

# Multimodal artificial intelligence models for radiology

Amara Tariq, PhD<sup>1</sup>, Imon Banerjee , PhD<sup>1</sup>, Hari Trivedi, MD<sup>2</sup>, Judy Gichoya , MD, MS<sup>\*,2</sup>

<sup>1</sup>Mayo Clinic, Phoenix, AZ, 85054, United States

<sup>2</sup>Department of Radiology and Imaging Sciences, Emory University, Atlanta, GA 30322, United States

\*Corresponding author: Judy Gichoya, MD, MS, Department of Radiology and Imaging Sciences, Emory University, 1364 Clifton Rd NE, Suite AG08, Atlanta, GA 30322, United States (judywawira@emory.edu)

## Abstract

Artificial intelligence (AI) models in medicine often fall short in real-world deployment due to inability to incorporate multiple data modalities in their decision-making process as clinicians do. Clinicians integrate evidence and signals from multiple data sources like radiology images, patient clinical status as recorded in electronic health records, consultations from fellow providers, and even subtle clues using the appearance of a patient, when making decisions about diagnosis or treatment. To bridge this gap, significant research effort has focused on building fusion models capable of harnessing multi-modal data for advanced decision making. We present a broad overview of the landscape of research in multimodal AI for radiology covering a wide variety of approaches from traditional fusion modelling to modern vision-language models. We provide analysis of comparative merits and drawbacks of each approach to assist future research and highlight ethical consideration in developing multimodal AI. In practice, the quality and quantity of available training data, availability of computational resources, and clinical application dictates which fusion method may be most suitable.

**Keywords:** multimodal AI; AI in radiology; vision-language models for radiology.

## Introduction

Multimodal data analysis is a routine part of clinical decision making. Clinicians often review multiple streams of heterogeneous data elements including imaging, labs, clinical notes, physical examination findings, medications, and comorbidities to make diagnostic or prognostic decisions. Early efforts in deep learning models in healthcare typically focused on single modalities—for example, imaging, labs, or a finite set of features from the electronic health record (EHR). Most radiology models today are developed using convolutional neural networks (CNNs) applied to a single image type like chest X-rays or head CTs.<sup>1,2</sup> Even basic elements like patient demographics are rarely included in the model pipeline. While image-only models can perform well on image-only tasks such as haemorrhage detection, there remains a fundamental gap for artificial intelligence (AI) models that operate on imaging alone to outperform physicians in clinical decision making as they don't have the same information available. When models are developed without including pertinent clinical information, they may lead to non-specific or inappropriate conclusions that limit their utility. For example, a clinically deployed model for intracranial haemorrhage at our institution notifies radiologists multiple times daily about acute haemorrhages for in-patients that are postoperative, for whom intracranial blood products are expected thus unnecessarily disrupting the radiologist workflow. Additionally, we observed a 10% performance drop in outpatients in a pulmonary embolism triage model compared to the reported FDA validation metrics. This is concerning as the outpatients are the most vulnerable when they have critical findings as they cannot receive timely intervention. Building single modality models without clinical context (available from multimodal

data) ultimately results in impractical models with limited clinical utility.

Advancements in computational methods have made it possible to merge clinical and imaging data when training models, as well as incorporate multiple imaging modalities into the same model (eg, combining mammograms with ultrasound and MRI for breast cancer screening<sup>3</sup>) These may even be combined with digital histopathology slides to predict treatment response and post-diagnosis prognostication.<sup>4-6</sup> The interest in multi-modal fusion has resulted in many targeted funding opportunities and release of many multimodal large datasets like INSPECT,<sup>7</sup> ROCO,<sup>8</sup> All of Us,<sup>9</sup> and UK Biobank.<sup>10</sup> Many new multi-modal fusion modelling techniques have also been developed that combine radiology images along with clinical information from the EHR to make diagnoses or predictions.<sup>11-15</sup> This survey paper reviews various computational approaches that can be applied to multimodal datasets in radiology based on model architecture and discusses them from technical and application perspectives including relevant ethical concerns.

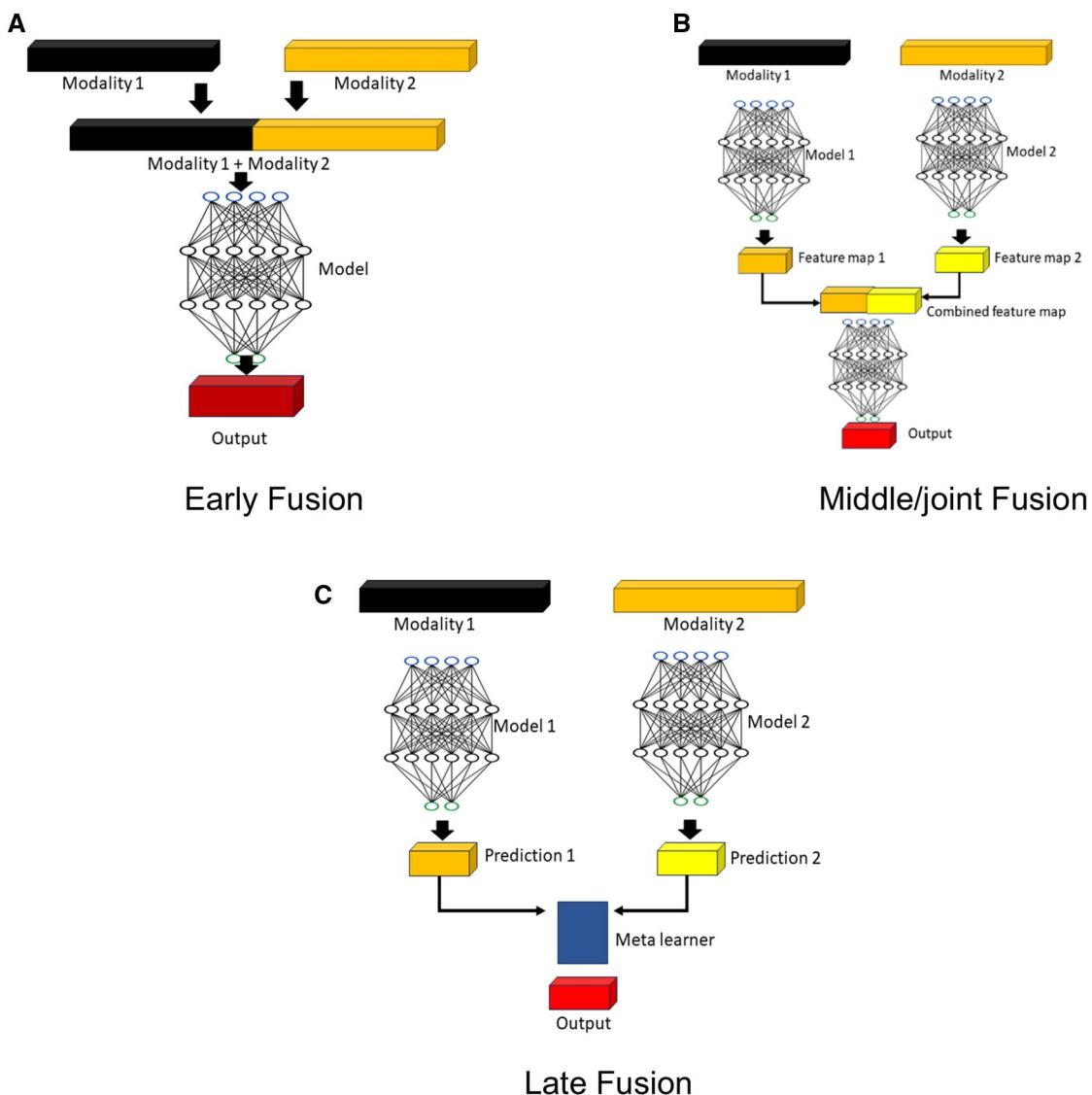
## Traditional fusion models

The earliest form of fusion models in machine learning combined tabular data from multiple sources (eg, lab test results, demographics), transformed or harmonized them, and trained an end-to-end model. These models are straightforward in terms of implementation and are explainable in terms of model-assigned importance to input features. However, introduction of complex, unstructured, and high dimensional data modalities like images and free text has motivated advanced research that combines straightforward fusion frameworks with complex pretrained feature extraction pipelines (Figure 1).<sup>11</sup>

Received: 26 May 2024; Revised: 1 November 2024; Accepted: 13 November 2024

© The Author(s) 2025. Published by Oxford University Press on behalf of the British Institute of Radiology.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.



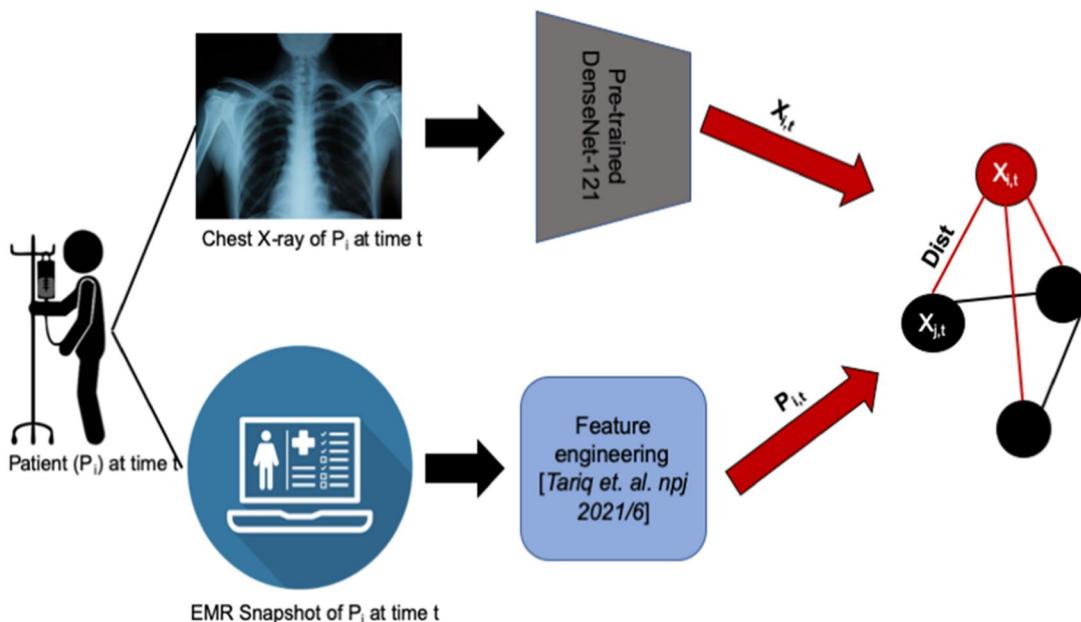
**Figure 1.** Traditional fusion models—(A) early fusion, (B) joint/middle fusion, and (C) late fusion. These architectures differ based on the point of merging multiple features early, middle, or late in the model pipeline.

*Late fusion* is a relatively simple framework in which scores/probabilities generated by different models are parsed via a meta-learner to calculate a weighted combination of the decision as summarized in Figure 1C. Each model is trained in isolation on different modalities and since only target probabilities are “fused,” no architectural innovation is needed to fuse different modalities like images and tabular data. However, this technique is limited in that the fusion model cannot learn complementary information from different modalities as the features or probabilities are already extracted and frozen from each modality *before* fusion. Although limited compared to more advanced methods, many models have successfully used this approach to combine radiology images like chest X-rays with clinical data, achieving greater performance compared to individual modality models.<sup>16,17</sup>

In contrast, *Early fusion* strategies fuse raw data or extracted features (often by straightforward concatenation) early in model training and then pass the combined features through a model with loss propagated back up to the fusion layer.<sup>11</sup> While this approach allows the model to extract

complementary information from 2 or more modalities, the heterogeneous nature of data, for example, images and tabular EHRs typically requires features to be extracted from unstructured data (images and text) instead of raw data before they can be fused. This suggests that the quality of the pre-trained feature extractor can significantly influence the performance of the final classification model, highlighting the necessity of careful selection of the feature extractor, and potential need to experiment with multiple variations of feature extraction. For example, ImageNet-trained CNN models are not ideal feature extractors for chest X-rays due to differences between natural images (ImageNet dataset) and chest X-rays. Instead features extracted from pipelines trained or fine-tuned on medical domain images have demonstrated better model performance.<sup>18</sup>

Finally, *joint fusion* combines components of early and late fusion and allows end-to-end training of models (ie, backpropagation of loss to the input layer), however, with independent parallel feature extractors that derive relevant features from each modality that are then combined using linear or nonlinear functions.<sup>11,12</sup> In joint fusion, the feature



**Figure 2.** Graph convolutional neural network adaptation to predict COVID outcomes using multimodal data—chest X-ray and tabular EHR. The image is adapted from Tariq et al.<sup>8</sup>

extraction backbone can be updated to better suit the downstream prediction task.

A limitation of all traditional fusion approaches is the inability to handle missing data, i.e., all modalities must be available during both training and inference. Early and joint fusion models may also overfit due to the high dimensionality of the fused feature sets and often underperform compared to the baseline single modality models. In addition, developers use feature selection methods and manual curation based on domain knowledge to overcome limitations of joint learning opportunity in late fusion models, however, this can limit widespread deployment of developed models where the same curation techniques may not apply. These limitations have led to the development of innovative fusion methods with capacity to integrate both explicit and implicit information with minimal human curation.

### Graph-based fusion models

While traditional fusion paradigms result in rigid models with restrictions on features that can be processed, graph convolutional neural networks (GCNs) provide an opportunity to fuse implicit information about clinical similarity between samples or patients with information extracted from multiple clinical modalities like images, text, or tabular data. This architecture has proven to be effective for several tasks where instances can form meaningful and informative “neighbourhoods.”<sup>14,19-22</sup> For example, information on Alzheimer disease progression is present both in brain MRIs and in the demographics and clinical history of a patient.<sup>22</sup> Instead of fusing demographic information directly with features extracted from brain images under an early or late fusion paradigm, graph-based models can learn the relationship between brain images of a specific patient (node feature vector) and brain images of other clinically similar patients (feature vectors of other connected nodes) as defined by neighbourhoods in the graph structure that connects patients with similar demographic characteristics (edge feature

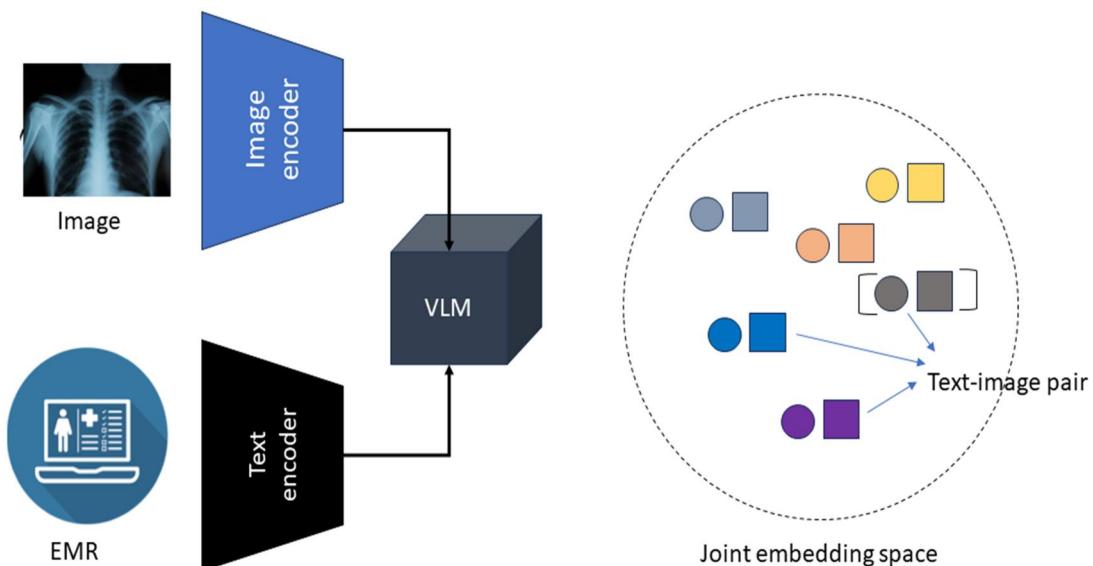
vectors). This approach has proven its merit in several use cases including diagnosis and clinical event prediction.<sup>14</sup>

GCNs have been adapted for a wide range of modalities including textual, tabular, and imaging data, indicating their modality-agnostic benefits. They also demonstrate better generalizability capabilities for missing data, especially if the missing data was only used for graph structure formation.<sup>23,24</sup> A trained GCN model expects a graph structure as input and does not place any restriction on the graph formation process. Hence, any missing features can be omitted from the graph formation process during external validation while leaving the pretrained GCN applicable. GCNs have also been modified to include temporal information by including temporal convolutional layers.<sup>19</sup> Such models allow incorporation of time-varying clinical information, like laboratory values or vital signs (Figure 2). For example, a temporal model was used to build edges of the graph such that patients (nodes) with similarly evolving clinical status were connected and used to predict the risk of blood transfusion for each patient.<sup>23</sup> To demonstrate the variable performance of fusion architectures, we evaluated GCN and traditional fusion for 2 different tasks; (a) Discharge from hospital and mortality prediction for COVID-19 patients by fusing chest X-rays and procedure and billing codes; (b) transfusion risk prediction for in-patients by fusing acquired features (laboratory test results and vital signs) and derived clinical features (billing and procedure codes and prescribed medications) [17]. These experiments were externally validated in addition to evaluation on internal held-out test sets. For the first task, models were trained using cohort of Emory University Hospital patients and externally validated on a cohort from Mayo Clinic. For the second task, the model was trained using patient cohort from Mayo Clinic and externally validated on the publicly available MIMIC-III dataset – see result summary in Table 1.

Although GCNs are powerful, unstructured modalities like images and text may still need pretrained feature extractors for generating node feature vectors. Careful selection of the

**Table 1.** Performance comparison between GCN and traditional fusion approach of late fusion across 2 different use cases—COVID-19 discharge and mortality prediction; and blood transfusion risk.

Model	Internal			External		
	Sensitivity	Specificity	AUROC	Sensitivity	Specificity	AUROC
<b>Hospital discharge prediction</b>						
Late fusion	69.7 [68.4-70.8]	65.5 [64.1-67.2]	74.5 [73.4-75.6]	51.2 [50.4-51.9]	61.3 [60.0-62.7]	58.1 [57.2-59.0]
GCN	71.1 [69.9-72.3]	69.6 [68.2-71.4]	77.1 [76.1-78.2]	64.8 [64.1-65.4]	57.4 [55.9-58.8]	64.6 [63.7-65.5]
<b>Mortality prediction</b>						
Late fusion	85.6 [83.0-88.5]	81.1 [79.8-82.8]	88.6 [87.2-90.2]	76.0 [74.0-78.3]	82.2 [81.1-83.4]	81.6 [80.3-82.9]
GCN	84.7 [82.3-87.9]	82.5 [81.1-83.9]	90.1 [89.0-91.3]	81.4 [79.6-83.5]	74.6 [73.2-76.0]	85.3 [84.2-86.5]
<b>Transfusion risk prediction</b>						
Late fusion	64.0 [62.1-66.0]	64.2 [62.9-65.5]	69.9 [68.6-71.2]	60.0 [50.0-75.0]	54.0 [50.8-57.9]	44.4 [31.3-56.0]
GCN	73.8 [72.0-75.6]	65.4 [64.1-66.7]	77.4 [76.3-78.5]	80.0 [66.0-95.0]	69.8 [66.7-73.6]	70.8 [62.5-84.7]



**Figure 3.** Joint embedding space generation using vision-language model (VLM)—image and text.

feature extraction pipeline is still required for traditional fusion architecture. GCN architectures may also suffer from a “homogenization” effect if too many graph convolutional layers are included in the model. For example, imaging features extracted from pretrained CNN pipelines for different chest X-rays have similar characteristics as they share overall image structures such as a dark background and human-shaped foreground object. When used as node features, the first layer of graph convolution will process one patient’s chest X-ray (node feature vectors) along with chest X-rays of its directly connected neighbours, often through weighted combinations. A second layer will involve node feature vectors of *neighbours of neighbours* of the node. As additional graph convolution layers are added, the processed forms of all chest X-rays (node features) will become increasingly similar and the GCN will lose discriminatory ability for differences between the images (eg, disease states).

**Model explainability** is another challenge of graph-based models. While feature weights can be used as explanations in early or late fusion models, GCN-based models require 2 forms of explanation—explanation of important node features and importance of graph neighbourhood or subgraphs. Explanatory frameworks like GNNExplainer have been proposed which randomly sample graph neighbourhoods and

estimate their effect on model output, thus assigning importance weights to graph edges.<sup>15,25</sup> However, clinical applicability of such explanations is still under-explored due to their complexity.

#### Joint embedding of multimodal data

In recent years, large transformer models which are trained under self-supervision have revolutionized the fields of natural language and image processing. Large models (billions of trainable parameters) and training data set sizes (billions of word tokens) enable these models to show “emergent” behaviour, ie, they can perform task they were not explicitly trained for, ie, zero-shot performance. These models are known for their utility as zero-shot or few-shot learners for a wide variety of text-related tasks such as language generation and vision tasks such as image-to-image translation.<sup>26-28</sup> Recently developed vision-language models (VLM) have shown surprisingly good performance on a variety of image and text-related downstream tasks like radiology report generation and visual question answering.<sup>29-31</sup> Research in this field is rapidly moving towards multi-modal large models such as VLM that have joint embedding spaces to process visual and textual data to learn text-image pair interdependence (Figure 3). VLMs are generally composed of 2

encoders—one for each modality—and layers that jointly process both image and text features. MedCLIP and MedViLL are examples of joint encoder architectures.<sup>29,30</sup> Transformer-based architectures have homogenized processes of both text and images as sequences of tokens (text) and image patches (images). For example, Vision and language transformer architecture (ViLT) uses 1 transformer-based encoder to encode an image-text pair.<sup>32</sup> While MedCLIP was trained for the task of image-text matching using a contrastive loss, training of recent VLMs like ViLT included tasks like masked word prediction, image patch-word alignment, and image-text retrieval.

The primary challenge for training VLMs is that the inherently large size of these models requires huge training datasets—in the range of million text-image pairs. Examples of open-source datasets used to train these models are Public Multimodal Dataset (PMD)<sup>33</sup> which contains about 70M image-text pairs, and several smaller datasets including Conceptual Captions,<sup>34</sup> WIT,<sup>35</sup> Localized Narratives,<sup>36</sup> RedCaps,<sup>37</sup> COCO.<sup>38</sup> Developing annotated datasets at this scale in the healthcare domain is challenging which makes training VLMs for this domain difficult. In addition, benchmark datasets to robustly evaluate VLM performance are missing. Researchers are tackling this problem by combining several imaging and image-text datasets from clinical domains including Open-I,<sup>39</sup> MIMIC-CXR,<sup>40</sup> and VQA-RAD.<sup>41</sup> MIMIC-CXR and VQA-RAD include image-text pairs from various imaging modalities (X-ray, CT, MR) and anatomical sites and have been used to train medical domain VLMs like MedViLL.<sup>30,32</sup>

As the availability of radiology datasets and computational resources improves, VLMs are increasingly being applied for several radiology tasks. The inherently self-supervised nature of VLM training where a model is optimized to learn correlation between image-text pairs, for example, radiology image and report pairs, is particularly suitable for radiology where the cost of manual annotation of images remains expensive. While VLMs were initially applied to 2-dimensional radiology images like chest X-ray,<sup>29</sup> they are now being increasingly used for 3-dimensional radiology volumes like CT.<sup>42,43</sup>

### Ethical considerations/pitfalls of multimodal models in radiology

Multimodal models exhibit similar limitations to single modality models including limited explainability, poor generalizability when encountering out-of-distribution datasets (especially for traditional fusion models), and limited interpretability. Additional concerns arise when applying multimodal models. Review of several radiology and pathology models demonstrates that most of these models rely on clinical trial datasets which tend to be very structured and have a very strict exclusion criterion.<sup>44</sup> This is because the radiology study must be matched to the pathology slide, and given varied appearance it's impossible to generate overlays and image registrations. Therefore, it is unlikely that these types of models that are generated from clinical trial databases will

demonstrate robustness when they encounter out-of-distribution data in a real-world setting.<sup>44</sup> Another challenge of these models arises in the feature engineering which can inject human biases in what is considered important. Feature extraction pipelines, for example, with radiomics may lack robustness and standardization making it difficult to reproduce in datasets with variable acquisition parameters.<sup>45</sup> Computational limitations may mandate to focus on the findings with high occurrences, for example, with radiogenomics—and this ignores the contribution of rare occurring cell types.

At baseline, deep learning models are challenging to interpret and explain due to the millions of model parameters. For multimodal models, traditional approaches may offer better explainability compared to VLMs. It is not clear if the large technical and data burden of multimodal data fusion outperforms single modality models or those that use simple regression approaches, and more research is necessary to fill this knowledge gap. The ability of deep learning models to encode “hidden characteristics” on images like self-reported race of patients is challenging as we combine different image types.<sup>46</sup> Finally, a challenge of the use of multimodal models for developing biomarkers<sup>47</sup> has shown that group differences are reported with demographic subgroups of race—a social and legal construct, that is captured differently across many care settings and is forbidden in some geographic regions. There is need to develop more subgroups from these multimodal data beyond Black and white patients—as there is a risk of strengthening historical biases that attempt to show different biological occurrences across different race groups.

### Conclusion

Large multimodal models like VLMs are pushing the boundaries of AI in radiology, but caution should be exercised to select the appropriate fusion approach for a task due to challenges of handling high-dimensional heterogeneous data in an end-to-end model. While existing pretrained VLMs are suitable for natural image tasks, careful finetuning may be required before their use in the healthcare domain. Graph-based fusion models present an alternative when limited training data are available and can also circumvent some of the challenges of missing data and model generalization. However, graph-based models have challenges with explainability. In cases where high-quality but limited multimodal data are available, traditional machine learning-based fusion models may be computationally efficient and explainable. Fusion models combining information from multiple sources are critical to delivering precision medicine by incorporating data from different modalities including imaging, text, tabular, and genomic sequences. As computational approaches advance, there will be several radiology toolkits available for harnessing multimodal data, as well as ethical pitfalls that must be addressed to ensure benefit for all patients. Table 2 summarises the pros and cons of various fusion strategies.

**Table 2.** Summary of pros and cons of various fusion strategies.

Approach	Pros	Cons
Traditional—early fusion	<ul style="list-style-type: none"> <li>• Explainable</li> <li>• Ability to extract complimentary information in the various modalities</li> </ul>	<ul style="list-style-type: none"> <li>• Must extract features from heterogeneous data (instead of using the raw data)—hence can be affected by the quality of feature extractor</li> <li>• Inability to handle missing data</li> <li>• Risk of overfitting due to high feature dimensionality</li> <li>• Human-based feature curation and engineering can be biased and is time-intensive</li> </ul>
Traditional—middle/joint fusion	<ul style="list-style-type: none"> <li>• Architectural innovation to support parallel extract features extractors. Backbone can be updated for better performance on the downstream task.</li> </ul>	<ul style="list-style-type: none"> <li>• Inability to handle missing data</li> <li>• Risk of overfitting due to high feature dimensionality</li> <li>• Human-based feature curation and engineering can be biased and is time-intensive</li> </ul>
Traditional—late fusion	<ul style="list-style-type: none"> <li>• Explainable</li> <li>• No architectural innovation to extract features</li> </ul>	<ul style="list-style-type: none"> <li>• Unable to learn complimentary information from different modalities</li> <li>• Inability to handle missing data</li> <li>• Human-based feature curation and engineering can be biased and is time-intensive</li> </ul>
Graph neural networks	<ul style="list-style-type: none"> <li>• Minimal curation required</li> <li>• Can handle missing data</li> <li>• Robust generalizability to new datasets and missing data</li> </ul>	<ul style="list-style-type: none"> <li>• Unstructured data may still need pretrained feature extractors</li> <li>• Homogenization effect may occur when too many graph layers are included in the model</li> <li>• Not explainable</li> </ul>
Joint embedding of multimodal data	<ul style="list-style-type: none"> <li>• Require minimal dataset curation</li> </ul>	<ul style="list-style-type: none"> <li>• Require large training datasets and compute</li> <li>• No benchmark datasets for robust VLM performance evaluation</li> </ul>

## Funding

J.G. is a 2022 Robert Wood Johnson Foundation Harold Amos Medical Faculty Development Program and declares support from Radiology Society of North America Health Disparities grant (#EIID2204), Lacuna Fund (#67), Gordon and Betty Moore Foundation, NIH Medical Imaging and Data Resource Center (MIDRC) grant under contracts 75N92020C00008 and 75N92020C00021, and National Heart, Blood and Lung Institute (NHLBI) Award Number R01HL167811.

## Conflicts of interest

None declared.

## References

1. Abdou MA. Literature review: efficient deep neural networks techniques for medical image analysis. *Neural Comput Appl.* 2022;34(8):5791-5812.
2. Anwar S, Majid M, Qayyum A, Awais M, Alnowami M, Khan M. Medical image analysis using convolutional neural networks: a review. *J Med Syst.* 2018;42(11):226.
3. Huang SC, Huo Z, Steinberg E, et al. INSPECT: a multimodal dataset for patient outcome prediction of pulmonary embolisms. *Adv Neural Inf Process Syst.* 2024;36.
4. Pelka O, Koitka S, Rückert J, Nensa F, Friedrich CM. Radiology objects in context (ROCO): a multimodal image dataset. In: *Intravascular Imaging and Computer Assisted Stenting and Large-Scale Annotation of Biomedical Data and Expert Label Synthesis: 7th Joint International Workshop, CVII-STENT 2018 and Third International Workshop, LABELS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 16, 2018, Proceedings* 3. Springer International Publishing; 2018:180-189.
5. Huang SC, Pareek A, Seyyedi S, Banerjee I, Lungren MP. Fusion of medical imaging and electronic health records using deep learning: a systematic review and implementation guidelines. *NPJ Digit Med.* 2020;3(1):136.
6. Huang SC, Pareek A, Zamani R, Banerjee I, Lungren MP. Multimodal fusion with deep neural networks for leveraging CT imaging and electronic health record: a case-study in pulmonary embolism detection. *Sci Rep.* 2020;10(1):22147.
7. Mohsen F, Ali H, El Hajj N, Shah Z. Artificial intelligence-based methods for fusion of electronic health records and imaging data. *Sci Rep.* 2022;12(1):17981.
8. Tariq A, Tang S, Sakhri H, et al. Fusion of imaging and non-imaging data for disease trajectory prediction for coronavirus disease 2019 patients. *J Med Imaging (Bellingham).* 2023;10(3):034004.
9. Tariq A, Lancaster L, Elugunti P, et al. Graph convolutional network-based fusion model to predict risk of hospital acquired infections. *J Am Med Inform Assoc.* 2023;30(6):1056-1067.
10. Zhou Y, Huang SC, Fries JA, et al. RadFusion: benchmarking performance and fairness for multimodal pulmonary embolism detection from CT and EHR. arXiv, arXiv:2111.11665, preprint: not peer reviewed.
11. Tariq A, Patel BN, Sensakovic WF, Fahrenholz SJ, Banerjee I. Opportunistic screening for low bone density using abdominopelvic computed tomography scans. *Med Phys.* 2023;50(7):4296-4307.
12. Tang S, Tariq A, Dunnmon JA, et al. Predicting 30-day all-cause hospital readmission using multimodal spatiotemporal graph neural networks. *IEEE J Biomed Health Inform.* 2023;27(4):2071-2082.
13. Cao M, Yang M, Qin C, et al. Using DeepGCN to identify the autism spectrum disorder from multi-site resting-state data. *Biomed Signal Process Control.* 2021;70:103015.
14. Kazi A, Shekarforoush S, Arvind Krishna S, et al. InceptionGCN: receptive field aware graph convolutional network for disease prediction. In: *Information Processing in Medical Imaging: 26th International Conference, IPMI 2019, Hong Kong, China, June 2-7, 2019, Proceedings* 26. Springer International Publishing; 2019:73-85.

15. Parisot S, Ktena SI, Ferrante E, et al. Disease prediction using graph convolutional networks: application to autism spectrum disorder and Alzheimer's disease. *Med Image Anal.* 2018; 48:117-130.
16. Tariq A, Su L, Patel B, Banerjee I. Prediction of transfusion among in-patient population using temporal pattern based clinical similarity graphs. *AMIA Annu Symp Proc.* 2024;2023:679-688.
17. Tariq A, Kaur G, Su L, Gichoya J, Patel B, Banerjee I. Generalizable model design for clinical event prediction using graph neural networks. medRxiv 23287599. <https://doi.org/10.1101/2023.03.22.23287599>, March 25, 2023, preprint: not peer reviewed.
18. Ying Z, Bourgeois D, You J, Zitnik M, Leskovec J. GNNExplainer: generating explanations for graph neural networks. *Adv Neural Inf Process Syst.* 2019;32:9240-9251.
19. Guo J, Li J, Li D, et al. From images to textual prompts: zero-shot visual question answering with frozen large language models. 2023; 10867-10877. [https://openaccess.thecvf.com/content/CVPR2023/papers/Guo\\_From\\_Images\\_to\\_Textual\\_Prompts\\_Zero-Shot\\_Visual\\_Question\\_Answering\\_With\\_CVPR\\_2023\\_paper.pdf](https://openaccess.thecvf.com/content/CVPR2023/papers/Guo_From_Images_to_Textual_Prompts_Zero-Shot_Visual_Question_Answering_With_CVPR_2023_paper.pdf)
20. Huang W, Abbeel P, Pathak D, Mordatch I. Language models as zero-shot planners: Extracting actionable knowledge for embodied agents. In: *International Conference on Machine Learning 2022*. PMLR; 2022:9118-9147.
21. Kojima T, Gu SS, Reid M, Matsuo Y, Iwasawa Y. Large language models are zero-shot reasoners. *Adv Neural Inf Process Syst.* 2022;35:22199-22213.
22. Wang Z, Wu Z, Agarwal D, Sun J. MedCLIP: Contrastive learning from unpaired medical images and text. *Proc Conf Empir Methods Nat Lang Process.* 2022;2022:3876-3887.
23. Moon JH, Lee H, Shin W, Kim YH, Choi E. Multi-modal understanding and generation for medical images and text via vision-language pre-training. *IEEE J Biomed Health Inform.* 2022;26(12):6070-6080.
24. Zhang J, Huang J, Jin S, Lu S. Vision-language models for vision tasks: a survey. *IEEE Trans Pattern Anal Mach Intell.* 2024;46(8):5625-5644.
25. Kim W, Son B, Kim I. ViLT: vision-and-language transformer without convolution or region supervision. In: *International Conference on Machine Learning 2021*. PMLR; 2021:5583-5594.
26. Singh A, Hu R, Goswami V, et al. FLAVA: a foundational language and vision alignment model. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2022*. IEEE; 2022:15638-15650.
27. Sharma P, Ding N, Goodman S, Soricut R. Conceptual captions: a cleaned, hypernymed, image alt-text dataset for automatic image captioning. 2018:2556-2565. <https://aclanthology.org/P18-1238.pdf>
28. Srinivasan K, Raman K, Chen J, Bendersky M, Najork M. WIT: wikipedia-based image text dataset for multimodal multilingual machine learning. SIGIR'21: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2021:2443-2449. <https://dl.acm.org/doi/10.1145/3404835.3463257>
29. Pont-Tuset J, Uijlings J, Changpinyo S, Soricut R, Ferrari V. Connecting vision and language with localized narratives. In: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings 16, Part V*. Springer International Publishing; 2020:647-664.
30. Desai K, Kaul G, Aysola Z, Johnson J. RedCaps: web-curated image-text data created by the people, for the people. arXiv, arXiv:2111.11431, preprint: not peer reviewed.
31. Lin TY, Maire M, Belongie S, et al. Microsoft COCO: common objects in context. In: *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings 13, Part V*. Springer International Publishing; 2014:740-755.
32. Demner-Fushman D, Kohli MD, Rosenman MB, et al. Preparing a collection of radiology examinations for distribution and retrieval. *J Am Med Inform Assoc.* 2016;23(2):304-310.
33. Johnson AE, Pollard TJ, Berkowitz SJ, et al. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Sci Data.* 2019;6(1):317.
34. Lau JJ, Gayen S, Ben Abacha A, Demner-Fushman D. A dataset of clinically generated visual questions and answers about radiology images. *Sci Data.* 2018;5:180251.
35. Ramirez AH, Sulieman L, Schlueter DJ, et al. The All of Us Research Program: data quality, utility, and diversity. *Patterns.* 2022;3(8):100570.
36. Matthews PM, Sudlow C. The UK Biobank. *Brain.* 2015;138(Pt 12):3463-3465.
37. Sowrirajan H, Yang J, Ng AY, Rajpurkar P. MoCo-CXR: MoCo pretraining improves representation and transferability of chest X-ray models. arXiv, arXiv:2010.05352, preprint: not peer reviewed.
38. Magnuska ZA, Roy R, Palmowski M, et al. Combining radiomics and autoencoders to distinguish benign and malignant breast tumors on US images. *Radiology.* 2024;312(3):e232554.
39. Simon GR, Jordan P, Sako C, et al. AI-based radiomic biomarkers to predict PD-(L)1 immune checkpoint inhibitor response within PD-L1 high/low/negative expression categories in stage IV NSCLC. 2023. [https://ascopubs.org/doi/10.1200/JCO.2023.41.16\\_suppl.1517](https://ascopubs.org/doi/10.1200/JCO.2023.41.16_suppl.1517)
40. Sako C, Jordan P, McCall R, et al. Multi-center real-world data curation and assessment of tumor growth rate and overall survival in advanced NSCLC treated with PD-(L)1 immune checkpoint inhibitor therapy. 2023. <https://digitalcommons.providence.org/publications/8002/>
41. Mutha P, Khorrami M, Sonuga B, et al. Correlation of the combination of CT-derived tumor texture and vessel tortuosity on survival outcomes for immunotherapy but not for chemotherapy in metastatic non-small cell lung cancer (mNSCLC): results from a CheckMate227 (CM227) subset. 2024. [https://ascopubs.org/doi/10.1200/JCO.2024.42.16\\_suppl.8610](https://ascopubs.org/doi/10.1200/JCO.2024.42.16_suppl.8610)
42. Xu Y, Sun L, Peng W, et al. MedSyn: text-guided anatomy-aware synthesis of high-fidelity 3-D CT images. *IEEE Trans Med Imaging.* 2024;43(10):3648-3660.
43. Blankemeier L, Cohen JP, Kumar A, et al. Merlin: a vision language foundation model for 3D computed tomography. arXiv, arXiv:2406.06512, preprint: not peer reviewed.
44. Orcutt X, Mamani R, Sondhi A, Cohen AB, Parikh RB. Evaluating generalizability of practice-changing randomized clinical trials in non-small cell lung cancer using machine learning-based in-silico trials. *J Clin Oncol.* 2023;141(16\_suppl):9130.
45. Bera K, Braman N, Gupta A, Velcheti V, Madabhushi A. Predicting cancer outcomes with radiomics and artificial intelligence in radiology. *Nat Rev Clin Oncol.* 2022;19(2):132-146.
46. Gichoya JW, Banerjee I, Bhimireddy AR, et al. AI recognition of patient race in medical imaging: a modelling study. *Lancet Digit Health.* 2022;4(6):e406-e414.
47. Choi D, Bakhtiari M, Pilcher W, et al. Single-cell analysis of debrided diabetic foot ulcers reveals dysregulated wound healing environment in non-Hispanic blacks. *J Invest Dermatol.* 2024;