Cancer Diagnosis and Prediction Based on Multimodal AI Algorithms

Han Zhou[1,a,*]

[1]Sydney Smart Technology College, Northeastern, University, Qinhuangdao, Hebei, China
a. zhouhan040808@gmail.com
**corresponding author*

Abstract: The integration of multimodal artificial intelligence (AI) has shown immense promise in enhancing cancer detection and diagnosis by leveraging diverse medical data, such as imaging, genomic, and clinical records. Traditional diagnostic methods, while effective in certain contexts, are limited by their inability to comprehensively capture the complex characteristics of diseases. Multimodal AI addresses these limitations by synthesizing data from multiple sources, leading to more precise and early-stage detection of cancer. This paper provides an in-depth analysis of key multimodal fusion methods, including feature-level fusion, decision-level fusion, and dataset-level fusion, each offering distinct advantages and challenges. By reviewing the current state of multimodal AI applications in cancer diagnostics, this paper highlights the strengths of these methods, explores their limitations, and discusses potential solutions for improving data privacy, evaluation standards, and explainability. Furthermore, the paper outlines future directions for multimodal AI, emphasizing its transformative potential in revolutionizing personalized cancer treatment and early intervention strategies.

Keywords: Feature Fusion, Decision Fusion, Dataset-Level Fusion, Explainability in Multimodal Diagnostics.

1. Introduction

Early detection and accurate diagnosis of cancer are crucial for improving patient outcomes and survival rates. Traditional unimodal approaches, such as imaging studies, pathological analyses, and molecular biology techniques, have shown effectiveness in certain contexts. However, these methods often struggle to comprehensively capture the complex characteristics of diseases and are influenced by technical limitations, clinician experience, and data quality. Against this backdrop, the application of multimodal artificial intelligence (AI) represents a significant innovation in the field of cancer diagnosis. Multimodal AI integrates medical data from diverse sources, such as imaging data, genomic information, and clinical records, enabling comprehensive analysis of diseases from multiple perspectives. This capability significantly enhances diagnostic accuracy, particularly in the early stages of cancer, by identifying subtle lesions and enabling earlier intervention [1].

Compared to traditional unimodal methods, multimodal AI offers the advantage of synthesizing diverse data types to produce more precise diagnostic results. For instance, by combining CT scans, MRI images, and genomic data, AI systems can detect early lesions and potential cancer risks that conventional methods might overlook. This multidimensional data fusion not only improves the sensitivity of early disease detection but also enhances the ability to diagnose various cancer types and account for individual patient differences, supporting the development of personalized treatment plans [2].

This paper provides a comprehensive review of the application and development of different multimodal feature fusion techniques in cancer detection. It will explore major feature fusion methods and analyze how they integrate information such as imaging, genomics, and clinical data to improve diagnostic accuracy. Furthermore, this study examines the strengths and limitations of these methods, summarizes the current state of multimodal AI in cancer diagnosis, and identifies existing challenges. Finally, potential solutions and future directions for this field will be proposed.

2. Multimodal Fusion

The integration of features from different modalities is a critical research topic in multimodal AI. In the field of cancer detection and prevention, mainstream multimodal AI primarily employs three methods: feature-level fusion, decision-level fusion, and dataset-level fusion.

2.1. Feature-Level Fusion

Feature-level fusion integrates data from various modalities (e.g., CT and MRI) during the feature extraction phase. Deep learning models, such as Convolutional Neural Networks (CNNs) or 3D-CNNs, are used to extract key features from each modality, which are then fused at deeper network layers to enable comprehensive lesion analysis [3]. The primary advantage of feature-level fusion lies in its holistic utilization of information and high accuracy, making it suitable for analyzing complex tumor characteristics at a deeper level. However, this method demands precise data alignment and is computationally intensive, posing higher requirements for system performance and resources [4].

2.2. Decision-Level Fusion

Decision-level fusion processes data from each modality independently, generating separate diagnostic results before combining them at the decision layer. This approach is more flexible in handling data and avoids the alignment challenges associated with feature-level fusion while requiring less computational power. However, decision-level fusion involves shallower information integration, which may result in the loss of certain feature details, potentially impacting diagnostic accuracy [5].

2.3. Dataset-Level Fusion

Dataset-level fusion directly integrates data from various modalities (e.g., CT, MRI, and genomic data) at the data layer to create a unified dataset for model processing. Unlike feature-level or decision-level fusion, this method operates at the raw data stage, combining or concatenating inputs to incorporate multimodal information. Its advantage lies in preserving the original information from each modality, making it suitable for scenarios with high data heterogeneity. However, this approach requires standardization and normalization when processing data of varying formats and scales, increasing the complexity and computational burden. It is particularly effective for tasks that involve high-dimensional data and require the retention of raw information, such as joint analyses of CT images, genomic data, and clinical records [6].

2.4. Applicability of Fusion Methods

Feature-level fusion is best suited for scenarios that demand deep analysis of interrelationships within multimodal data. For instance, integrating CT, MRI, and genomic data during the feature extraction phase allows models to learn subtle inter-modal feature relationships, making it ideal for high-accuracy cancer detection and segmentation tasks.

Decision-level fusion is more appropriate for cases where multimodal data exhibit a high degree of independence or where data acquisition difficulty varies. For example, independent models can be trained on imaging, genomic data, and clinical text, and their diagnostic results can be aggregated at the decision layer. This approach is suitable for rapid diagnosis and applications involving heterogeneous data. Dataset-level fusion excels in scenarios requiring the integration of raw data from diverse modalities while maintaining information integrity. For example, combining CT and MRI images with genomic data and clinical records into a unified dataset enables models to learn interrelations among modalities at a single input level. This makes it well-suited for tasks involving highly heterogeneous data and comprehensive analyses, such as early tumor screening and multidimensional diagnosis.

3. Applicability of Fusion Methods

Feature-level fusion is ideal for tasks requiring deep analysis of multimodal data relationships, such as high-accuracy cancer detection and segmentation. Decision-level fusion is suited for tasks where data are independent or when quick diagnoses are needed, such as rapid imaging or genomic analysis. Dataset-level fusion excels in tasks needing raw data integration and the preservation of modality-specific information, like early tumor screening and multidimensional diagnosis. Multimodal AI-assisted cancer diagnosis and prediction

3.1. Feature-Level Fusion

Feature-level fusion integrates data by combining features extracted from different modalities during the feature extraction stage. A study by Cui et al. utilized stacked autoencoders (SAE) to process imaging data from the Cancer Genome Atlas (TCGA) dataset and genomic data, effectively reducing dimensionality and denoising the data. Extracted features were fused into a unified representation and optimized for classification using a multilayer perceptron (MLP), achieving a 15% improvement in diagnostic accuracy and sensitivity [7]. Another example is Zhang et al., which employed 3D convolutional neural networks (3D-CNNs) to integrate MRI modalities, such as T1, T2, and DWI, for cervical cancer detection. This approach captured intricate 3D spatial relationships between modalities, significantly improving Dice coefficients and sensitivity [8]. However, both methods highlight challenges in computational complexity and the reliance on high-quality, well-aligned datasets, underscoring the need for efficient preprocessing strategies in practical applications.

3.2. Decision-Level Fusion

Decision-level fusion aggregates predictions from independently trained models, offering a flexible way to handle heterogeneous data. Zhang et al. proposed a hybrid framework where CNN, Long Short-Term Memory (LSTM), and Gated Recurrent Unit (GRU) models processed imaging, genomic, and clinical data, respectively. Predictions were integrated using a weighted voting mechanism, achieving 98% accuracy in breast cancer survival prediction [9]. Shen et al. applied decision-level fusion to CT imaging and clinical time-series data for lung cancer survival prediction, using CNN and Recurrent Neural Network (RNN) models with a weighted aggregation strategy. The approach yielded 96% accuracy and high F1 scores, demonstrating its robustness [10]. Despite these strengths, both studies acknowledge a lack of deep cross-modal interactions, limiting the approach in tasks requiring high inter-modality correlation.

3.3. Dataset-Level Fusion

Dataset-level fusion processes multimodal raw data in an integrated manner to allow for end-to-end learning. Lee et al. demonstrated this method by fusing molecular imaging and genomic data, achieving a sensitivity of 92% and specificity of 89%. Their framework relied on deep neural networks to align and standardize raw inputs, though preprocessing complexity presented challenges [11]. Similarly, Rahman et al. developed a stacked autoencoder-based framework that integrated MRI and clinical data for breast cancer diagnosis. Their model achieved sensitivity and specificity rates of 93% and 91%, respectively [12]. While these studies validate the potential of dataset-level fusion in cancer detection, both emphasize the critical need for robust preprocessing pipelines to address data misalignment and standardization issues.

4. Current Limitations and Future Prospects

The practical application of multimodal AI in the medical field still faces many challenges, particularly in the areas of data privacy, evaluation standard completeness, and explainability.

4.1. Data Privacy

Medical data typically contains sensitive information, such as patient histories, imaging, and genomic data, which requires special attention to patient privacy during data sharing and storage. In multimodal AI, where models extract information from various types of data (e.g., imaging, text, and genomic data), data integration increases the risk of privacy breaches. Several methods can protect patient data privacy [13].

Federated learning enables healthcare institutions to collaboratively train models without sharing actual data. By training models locally and only sharing parameter updates, data privacy is ensured. Federated learning has been applied in several medical AI projects, effectively mitigating privacy concerns. It focuses on solving the issue of non-shared data in multi-center collaborations, reducing the risk of privacy breaches through local training [14].

Differential privacy introduces random noise into the data to obscure personal information, preventing privacy leakage. When combined with federated learning or centralized training datasets, differential privacy can further reduce privacy risks. It focuses on protecting the privacy of sensitive data by masking personal information and enhancing the security of data sharing.

Homomorphic encryption allows computation on encrypted data without the need for decryption. Although it incurs significant computational overhead, this encryption method provides strong data security for highly sensitive scenarios, ensuring that data is not intercepted or tampered with during transmission and processing.

4.2. Evaluation Standards

Evaluation standards for multimodal AI applications in healthcare are complex. While traditional metrics like accuracy, sensitivity, specificity, and the Dice coefficient assess model performance effectively, they are insufficient in evaluating the comprehensive performance of multimodal tasks. For example, existing evaluation standards are not adequate for evaluating cross-modal analytical abilities, adaptability to different modality data, and model generalization [15].

In addition to basic classification and segmentation metrics, dimensional assessments of feature contribution and modality consistency can be introduced. For example, by calculating the contribution of each modality's features to the final diagnosis, these evaluations examine whether the model effectively integrates multimodal data. This method aims to address the gaps in current standards regarding multimodal data fusion and overall performance evaluation, focusing on the adaptability and contribution of each modality [16].

Testing model robustness and consistency on multi-center datasets ensures the model’s performance across diverse data environments. Establishing cross-institutional evaluation standards ensures stable diagnostic results across different devices, patient populations, and imaging conditions, solving generalization issues. This approach enhances the model’s applicability through validation across multiple data sources.

4.3. Explainability

Explainability is critical in medical AI, especially for high-risk tasks such as cancer detection, where clinicians need to understand the rationale behind model decisions. However, multimodal AI models often feature highly complex network structures, and the process of fusing features from different modalities makes it difficult to interpret model decisions. This "black-box" nature hinders clinician trust and can create uncertainties in clinical applications.

Explanation tools such as SHAP and LIME calculate each feature's contribution to the model's predictions, generating visualizations that help clinicians understand the key features the model focuses on. For instance, SHAP can reveal which modality or feature plays a crucial role in cancer diagnosis, improving the model's transparency [17].

Grad-CAM is a heatmap method for neural networks that uses backpropagation to locate areas of an image that the model is focusing on, helping clinicians understand the basis of image analysis. This method can also be applied to multimodal imaging data to show areas of interest across different modalities [18].

By incorporating modality weight learning modules during training, the model can automatically learn and display the relative weight of each modality in its predictions. This method provides insights into the contribution of different modalities, helping clinicians understand the importance of each data type in decision-making.

5. Conclusion

Multimodal AI has the potential to significantly advance cancer diagnosis by integrating diverse data sources to improve diagnostic accuracy and speed. Methods such as feature-level fusion, decision-level fusion, and dataset-level fusion each bring unique strengths to the table, offering flexible solutions to the challenges of data integration and analysis. Despite its promise, the application of multimodal AI in healthcare is not without challenges, particularly in data privacy, evaluation standards, and explainability. Addressing these issues through innovations in federated learning, differential privacy, and enhanced evaluation frameworks will be essential for the broader adoption of these technologies. Future advancements in AI explainability and privacy-preserving techniques will further enhance the trust and utility of multimodal AI in clinical settings. As these technologies continue to evolve, multimodal AI is poised to play a pivotal role in early cancer detection, personalized treatment, and ultimately improving patient outcomes across diverse healthcare environments.

References

[1] *Han, J., & Lee, K. (2022). Enhancing diagnostic accuracy with multimodal AI. Journal of Medical Systems, 46(3), 52. https://doi.org/10.1007/s10916-022-01711-2*

[2] *Bommasani, R., Aghi, K., & Achiam, C. (2023). Multimodal learning for clinical applications. Nature Biomedical Engineering, 7(2), 145–156.*

[3] *Singh, S., & Thomas, T. (2021). Feature-level fusion in medical imaging: A review. IEEE Transactions on Medical Imaging, 40(6), 1201–1215.*

[4] Tripathi, R., Achiam, C., & Chan, S. (2024). Advances in histopathology-based fusion methods. IEEE Transactions on Biomedical Engineering, 71(1), 56–68.

[5] Ahmed, M., & Salahuddin, Z. (2022). Decision-level fusion for multimodal medical diagnosis. PLOS Computational Biology, 18(3), e1010001.

[6] Smith, J., & Brown, A. (2023). Dataset fusion approaches in medical AI. Nature Reviews Machine Learning, 2(4), 123–135. https://doi.org/10.1038/s42256-023-00101-3

[7] Cui, C., Zhang, J., Liu, T., & Wang, H. (2023). Deep multi-modal fusion of image and non-image data in disease diagnosis and prognosis. arXiv. https://doi.org/10.48550/arXiv.2203.15588

[8] Zhang, Y., Liu, C., & Wei, L. (2023). Multimodal cancer diagnosis using 3D-CNN for feature-level fusion. Springer Biomedical Engineering. https://doi.org/10.1007/s10916-023-00635-7

[9] Zhang, T., & Li, Y. (2023). A hybrid deep learning framework with decision-level fusion for breast cancer survival prediction. MDPI Cancers, 15(2), 245.

[10] Shen, Z., & Wu, P. (2023). Decision-level fusion using multimodal AI in lung cancer survival prediction. IEEE Transactions on Biomedical Engineering, 70(4), 1231–1245.

[11] Lee, H., & Zhang, X. (2023). Artificial intelligence-based multimodal molecular imaging fusion for cancer detection. Nature Biomedical Engineering.

[12] Rahman, M., & Davis, J. (2022). Multimodal medical data integration for early cancer detection. Journal of Medical Informatics, 29(3), 245–259. https://doi.org/10.1016/j.jmi.2022.03.004

[13] Shen, J., Zhao, Y., Huang, S., & Ren, Y. (2024). Secure and flexible privacy-preserving federated learning based on multi-key fully homomorphic encryption. Electronics, 13(22), 4478.

[14] Koskela, M., & Suojanen, P. (2023). Protecting data from all parties: Combining homomorphic encryption and differential privacy in federated learning. arXiv.

[15] Wu, X., & Zhang, H. (2023). Cross-center dataset testing for robust evaluation of multimodal AI systems in healthcare. IEEE Transactions on Biomedical Engineering, 70(4), 1311–1320.

[16] Zhao, P., & Li, M. (2022). Beyond traditional metrics: Multidimensional evaluation standards for multimodal AI in healthcare. Journal of Medical Imaging, 29(3), 213–224.

[17] Belt, S., & Granger, D. (2023). SHAP and LIME tools for explainable AI in healthcare: Improving transparency and trust. Communications of the ACM, 66(7), 68–77.

[18] Chen, W., & Zhou, L. (2023). Visualizing model decisions with Grad-CAM: Enhancing explainability in multimodal medical AI. Journal of Artificial Intelligence in Medicine, 50(2), 120–133. https://doi.org/10.1016/j.artmed.2023.01.004