Vol. 12 (2026): Proceedings of the 2026 International Conference on Life Sciences and Health Engineering (IC-LSHE 2026)
E-ISSN: 3105-0522 P-ISSN：3105-0514
DOI: https://doi.org/10.70267/lshe.20260204

The Application and Progress of Multimodal Models in the Medical Field

Qiyuan Zhang*

Zhongnan University of Economics and Law, Department of College of Information Engineering, Wuhan, China

**Corresponding author: Qiyuan Zhang*

Abstract

With the rapid development of artificial intelligence, especially generative models and multimodal large models, medical artificial intelligence has gradually moved from the early era of single-modal image recognition and text classification to the era of multimodal modeling. Medical data is inherently multimodal, including various modalities such as images, clinical texts, structured data, and genetic and signal information. For instance, in the diagnosis of lung diseases, multimodal models can integrate chest CT images with patients' electronic health records (EHR) and medical history texts to quickly locate the lesions and generate preliminary diagnostic suggestions; in the orthopedic treatment scenarios, the models can combine X-ray images with surgical record texts to assist doctors in formulating personalized surgical plans. How to effectively integrate these modalities, and conduct tasks such as diagnostic assistance, report generation, multi-round questioning, pathological explanation and reasoning, has been a focus of research in recent years. This paper systematically reviews the development path of medical multimodal models, summarizes the changes in the capabilities of mainstream methods and the limitations of datasets, and looks forward to the challenges and future trends for practical deployment in clinical settings.

Keywords

multimodal data fusion, healthcare, artificial intelligence, clinical implementation, cross-modal modeling

1. Introduction

In modern medical practice, digital diagnosis and treatment technologies have facilitated the rapid accumulation of massive heterogeneous clinical data, including imaging (such as CT, MRI, etc.), free text (medical records, consultation opinions), structured data (signs and test values in electronic health records), genomic information, and physiological signals (electrocardiograms, blood oxygen curves), etc. Each data modality carries unique clinical semantics. For instance, imaging reveals the shape of lesions, while genomic information indicates genetic characteristics. The combination of these two can overcome the limitations of individual data and provide support for precise diagnosis and personalized treatment, which is a key step in transitioning from empirical medicine to precision medicine.

However, cross-modal modeling and inter-modal reasoning have long been challenges for medical AI. Different modalities have significant differences in semantics, structure, and scale: images convey information through visual features, while texts describe diseases in natural language, and semantic systems are difficult to unify; structured data formats are well-organized, but images and texts have no fixed structure, and pre-processing methods are difficult to be universally applicable; genomic data reaches the GB level, while medical record texts are only the KB level, the disparity in data volume makes storage and computing difficult to coordinate, and these problems severely restrict the application of medical AI in cross-modal integration, and technological breakthroughs are needed to advance it.

JN Acosta pointed out that with the accumulation of biomedical data and the decline in sequencing costs, the foundation for the development of multimodal AI is gradually becoming complete. Multimodal AI has great application, challenges, and opportunities in the health field [1]. Mesko B focused on the attention brought by ChatGPT at the end of 2022 to large language models (LLMs) and the early single-modal limitations. Due to the medical multimodal characteristics, he explored the potential of LLMs to develop towards multimodality[2]. Michael Moor proposed Med-Flamingo, a multimodal few-shot learner applicable to the medical field, which solved the problem of data scarcity and took a crucial step towards the development of medical generative visual language models (VLMs) in the multi-modal direction[3]. Multimodal large language models (MLLMs) perform well in general multimodal tasks, but in the medical field, due to training costs, data, and other limitations, there are also difficulties in visual localization tasks. Jinlong He proposed the PFMVG model for this[4]. Fenglin Liu proposed a multimodal, multi-domain, and multilingual basic model suitable for solving rare or newly emerging diseases and non-English language problems[5].

“Multimodal Data Processing” as a research field can be traced back to the human-computer interaction (HCI) studies in the 1980s. After entering the medical field, the development of this area has gone through stages from early text-image matching and shallow feature fusion, to recent systematic improvements such as multimodal large language models (MLLMs), generative models (such as Diffusion), and general cross-modal agents (such as XGeM).

Although existing research has made significant progress in the technical path, there are still issues such as geographical limitations of data sets, insufficient model interpretability, and templateization of text generation. Based on this, this paper focuses on the development and application of medical multimodal models, using “fusion stage” as the core classification basis to sort out mainstream technical methods, systematically summarizing the performance characteristics, advantages and disadvantages, and data set limitations of each method, while analyzing the challenges faced in clinical implementation, and finally looking forward to future technical optimization directions, providing theoretical references for the transformation of medical multimodal models from “method innovation” to “clinical practicality”.

2. Overview of Mainstream Methods

2.1 Early Fusion

In the practice of multimodal data fusion in the medical field, early fusion is one of the core strategies, focusing on the initial stage of the data processing flow. Specifically, it involves systematic integration of heterogeneous modal data such as medical images, electronic medical records, clinical narratives, and physiological signals at the pre-stage of feature extraction and encoding. It constructs a multimodal fusion feature space to input the integrated data into a deep learning model for end-to-end learning. The core logic is to enable the model to fully explore the fine-grained correlations between cross-modal data at the early stage of information processing, accurately capture the complementary knowledge of different modalities, and provide comprehensive multimodal information support for clinical decision-making tasks such as disease diagnosis and prognosis prediction [6]. It directly concatenates the original features of different modalities (such as pixel values of images and word vectors of text) into a single vector and inputs it into a deep learning model (such as convolutional neural network CNN or fully connected network). Shared neural network layers (such as CNN or Transformer’s encoding layer) are used to uniformly encode the multimodal data and generate cross-modal feature representations.

2.2 Mid-Level Fusion

Mid-level Fusion refers to a strategy where features are independently extracted from each modality data to form an intermediate layer representation (such as high-order visual features of images, semantic vectors of text), and then fused through a cross-modal interaction mechanism. Its core lies in balancing modality specificity and correlation, supporting complex clinical tasks such as diagnostic assistance and report generation. The intermediate features of each modality are weighted and fused through an attention module (such as multi-head attention mechanism), highlighting key information. The features of different modalities are mapped to a unified semantic space (such as through fully connected layers or contrastive learning), and then fused.

2.3 Late Fusion

In the field of multimodal data fusion in healthcare, Late Fusion is a strategy where the fusion is carried out after each modality data has been independently processed and generates decision results. That is, models are first constructed for different modalities such as medical images and text medical records, and prediction results are output. Then, these results are integrated through methods such as weighted fusion and voting mechanisms to form the final decision. Its core lies in preserving the independent processing links of each modality data, reducing cross-modal interference, and simultaneously enhancing the decision robustness through the integration at the result level, providing comprehensive support based on independent judgments of multiple modalities for clinical tasks [7].

3. Challenges and Prospects

3.1 Challenges

The current application of multimodal models in the medical field has made certain progress in terms of technical paths and practical task empowerment, but still faces many global and deep-seated challenges.

3.1.1 Challenges at the Data Level

The challenges at the application level lie in the data level. The current layout and configuration of medical data are far from meeting the continuous optimization requirements of multimodal models. There are bottlenecks that urgently need to be overcome in aspects such as the balance of data sources, the completeness of sharing mechanisms, and the standardization of annotation systems. This largely restricts the further improvement and wide applicability of model performance. Although the medical field has abundant data, high-quality, large-scale, and multimodal correlated data sets are scarce. The scale of publicly available medical image data sets is much smaller than that of general data sets, and there is a significant difference in volume. The data volume and disease type distribution in different regions and medical institutions are uneven. For example, data is scarce in remote areas, and rare disease data is even more scarce, resulting in insufficient training data for model training, making it difficult to learn comprehensive features and limiting the generalization ability [8]. Multimodal data covers a large amount of sensitive information of patients, and the difficulty of privacy protection during data sharing and use is high. Once leaked, it will cause serious damage to patients' rights and interests. Moreover, multimodal data increases the risk of patient re-identification. Although privacy protection technologies are developing, they still face many challenges, such as the balance between data encryption and model training efficiency.

3.1.2 Challenges at the Model Level

In terms of the development dimension of the model itself, there is still considerable room for improvement in the core capabilities of multimodal models in medical scenarios. The interpretability issue of the model has always been a key bottleneck hindering its deep implementation in clinical practice. Medical decisions require high interpretability to gain clinical trust, but most current multimodal models operate in a "black box" manner, making it difficult to clearly explain the decision-making process and basis. When diagnosing diseases, doctors cannot understand why the model reaches a certain conclusion, making it difficult to form a reasoning logic that conforms to the rigorous requirements of the medical industry. This makes it difficult for the model to fully gain the widespread trust and recognition of medical professionals when assisting clinical decision-

making [9]. At the same time, in terms of the quality control of generated content, the model lacks stable and unified standards. Some generated results have a gap from clinical actual needs and cannot fully adapt to diverse medical application scenarios. The model's robustness is insufficient. The clinical environment is complex and variable, and data has noise and missing values. The existing models have poor adaptability to data changes and environmental interference. When encountering incomplete or interfered data cases, the model performance is likely to drop significantly, affecting the stability of diagnosis. For example, in emergency scenarios, some patient check data is missing, and the model's diagnostic accuracy is severely affected. The lack of deep multimodal fusion leads to insufficient exploration of complex and deep correlations between different modalities. In cases of inconsistent text and images, the accuracy of judgment is low. In the medical field, this is manifested as the inability to precisely integrate information when integrating data such as images, text, and genes, affecting the diagnostic accuracy [10].

3.1.3 Challenges at the Application Level

Furthermore, from the perspective of the overall application ecosystem, the integration of multimodal models with the existing medical system is not yet deep enough. There are many areas that require coordination and improvement, such as the connection with clinical work processes, the collaboration of technical applications among different medical institutions, and the compatibility of cross-scenario applications. The lack of industry standards, the absence of unified industry standards for multimodal data formats, fusion methods, and verification processes, and the difficulty in comparing and integrating models developed by different institutions all hinder the promotion and application of the technology. Each medical institution and research team develops models according to their own standards, resulting in poor interoperability between models and limiting multi-center research and large-scale clinical applications. The low clinical adaptability of the models makes it difficult to connect with existing hospital information systems (HIS), laboratory information systems (LIS), etc., and the integration with clinical actual work processes is low. This leads to inconvenience for doctors when using the models, increases the workload, and reduces the enthusiasm for applying the models. These problems collectively prevent the application effect of multimodal models in the medical field from being fully exerted and from realizing its inherent value potential.

3.2 Prospects

Looking ahead, the development prospects of multimodal models in the medical field are extremely promising, and they will undoubtedly play an increasingly important role in promoting the intelligent transformation of the medical industry. In terms of technological innovation, in the future, continuous in-depth exploration will be carried out to enhance the generalization ability of the models. By building a more comprehensive and diverse technical system, the limitations of current models in application scenarios and data dependence will be broken, enabling the models to better adapt to different medical environments and different disease types, and achieving wider clinical applicability.

In terms of optimizing the core capabilities of the model, interpretability and security will become the key research directions in the future. By continuously innovating technical methods and improving the reasoning mechanism and security guarantee system of the model, the model can not only possess efficient auxiliary diagnostic capabilities but also present the decision basis in a clearer and more understandable way, effectively enhancing the trust of medical workers in the model, and laying a solid foundation for the large-scale application of the model in clinical practice.

From the perspective of industry application, in the future, multimodal models will further integrate deeply with all aspects of the medical industry, playing a significant role in various fields such as disease diagnosis, treatment plan formulation, medical resource allocation, medical education and research. Through the collaborative development with the existing medical system, the service process will be continuously optimized, and the quality and efficiency of medical services will be improved, providing strong support for achieving higher-level medical and health services, and promoting the medical industry to move towards a more intelligent, precise and efficient direction.

4. Conclusion

This article focuses on the application and progress of multimodal models in the medical field, systematically reviewing their development path. Based on the "integration stage" as the core classification criterion, the mainstream methods are classified into three categories: early, mid-term, and late integration. The characteristics, advantages and disadvantages, as well as data set limitations of each method are analyzed. Research has shown that multimodal models can integrate various multimodal medical data such as images and text, and have demonstrated practical value in scenarios like lung disease diagnosis and orthopedic treatment. However, their clinical implementation faces multiple challenges at the levels of data, models, and applications. In the future, it is necessary to focus on the collaborative optimization of "data-model-application", to break through data bottlenecks, enhance model interpretability and security, and deepen integration with the medical system, in order to facilitate the development of the medical industry towards intelligence and precision, and provide technical support for improving medical service quality.

References

[1] Acosta, J. N., Falcone, G. J., Rajpurkar, P. and Topol, E. J. Multimodal biomedical AI. Nature Medicine. 2022, 28(9), pp. 1773-1784. https://doi.org/10.1038/s41591-022-01981-2.

[2] Meskó, B. The Impact of Multimodal Large Language Models on Health Care's Future. Journal of Medical Internet Research. 2023, 25, p. e52865. https://doi.org/10.2196/52865.

[3] Moor, M., Huang, Q., Wu, S., Yasunaga, M., Dalmia, Y., Leskovec, J., Zakka, C., Reis, E. P. and Rajpurkar, P., 2023. Med-Flamingo: a Multimodal Medical Few-shot Learner. *In:* Stefan, H., Antonio, P., Divya, S., et al. (eds.) *Proceedings of the 3rd Machine Learning for Health Symposium.* Proceedings of Machine Learning Research: PMLR.

[4] He, J., Li, P., Liu, G. and Zhong, S. Parameter-Efficient Fine-Tuning Medical Multimodal Large Language Models for Medical Visual Grounding. In 2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI), Houston, TX, 2025; pp. 1-5. https://doi.org/10.1109/ISBI60581.2025.10981029.

[5] Liu, F., Li, Z., Yin, Q., Huang, J., Luo, J., Thakur, A., Branson, K., Schwab, P., Yin, B., Wu, X., et al. A multimodal multidomain multilingual medical foundation model for zero shot clinical diagnosis. npj Digital Medicine. 2025, 8(1), p. 86. https://doi.org/10.1038/s41746-024-01339-7.

[6] Gadzicki, K., Khamsehashari, R. and Zetzsche, C. Early vs Late Fusion in Multimodal Convolutional Neural Networks. In 2020 IEEE 23rd International Conference on Information Fusion (FUSION), Rustenburg, South Africa, 2020; pp. 1-6. https://doi.org/10.23919/FUSION45008.2020.9190246.

[7] Escalante, H. J., Hérnadez, C. A., Sucar, L. E. and Montes, M., 2008. Late fusion of heterogeneous methods for multimedia image retrieval. *Proceedings of the 1st ACM international conference on Multimedia information retrieval.* Vancouver, British Columbia, Canada: Association for Computing Machinery.

[8] Liu, F., Zhu, T., Wu, X., Yang, B., You, C., Wang, C., Lu, L., Liu, Z., Zheng, Y., Sun, X., et al. A medical multimodal large language model for future pandemics. npj Digital Medicine. 2023, 6(1), p. 226. https://doi.org/10.1038/s41746-023-00952-2.

[9] AlSaad, R., Abd-alrazaq, A., Boughorbel, S., Ahmed, A., Renault, M.-A., Damseh, R. and Sheikh, J. Multimodal Large Language Models in Health Care: Applications, Challenges, and Future Outlook. Journal of Medical Internet Research. 2024, 26, p. e59505. https://doi.org/10.2196/59505.

[10] Sun, K., Xue, S., Sun, F., Sun, H., Luo, Y., Wang, L., Wang, S., Guo, N., Liu, L., Zhao, T., et al. Medical multimodal foundation models in clinical diagnosis and treatment: Applications, challenges, and future directions. Artificial Intelligence in Medicine. 2025, 170, p. 103265. https://doi.org/10.1016/j.artmed.2025.103265.

Funding
This research received no external funding.

Conflicts of Interest
The authors declare no conflict of interest.

Acknowledgment

This paper is an output of the science project.