International Journal of Scientific Research & Engineering Trends
Volume 12, Issue 5, Sep-Oct-2026, ISSN (Online): 2395-566X

Multimodal Ai Integrating Text, Images, Audio And Video For Intelligent Healthcare Diagnosis: A Review Of Fusion Methods And Foundation Models

Khot Shruti Rahul[1] Prof. M. C. Kolse[2] Prof. S. K. Said[3]
Department Of Artificial Intelligence And Data Science Jaihind College Of Engineering Pune, Maharashtra, India

Abstract—Clinical diagnosis rarely relies on one source of information: doctors combine medical images, symptoms described by patients, clinical reports, laboratory results, electronic health records, and sometimes audio or video such as heart sounds or gait recordings. Multimodal Artificial Intelligence integrates such heterogeneous data to improve the accuracy, efficiency and reliability of diagnosis. This paper reviews multimodal AI for intelligent healthcare diagnosis assistance. It introduces the taxonomy of multimodal machine learning, covering representation, alignment and fusion, and compares early, intermediate and late fusion strategies. Applications are reviewed in disease detection, image interpretation with reports, personalized treatment and patient monitoring, together with vision–language models that learn from paired images and text and generalist medical foundation models that accept many input types. Public multimodal datasets are discussed. A five-layer architecture comprising multimodal data acquisition, modality-specific encoders, fusion, diagnostic reasoning and clinician-facing explanation is proposed, with research gaps, objectives, a comparison of fusion strategies and challenges including missing modalities, data alignment, privacy, interpretability and validation. The review concludes that multimodal AI can provide more comprehensive patient analysis and faster decisions when it is rigorously validated and clinically supervised.

Index Terms—Multimodal AI, data fusion, healthcare diagnosis, vision–language models, medical foundation models, electronic health records, deep learning

I. INTRODUCTION

A radiologist reading a chest X-ray benefits from knowing the patient's symptoms and history, and a cardiologist interprets an ECG alongside blood tests and examination findings. Most medical AI systems, however, analyse a single type of data, such as an image or a set of laboratory values, and therefore miss information that clinicians routinely combine.

Multimodal AI addresses this limitation. Acosta *et al.* [1] reviewed multimodal biomedical AI and argued that integrating imaging, genomics, clinical records and sensor data could enable more precise diagnosis and personalized care.

Recent generalist medical models [2] go further, accepting many types of input and producing free-text explanations.This paper reviews multimodal AI for healthcare diagnosis, proposes an architecture and discusses challenges.

II. RELATED WORK

Huang *et al.* [3] systematically reviewed deep learning studies that fuse medical images with electronic health records and found that fused models generally outperformed single-modality models; they proposed implementation guidelines for early, joint and late fusion.

Vision–language pre-training has transformed the field. CLIP [4] learned visual concepts from image–text pairs, and Wang *et al.* [5] adapted contrastive learning to medical images and reports in MedCLIP, handling unpaired data efficiently. Johnson *et al.* [6] released MIMIC-CXR, chest radiographs paired with free-text reports, enabling such training.

General vision–language models including Flamingo [7], BLIP-2 [8] and LLaVA [9] combine image encoders with language models for question answering about images, and medical variants follow the same pattern. Moor *et al.* [2] proposed generalist medical AI foundation models, and Tu *et al.* [10] demonstrated a single model handling clinical language, imaging and genomics tasks.

For audio and video, multimodal affective and behavioural analysis [11] shows that combining modalities improves robustness, a principle applied to heart-sound, cough and gait analysis. Table I summarizes these works.

© 2026 IJSRET

TABLE I Summary of Literature Survey

Ref.	Author(s) / Year	Technique / Approach	Key Contribution	Limitation / Gap
[12]	Baltrušaitis et al., 2019	Multimodal ML taxonomy	Five core challenges defined	General survey
[3]	Huang et al., 2020	Review of image–EHR fusion	Fusion outperforms single modality	Few prospective studies
[1]	Acosta et al., 2022	Multimodal biomedical AI review	Roadmap for integrated care	Data access barriers
[6]	Johnson et al., 2019	MIMIC-CXR dataset	Paired X-rays and reports	Single institution
[5]	Wang et al., 2022	MedCLIP contrastive learning	Efficient medical image–text learning	Report quality dependent
[9]	Liu et al., 2023	LLaVA visual instruction tuning	Image question answering	General domain
[2]	Moor et al., 2023	Generalist medical AI	Flexible multimodal medical models	Conceptual, safety open
[10]	Tu et al., 2024	Generalist biomedical model	One model, many modalities	Not clinically deployed

2.1 Preliminaries

a. Modalities In Healthcare

Fig. 1 summarizes healthcare modalities (images, text, structured records, audio and video, physiological signals) and the fusion strategies used to combine them.

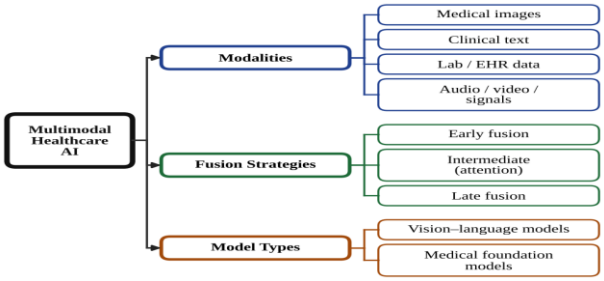


Fig. 1. Healthcare modalities and fusion strategies.

2.2 Multimodal Machine Learning

Baltrušaitis *et al.* [12] identified five core challenges: representation, translation, alignment, fusion and co-learning. In healthcare, fusion and alignment are especially important because data arrive at different times and in different formats.

2.3 Problem Statement And Objectives

a. Identified Research Gaps

- Real patients often have missing modalities, which many models cannot handle.
- Aligning data collected at different times and resolutions is difficult.
- Large paired multimodal datasets from Indian hospitals are scarce.
- Explanations that show how each modality contributed to a decision are limited.
- Generalist models need extensive safety evaluation before clinical use.

2.4 Objectives Of The Review

- To review multimodal machine learning concepts and fusion strategies.
- To examine multimodal applications and foundation models in healthcare.
- To propose a multimodal diagnosis assistance architecture.
- To compare early, intermediate and late fusion.
- To identify challenges and future research directions.

2.5 Proposed Framework

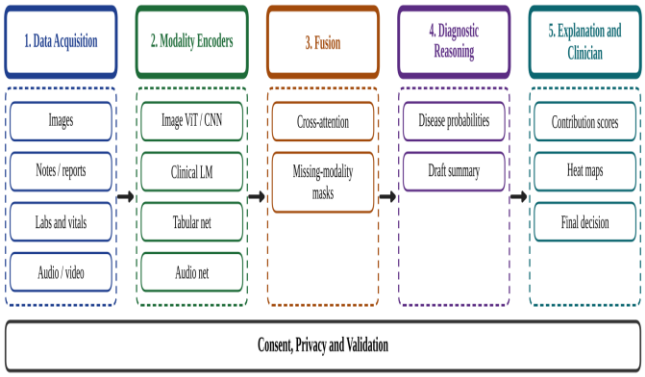


Fig. 2 shows the proposed architecture.

© 2026 IJSRET

Fig. 2. Proposed multimodal diagnosis assistance architecture.

2.6 Data Acquisition Layer

Images, clinical notes, laboratory values, vital signs and optional audio or video are collected with patient consent and linked by encounter.

2.7 Modality Encoder Layer

A CNN or vision transformer encodes images, a clinical language model encodes text, a tabular network encodes structured data and spectrogram networks encode audio.

2.8 Fusion Layer

Cross-attention fuses modality embeddings into a joint representation; missing modalities are handled by masking.

2.9 Diagnostic Reasoning Layer

Classifiers estimate disease probabilities, and a language model generates a draft summary grounded in the inputs.

2.10 Explanation And Clinician Layer

Clinicians see predictions, modality contribution scores, image heat maps and highlighted text, and make the final decision.

III. WORKING METHODOLOGY

Fig. 3 describes the workflow. In intermediate fusion with attention, the joint representation is a weighted combination of modality embeddings

$$z = \Sigma_m \alpha_m \cdot h_m \quad (1)$$

where h_m is the embedding of modality m and α_m its learned importance for the current patient. Contrastive vision–language pre-training aligns image and text embeddings by maximizing similarity of matched pairs relative to mismatched pairs.

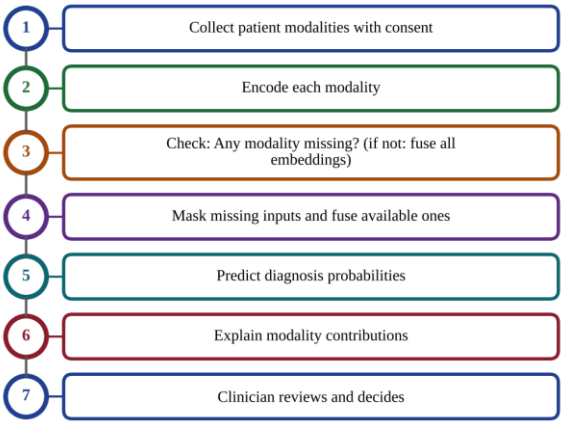


Fig. 3. Multimodal diagnosis workflow.

Models are evaluated with AUC, sensitivity and specificity, and with ablation studies that remove each modality to measure its contribution.

IV. DISCUSSION

4.1 Comparison Of Approaches

Table II compares fusion strategies. Early fusion captures low-level interactions but requires aligned inputs, late fusion is robust to missing modalities, and intermediate attention-based fusion offers the best balance.

TABLE II Comparison of Fusion Strategies

Strategy	Interaction Capture	Missing-Data Robustness	Complexity
Early fusion	High	Low	Low
Intermediate (attention)	High	Medium–High	Medium
Late fusion	Low	High	Low
Foundation model	Very high	Medium	Very high

4.2 Open Challenges And Future Work

Table III summarizes challenges. Future directions include robust handling of missing modalities, federated multimodal learning across hospitals, wearable data integration and regulatory evaluation of generalist medical models.

TABLE III Challenges and Mitigation

Challenge	Mitigation
Missing modalities	Masking; modality dropout training
Data alignment	Time-aware encoders; encounter linking
Privacy	Federated learning; de-identification
Interpretability	Modality attribution; heat maps
Validation	External, prospective clinical studies

V. CONCLUSION AND FUTURE WORK

Multimodal AI brings machines closer to how clinicians think, by combining images, text, structured data and signals. The reviewed literature shows that fusion improves diagnostic performance and that vision–language and generalist foundation models open new possibilities. Addressing missing data, alignment, interpretability, privacy and validation will be essential for safe clinical adoption.

© 2026 IJSRET

5.1 Acknowledgment

The authors express their sincere gratitude to the Department of Artificial Intelligence and Data Science, Jaihind College of Engineering, Pune, for providing the facilities, guidance and encouragement needed to carry out this work. The authors also thank the faculty members and colleagues whose valuable suggestions helped improve the quality of this review.

REFERENCES

1. J. N. Acosta, G. J. Falcone, P. Rajpurkar, and E. J. Topol, "Multimodal biomedical AI," Nat. Med., vol. 28, pp. 1773–1784, 2022.
2. M. Moor et al., "Foundation models for generalist medical artificial intelligence," Nature, vol. 616, pp. 259–265, 2023.
3. S.-C. Huang, A. Pareek, S. Seyyedi, I. Banerjee, and M. P. Lungren, "Fusion of medical imaging and electronic health records using deep learning: A systematic review and implementation guidelines," npj Digit. Med., vol. 3, Art. no. 136, 2020.
4. A. Radford et al., "Learning transferable visual models from natural language supervision," in Proc. ICML, 2021, pp. 8748–8763.
5. Z. Wang, Z. Wu, D. Agarwal, and J. Sun, "MedCLIP: Contrastive learning from unpaired medical images and text," in Proc. EMNLP, 2022, pp. 3876–3887.
6. A. E. W. Johnson et al., "MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports," Sci. Data, vol. 6, Art. no. 317, 2019.
7. J.-B. Alayrac et al., "Flamingo: A visual language model for few-shot learning," in Proc. NeurIPS, vol. 35, 2022, pp. 23716–23736.
8. J. Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in Proc. ICML, 2023, pp. 19730–19742.
9. H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in Proc. NeurIPS, vol. 36, 2023.
10. T. Tu et al., "Towards generalist biomedical AI," NEJM AI, vol. 1, no. 3, 2024.
11. S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," Inf. Fusion, vol. 37, pp. 98–125, 2017.
12. T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," IEEE Trans. Pattern Anal. Mach. Intell., vol. 41, no. 2, pp. 423–443, 2019.

© 2026 IJSRET