Multimodal Medical Classification

Introduction

Multimodal classifiers combine multiple sources, such as images and text, to improve prediction accuracy. Clinical diagnosis often requires integrating imaging studies (e.g., chest X-rays, CT scans) with textual information like radiology reports or clinical notes [1]. These modalities offer complementary insights: images capture visual manifestations of disease, while text provides clinical context and explicit findings.

Models integrating vision and text outperform uni-modal models [2]. For instance, textual reports can highlight subtle findings on X-rays, guiding attention, while images can clarify vague text. By fusing both sources, classifiers improve disease prediction. This review surveys methods focusing on fusion strategies (early/late/joint), pretrained vision–language backbones (CLIP, LLaVA, Flamingo), contrastive learning, representative datasets, benchmark results, and ongoing challenges.

Existing Approaches

Image-only Classification Methods

Disease detection using only images of the area of concern can yield high performance, but such methods cannot incorporate additional text-based features—such as patient history—that often help doctors make better diagnoses. CheXNet [3] is a deep learning Convolutional Neural Network (CNN) architecture to diagnose pneumonia using chest X-rays. The model is a 121-layer CNN that can output the probability of pneumonia. CheXNet yielded a strong performance in pneumonia classification, outperforming radiologist performance on the same dataset. However, while CNNs can yield strong performance on image classification tasks, they cannot incorporate patient history. CheXNet can highlight areas of an X-ray which indicate the pathology identified, explaining its classification decision. Vision transformers (ViT) can also be used to detect diseases using imaging. A ViT was trained on detecting Alzheimer's disease using brain imaging achieving nearly 96.2% accuracy on the test set [4]. ViTs can achieve strong performance as they learn using patches of the images to learn relationships between the patches. It outperforms a pretrained CNN model, ResNet18, achieving a 88.5% accuracy on the same test set [4]. However, ViTs are very deep transformer models and require a significantly large dataset for pretraining. Training on a modest size dataset would lead to overfitting. Like CNNs, ViTs also cannot incorporate text-based features such as the patient's history.

Text-only Classification Methods

When building multimodal classifiers that combine images with text, it helps to start with a baseline that looks at text alone. Previous research has shown that trained physicians can get around 65% accuracy when making diagnoses using only patient questionnaires, without seeing any medical images [5]. More recently, large language models (LLMs) have been tested on similar tasks and were able to reach about 59.1% accuracy when the correct answer was counted as long as it appeared in the top ten predictions [6]. These results highlight both the value and the limits of text-only methods. For example, the LLM study reports accuracy in terms of "top-10," which means the correct answer might be buried several guesses down rather than being the first choice [6]. In real clinical settings, that kind of performance wouldn't be enough on its own. Similarly, the 65% physician result came from doctors with significant training in general ambulatory care. Less experienced physicians performed much worse, showing that the baseline depends heavily on expertise [5]. Altogether, these baselines suggest that while text-only reasoning—whether from humans or algorithms—can provide useful diagnostic clues, it isn't reliable enough to stand alone. This is where multimodal systems come in. By combining images like radiographs or pathology slides with patient-reported symptoms and notes, we can build models that are not only more accurate but also better at capturing the full picture of a case. Multimodal methods have the potential to go beyond the limitations of a single input type and provide stronger, more trustworthy medical AI tools.

Multimodal Fusion Methods

Fusion Strategies and Model Architectures Methods for multimodal classification differ in how they merge image and text features [2]:

- Early fusion (feature-level): concatenates image and text embeddings into a joint representation.
- Late fusion (decision-level) combines outputs of independent image-only and text-only classifiers.
- Joint fusion architectures allow deep cross-modal interaction via shared latent spaces or attention mechanisms.

For example, Gapp et al. fuse chest X-ray and report embeddings using a LLaMA-II backbone, exploring early, late, and mixed fusion pipelines for thoracic disease classification [7]. Vision-language transformers like Med-Flamingo use cross-attention to weigh relevant text for each image region [8]. LLaVA-Med uses a projection/prefix-style fusion [9]. CoD-VQA leverages richer modalities to enhance underrepresented ones, reducing modality-specific bias [10].

Large VLMs such as BiomedCLIP and OpenFlamingo achieve strong zero-/few-shot performance without fine-tuning [11], and Med-Flamingo adapts quickly to new radiology tasks [8]. These strategies use complementary data to boost accuracy, although they require carefully aligned training data and can be sensitive to modality-specific noise.

Contrastive Learning Contrastive learning approaches align image and text embeddings. Wang et al. demonstrate improved downstream classification with contrastive training on chest X-rays and reports [12]. Lei et al. introduce CLIP-Lung uses radiology prompts and disease-specific attributes for lung nodule malignancy prediction, achieving SOTA performance on LIDC-IDRI [13]. These methods also provide a strong initialization for zero- or few-shot adaptation and help models learn meaningful cross-modal embeddings even when labeled examples are scarce.

Datasets Key multimodal datasets include:

- Chest X-ray: OpenI and MIMIC-CXR contain thousands of frontal X-rays with paired radiology reports and labels for common findings [7].
- Lung CT nodules: LIDC-IDRI provides lung CT scans with annotated nodules. Recent work augments LIDC with textual nodule descriptions to classify nodules as benign or malignant [13].
- Ultrasound: UDIAT (breast ultrasound) has images with diagnostic labels. Byra et al. show that even simple text descriptors combined with CLIP enable few-shot ultrasound classification [14].
- Histopathology: Emerging datasets pair pathology images with clinical summaries; for example, multimodal breast cancer challenges include histology slides accompanied by relevant text [15].

These examples cover typical classification tasks (single-label and multi-label) in radiology and related fields. Most current models are evaluated on such radiology benchmarks, but diversity across modalities (e.g. MRI, pathology) and tasks remains limited, highlighting the need for broader multimodal datasets.

State-of-the-Art Models

Recent multimodal classifiers achieve strong results on benchmark tasks:

- LLaMA-II fusion [7]: AUC 0.971 on OpenI chest X-ray classification using fused image-report inputs, surpassing unimodal baselines.
- CLIP-Lung [13]: SOTA on LIDC-IDRI nodule classification with textual knowledge guidance.
- MedCLIP [12]: Outperforms previous self-supervised methods in image–text retrieval and classification.
- Large VLMs: BiomedCLIP and OpenFlamingo achieve competitive zero-/few-shot performance compared to CNNs without the need for fine-tuning [11], while Med-Flamingo adapts quickly to new radiology tasks [8].

In general, vision–language models (both fine-tuned and zero-shot) consistently outperform uni-modal baselines on these benchmarks [12], demonstrating the benefit of multimodal integration.

Challenges and Research Gaps

Key challenges remain in multimodal medical classification:

- Data scarcity, modality, and domain shift: High-quality paired image–text examples are limited and often skewed toward certain modalities (e.g., more X-rays than matched reports), which can bias training. Models trained on a single dataset or domain may also fail to generalize to other modalities (e.g., X-ray to CT) [2]. Ensuring robust performance across imaging types, domains, and patient populations remains an open challenge.
- Interpretability and trust: Modern multimodal models (especially large transformers) are largely opaque. Works like attention maps in MedFuseNet offers partial insight [16], but systematically explaining how image and text combine to yield a prediction remains difficult. This lack of transparency hinders clinical acceptance.
- Bias and hallucination: VLMs may over-rely on textual priors or external knowledge, sometimes ignoring the image and producing “hallucinated” findings [8], [17], e.g., zero-shot ECG VQA models overpredict normal rhythms [18]. Developing methods to ensure predictions are truly grounded in data is crucial.
- Resource and annotation costs: Training and fine-tuning large multimodal models require significant compute, and curating labeled image–text datasets in healthcare is labor-intensive. These practical constraints slow the development and evaluation of new methods.

Addressing these gaps is an active research area. For example, retrieval-augmented models like MMed-RAG incorporate external knowledge to reduce hallucination [17]. Expanding multimodal datasets and developing techniques to interpret multimodal reasoning will be key to deploying reliable clinical classifiers in practice.

Conclusion

CNNs and ViTs can identify patterns in images that indicate the presence of a pathology. However, other relevant textual features, such as patient history, cannot be incorporated into an image-based classifier [3]. Standalone text-only models also cannot classify diseases with high accuracy, as seen in [6].

Using both text and images adds valuable context for improved prediction on medical datasets. Multimodal approaches include VLMs that accept image and text tokens in a single sequence [8], as well as models with dedicated layers for text, image, and their fusion. Our work on predicting answers for images and questions in the OmniMedVQA [19] motivates our exploration of using multimodal models to incorporate all available information for prediction.

References

[1] I. U. Haq, M. Mhamed, M. Al-Harbi, H. Osman, Z. Y. Hamd, and Z. Liu, "Advancements in medical radiology through multimodal machine learning: A comprehensive overview," *Bioengineering (Basel)*, vol. 12, no. 5, p. 477, 2025, doi: 10.3390/bioengineering12050477.

[2] Z. Sun *et al.*, "A scoping review on multimodal deep learning in biomedical images and texts," *Journal of Biomedical Informatics*, vol. 146, p. 104482, 2023, doi: 10.1016/j.jbi.2023.104482.

[3] P. Rajpurkar *et al.*, "CheXNet: Radiologist-level pneumonia detection on chest x-rays with deep learning," *CoRR*, vol. abs/1711.05225, 2017, Available: http://arxiv.org/abs/1711.05225

[4] Y. Lyu, X. Yu, D. Zhu, and L. Zhang, "Classification of alzheimer's disease via vision transformer: Classification of alzheimer's disease via vision transformer," in *Proceedings of the 15th international conference on PErvasive technologies related to assistive environments*, in PETRA '22. New York, NY, USA: Association for Computing Machinery, 2022, pp. 463–468. doi: 10.1145/3529190.3534754.

[5] T. Uehara *et al.*, "Accuracy of diagnoses predicted from a simple patient questionnaire stratified by the duration of general ambulatory training: An observational study," *International Journal of General Medicine*, vol. 7, pp. 13–19, 2013, doi: 10.2147/IJGM.S53800.

[6] T. Tu, M. Schaekermann, A. Palepu, *et al.*, "Towards conversational diagnostic artificial intelligence," *Nature*, vol. 642, no. 8029, pp. 442–450, 2025, doi: 10.1038/s41586-025-08866-7.

[7] C. Gapp, E. Tappeiner, M. Welk, and R. Schubert, "Multimodal medical disease classification with LLaMA II," in *Proceedings of austrian symposium on AI, robotics, and vision 2024*, 2024. doi: 10.15203/99106-150-2-07.

[8] M. Moor *et al.*, "Med-flamingo: A multimodal medical few-shot learner," in *Proceedings of the 3rd machine learning for health symposium*, in Proceedings of machine learning research, vol. 225. PMLR, 2023, pp. 353–367. Available: https://proceedings.mlr.press/v225/moor23a.html

[9] C. Li *et al.*, "LLaVA-med: Training a large language-and-vision assistant for biomedicine in one day," in *Advances in neural information processing systems*, Curran Associates, Inc., 2023, pp. 28541–28564. Available: https://proceedings.neurips.cc/paper_files/paper/2023/file/5abcdf8ecdcacba028c6662789194572-Paper-Datasets_and_Benchmarks.pdf

[10] Q. Lu, S. Chen, and X. Zhu, "Collaborative modality fusion for mitigating language bias in visual question answering," *Journal of Imaging*, vol. 10, no. 3, p. 56, 2024, doi: 10.3390/jimaging10030056.

[11] M.-H. Van, P. Verma, and X. Wu, "On large visual language models for medical imaging analysis: An empirical study," in *2024 IEEE/ACM conference on connected health: Applications, systems and engineering technologies (CHASE)*, 2024, pp. 172–176. doi: 10.1109/CHASE60773.2024.00029.

[12] Z. Wang, Z. Wu, D. Agarwal, and J. Sun, "MedCLIP: Contrastive learning from unpaired medical images and text," in *Proceedings of the 2022 conference on empirical methods in natural language processing*, Association for Computational Linguistics, Dec. 2022, pp. 3876–3887. doi: 10.18653/v1/2022.emnlp-main.256.

[13] Y. Lei, Z. Li, Y. Shen, J. Zhang, and H. Shan, "CLIP-lung: Textual knowledge-guided lung nodule malignancy prediction," in *Medical image computing and computer assisted intervention – MICCAI 2023*, Springer Nature Switzerland, 2023, pp. 403–412. doi: 10.1007/978-3-031-43990-2_38.

[14] M. Byra, M. F. Rachmadi, and H. Skibbe, "Few-shot medical image classification with simple shape and texture text descriptors using vision-language models." 2023. doi: 10.48550/arXiv.2308.04005.

[15] F. Abdullakutty, Y. Akbari, S. Al-Maadeed, A. Bouridane, I. M. Talaat, and R. Hamoudi, "Histopathology in focus: A review on explainable multi-modal approaches for breast cancer diagnosis," *Frontiers in Medicine*, vol. 11, p. 1450103, 2024, doi: 10.3389/fmed.2024.1450103.

[16] D. Sharma, S. Purushotham, and C. K. Reddy, "MedFuseNet: An attention-based multimodal deep learning model for visual question answering in the medical domain," *Scientific Reports*, vol. 11, no. 1, p. 19826, 2021, doi: 10.1038/s41598-021-98390-1.

[17] P. Xia *et al.*, "MMed-RAG: Versatile multimodal RAG system for medical vision language models," in *Proceedings of the 2025 international conference on learning representations (ICLR)*, 2025. Available: https://proceedings.iclr.cc/paper_files/paper/2025/file/a559a5a8aa5ae6682ced009ad97cdb16-Paper-Conference.pdf

[18] T. Seki, Y. Kawazoe, H. Ito, Y. Akagi, T. Takiguchi, and K. Ohe, "Assessing the performance of zero-shot visual question answering in multimodal large language models for 12-lead ECG image interpretation," *Frontiers in Cardiovascular Medicine*, vol. 12, p. 1458289, 2025, doi: 10.3389/fcvm.2025.1458289.

[19] Y. Hu *et al.*, "OmniMedVQA: A new large-scale comprehensive evaluation benchmark for medical LVLM," *arXiv preprint arXiv:2402.09181*, 2024.