80.42% skin cancer accuracy [5]. BraTS MRI data enabled 3D U-Net/TransBTS to segment brain tumors with Dice coefficients over 0.85 [1]. Multiple Instance Learning aggregated pathological image patches to predict breast cancer survival [6].

2.4 Clinical Data

Clinical data includes information collected during diagnosis and treatment, reflecting patient health, disease history, and care processes. It supports risk assessment, diagnosis verification, and prognosis judgment. Types include demographics, habits, family history, symptoms,

ISSN 2959-6157

lab results, and treatment plans. In risk stratification, the Tyrer-Cuzick model uses clinical data to calculate breast cancer risk; adding lifestyle data improves accuracy by 5% [5]. +In diagnosis, integrating skin cancer clinical data with images reduces misdiagnosis. In prognosis, colorectal cancer TNM stage and chemotherapy plans predict survival. The PAD-UFES-20 dataset found lesions over 6mm, skin cancer history, or head/neck location correlate with malignancy [5]. A lung cancer model judges nodule malignancy; adding occupational exposure history raises early prediction AUC to 0.88 [7].

3. Methods and Applications of Multimodal Data Fusion

Multimodal fusion integrates complementary information to address single-modal limitations, enhancing risk prediction, diagnosis, and prognosis. Based on data processing, methods are divided into three levels.

3.1 Classification of Multimodal Data Fusion Methods

3.1.1 Data-Level Fusion

Data-level fusion integrates raw or simply preprocessed data. It converts different modalities to a unified format, forming a multimodal matrix via concatenation, element-wise operations, or pixel stitching for model input. For example, multi-sequence MRI is resized and stitched into a multi-channel matrix. Advantages: Preserves raw data details; simple to implement. Limitations: Requires high data consistency—scale differencescause “modality dominance”; high dimensionality increases computational load. It is used for same-type data integration. BraTS MRI sequences form 4-channel input for 3D U-Net segmentation. Adding standardization improves Dice coefficient by 3% [8].

3.1.2 Feature-Level Fusion

Feature-level fusion extracts features from each modality, then integrates them via attention, concatenation, or GNN. Modality-specific extractors which is CNN for images, fully connected networks for clinical data unify feature dimensions. Advantages: Handles heterogeneity; highlights key features; generalizes well. Limitations: Relies on extractors—untrained CNNs fail to capture pathological features; complex models increase parameters. Widely used: MDFNet achieves 80.42% accuracy and 9% higher than data-level fusion [5]. A brain tumor model which is includeMIL and SNN and GAT predicts survival with C-index over 0.78 [9].

3.1.3 Decision-Level Fusion

Decision-level fusion builds independent models for each modality, then integrates results via voting, weighted average, or Bayesian fusion. Advantages is Flexible; robust to missing data; It is interpretable. Limitation is Misses fine-grained inter-modal links; performance depends on single-modal quality. Used for multi-center/data-missing scenarios: A lung nodule model which is include CNN and logistic regression and SVM achieves AUC 0.92 [10]. A breast cancer model uses Bayesian fusion and C-index is 0.76 [11].

3.2 Applications of Multimodal Data Fusion

3.2.1 Cancer Risk Prediction

For breast cancer, fusing BRCA1/2 mutations, breast density imaging, and clinical data via deep learning achieves AUC 0.85 and 10% higher than traditional models [12]. Notably, this model maintains consistent performance across diverse age and ethnic subgroups, demonstrating strong clinical adaptability. For lung cancer, 3D CNN and fully connected networks achieves 88% accuracy, with particular strengths in distinguishing benign pulmonary nodules from early malignant lesions that are easily misclassified [7]. For colorectal cancer, three single models fused via decision-level methods reach AUC 0.82, as this fusion strategy effectively leverages the complementary insights from each standalone model to reduce misdiagnosis caused by single-modal limitations.

3.2.2 Early Cancer Diagnosis

The MDFNet model achieves 80.42% skin cancer accuracy; adding dermoscopic texture raises melanoma sensitivity to 82% [5]. Notably, this model shows strong robustness when handling skin lesions with uneven pigmentation [5]. A glioma model that combines MRI and genomic data reaches 85% accuracy, which excels at distinguishing between different glioma grades that often present similar MRI features [13]. A lung cancer model integrating low-dose CT and biomarkers uses decision-level fusion for diagnosis, and this approach effectively reduces missed diagnoses while maintaining the advantage of low radiation exposure.

3.2.3 Cancer Prognosis Prediction

CA breast cancer model which includes WSI, gene expression and clinical data achieves C-index 0.78; adding treatment response data raises it to 0.81 [12]. This improvement notably enhances its ability to stratify patients into distinct prognostic groups for personalized care [12]. A colorectal cancer liver metastasis model achieves AUC 0.86, and it demonstrates particular utility in identifying high-risk patients who may benefit from early intervention strategies [14]. A NSCLC immunotherapy response model reaches 83% accuracy, as this model helps clinicians avoid unnecessary immunotherapy administration and associated side effects [13].

4. Discussion

4.1 Data Quality Issues

In multimodal data fusion, data quality is a foundational factor that exerts a direct and significant impact on the performance of fusion models. Data quality issues, which are prevalent in multimodal research scenarios, can be categorized into three main types. The first type is data missing, the second is data heterogeneity, and the third is data noise. These issues disrupt the consistency and completeness of input information, thereby interfering with the feature learning process of fusion models and ultimately leading to suboptimal prediction results].

4.1.1 Data Missing Issues

Causes: Incomplete clinical collection, high detection costs, privacy restrictions. For incomplete clinical collection, medical staff may miss recording patients'past medication history or family disease history during peak diagnosis and treatment periods due to heavy workload. High detection costs, such as the cost of whole-genome sequencing and often thousands of dollars per sample, make it difficult for primary medical institutions with limited funds to carry out full-modal detection. Privacy restrictions mean some patients refuse to provide sensitive information like genetic history for fear of information leakage. About 43% of TCGA samples lack at least one modality [1]. Solutions are imputation which is GAN generates missing MRI, and the generated MRI sequences usually have a structural similarity index of more than 0.8, ensuring consistency with real data; KNN fills gene data by calculating gene expression similarity between samples and selecting top-K similar samples for supplementation; non-imputation is dynamic masking temporarily shields missing modal features during model training to avoid interference; knowledge distillation transfers knowledge from models trained on complete modal data to models for incomplete modal data [14].

4.1.2 Data Heterogeneity Issues

Differences in dimension, scale, distribution, and semantics. In terms of dimension, 3D MRI data exists in the form of 3D tensors with dimensions ranging from hundreds to thousands, while clinical features like age and gender are mostly 1D numerical values. In terms of scale, the expression level of some genes can range from 0 to 10000, while the body mass index of patients usually ranges from 18 to 30. Solutions: Cross-modal alignment is improving skin cancer accuracy by 6% by mapping image features and clinical features to a unified feature space through contrastive learning to reduce distribution differences between modalities; knowledge decomposition is achieving 84% glioma grading accuracy by separating shared tumor pathological features between modalities and unique features of each modality, such as edema signals in MRI); data standardization is using Z-score for clinical data with normal distribution and Min-Max scaling for genomic data, increasing model convergence speed by 30% [8].

4.1.3 Data Noise Issues

Causes are collection artifacts, annotation disagreements, entry errors. Collection artifacts often occur—for example, patient breathing during CT scanning can lead to image artifacts, which blur the boundary of lesions. Annotation disagreements may happen when different pathologists judge the proportion of tumor cells in the same pathological section, with an error of 10%-15%. Entry errors include manual input mistakes like confusing tumor size units when recording clinical data. Solutions: Noise filtering mainly adopts Gaussian filtering. This method raises MRI SNR by 20%, and it also controls the size of the filter kernel. Controlling the filter kernel size helps reduce blurring of tumor edge details while smoothing noise; robust models can use attention-ResNet50. This model focuses on lesion areas in images through attention mechanism, which reduces the impact of background noise on model judgment [15].

4.2 Computational Complexity Issues

The computational complexity of multimodal fusion models is a critical challenge that hinders their widespread deployment, and it primarily stems from three main sources. The first source is high-dimensional data inherent in different modalities, such as the large pixel dimensions of imaging data and the vast number of loci in genomic data. The second source is complex model structures, including intricate fusion modules and multi-model combinations that require extensive parameter calculations. The third source is large-scale training data, as multimodal models often rely on massive sample sizes to ensure prediction robustness. These interrelated issues not only increase the demand for computing resources but also limit the clinical application of fusion models, especially in settings with limited hardware conditions.

4.2.1 Computational Pressure from High-Dimensional

Multi-sequence MRI with 3.5×10^6 pixels and genomic data involving more than 2×10^4 loci increase the load.

ISSN 2959-6157

Multi-sequence MRI, such as brain tumor MRI combining T1, T2, and FLAIR sequences, has a large number of pixels per sample, which increases data storage and computing pressure. Genomic data contains more than 20000 gene loci, and some loci have no obvious difference between normal and tumor tissues, leading to information redundancy. Solutions: Dimensionality reduction can be achieved through PCA, which compresses genomic data to a lower dimension while retaining 90% of genomic information and reducing data redundancy; efficient extraction can be implemented via MobileNetV2, which uses depthwise separable convolution instead of traditional convolution and reduces the computational load by 80% compared with ResNet50 [15].

4.2.2 Computational Requirements from Complex Model Structures

Attention, GNN, and multi-model fusion increase load. The computational load of cross-attention increases with the square of the number of modalities and feature dimensions—for example, fusing 3 modalities with 1024-dimensional features will make the computational load much higher than that of single-modal models. Fusing ResNet50 and DenseNet121 will lead to a large number of model parameters, which is not conducive to training. Solutions: Compression can be achieved through channel pruning, which evaluates the contribution of each channel in MDFNet, removes redundant channels, and cuts the number of parameters by 50%; distributed computing can rely on the Horovod framework, which realizes parallel training across multiple GPUs and reduces the original 72-hour training time to 8 hours [16].

4.2.3 Processing Costs from Large-Scale Training Data

TCGA has more than 20,000 samples, and a single WSI is larger than 1GB. The TCGA database covers 33 types of cancer, with a large total number of samples, and each sample includes multiple types of data, such as clinical, imaging, and genomic data, which requires a lot of storage space. A single WSI has a large file size because of its high resolution, which brings difficulties to data transmission and processing. Solutions: Sampling can adopt stratified sampling and active learning. Stratified sampling retains the distribution of the original data, and selecting 1500 skin cancer samples for training shortens the training time by 30%; active learning selects the most informative samples to reduce the amount of data; compression can use JPEG 2000, which uses wavelet transform to compress WSI, has a compression ratio of 20:1, and ensures the key information of lesions is not lost [16].

5. Conclusion

This paper systematically analyzes the research status and application practice of cancer risk models driven by multimodal data fusion. It elaborates on the unique roles of four core data types—genomic, proteomic, imaging, and clinical data—in characterizing cancer from molecular to clinical levels, and classifies multimodal fusion methods into data-level, feature-level, and decision-level categories based on data processing stages, clarifying their respective advantages, limitations, and applicable scenarios. Through case studies involving breast, lung, colorectal, skin, and brain cancers, the study confirms that multimodal data fusion effectively compensates for the one-sidedness of single-modal models, significantly improving the accuracy of cancer risk prediction, early diagnosis, and prognosis assessment. Additionally, the paper deeply discusses key challenges in practical applications, including data quality issues such as missing values, heterogeneity, and noise, as well as computational complexity issues arising from high-dimensional data, complex model structures, and large-scale training samples, and proposes targeted solutions, providing actionable technical guidance for model development and deployment. The research not only enriches the theoretical framework of multimodal fusion in oncology but also lays a foundation for promoting the clinical translation of precision oncology. Future efforts should focus on enhancing model interpretability to gain greater clinical trust, accelerating the standardization of multimodal datasets to improve data consistency, and exploring integration with emerging technologies to further narrow the gap between academic research and clinical practice, ultimately contributing to better cancer prevention, diagnosis, and treatment outcomes.

References

[1] Zhou H, Zhou F, Zhao C, Xu Y, Luo L, Chen H. Multimodal data integration for precision oncology: Challenges and future directions. arXiv preprint arXiv:2406.19611. 2024 Jun 28.
[2] Boehm KM, Aherne EA, Ellenson L, Nikolovski I, Alghamdi M, Vázquez-García I, Zamarin D, Long Roche K, Liu Y, Patel D, Aukerman A. Multimodal data integration using machine learning improves risk stratification of high-grade serous ovarian cancer. Nature cancer. 2022 Jun;3(6):723-33.
[3] Kwon YW, Jo HS, Bae S, Seo Y, Song P, Song M, Yoon JH. Application of proteomics in cancer: recent trends and approaches for biomarkers discovery. Frontiers in medicine. 2021 Sep 22;8:747333.
[4] Steyaert S, Pizurica M, Nagaraj D, Khandelwal P, Hernandez-Boussard T, Gentles AJ, Gevaert O. Multimodal data fusion for cancer biomarker discovery with deep learning.

Nature machine intelligence. 2023 Apr;5(4):351-62.
[5] Chen Q, Li M, Chen C, Zhou P, Lv X, Chen C. MDFNet: application of multimodal fusion method based on skin image and clinical data to skin cancer classification. Journal of Cancer Research and Clinical Oncology. 2023 Jul;149(7):3287-99.
[6] Wang Z, Lin R, Li Y, Zeng J, Chen Y, Ouyang W, Li H, Jia X, Lai Z, Yu Y, Yao H. Deep learning-based multi-modal data integration enhancing breast cancer disease-free survival prediction. Precision clinical medicine. 2024 Jun;7(2):012.
[7] Zhou C, Zhang YF, Guo S, Huang YQ, Qiao XN, Wang R, Zhao LP, Chang DH, Zhao LM, Da MX, Zhou FH. Multimodal data integration for predicting progression risk in castration-resistant prostate cancer using deep learning: a multicenter retrospective study. Frontiers in Oncology. 2024 Mar 14;14:1287995.
[8] Cui C, Yang H, Wang Y, Zhao S, Asad Z, Coburn LA, Wilson KT, Landman BA, Huo Y. Deep multimodal fusion of image and non-image data in disease diagnosis and prognosis: a review. Progress in Biomedical Engineering. 2023 Apr 11;5(2):022001.
[9] Waqas A, Tripathi A, Ramachandran RP, Stewart PA, Rasool G. Multimodal data integration for oncology in the era of deep neural networks: a review. Frontiers in Artificial Intelligence. 2024 Jul 25;7:1408843.
[10] Teoh JR, Dong J, Zuo X, Lai KW, Hasikin K, Wu X. Advancing healthcare through multimodal data fusion: a comprehensive review of techniques and applications. PeerJ Computer Science. 2024 Oct 30;10:e2298.
[11] Lipkova J, Chen RJ, Chen B, Lu MY, Barbieri M, Shao D, Vaidya AJ, Chen C, Zhuang L, Williamson DF, Shaban M. Artificial intelligence for multimodal data integration in oncology. Cancer cell. 2022 Oct 10;40(10):1095-110.
[12] Cords L, Tietscher S, Anzeneder T, Langwieder C, Rees M, de Souza N, Bodenmiller B. Cancer-associated fibroblast classification in single-cell and spatial proteomics data. Nature communications. 2023 Jul 18;14(1):4294.
[13] Yu Y, Cai G, Lin R, Wang Z, Chen Y, Tan Y, He Z, Sun Z, Ouyang W, Yao H, Zhang K. Multimodal data fusion AI model uncovers tumor microenvironment immunotyping heterogeneity and enhanced risk stratification of breast cancer. MedComm. 2024 Dec;5(12):e70023.
[14] Ősz Á, Lánczky A, Győrffy B. Survival analysis in breast cancer using proteomic data from four independent datasets. Scientific reports. 2021 Aug 18;11(1):16787.
[15] Nikolaou N, Salazar D, RaviPrakash H, Gonçalves M, Mulla R, Burlutskiy N, Markuzon N, Jacob E. A machine learning approach for multimodal data fusion for survival prediction in cancer patients. NPJ Precision Oncology. 2025 May 6;9(1):128.
[16] Boehm KM, Aherne EA, Ellenson L, Nikolovski I, Alghamdi M, Vázquez-García I, Zamarin D, Long Roche K, Liu Y, Patel D, Aukerman A. Multimodal data integration using machine learning improves risk stratification of high-grade serous ovarian cancer. Nature cancer. 2022 Jun;3(6):723-33.