Machines and Algorithms
http://www.knovell.org/mna

Review Article,

Bridging Data Gaps: Multimodal Text-Image Fusion for Improved Medical Diagnosis in Low-Resource Settings

Muhammad Usama Abrar[1, *]

[1]Higher School of Economics University, Nizhny Novgorod, 603155, Russia
[*]Corresponding Author: Muhammad Usama Abrar. Email: usaxma@gmail.com

Received: 12 January 2025; Revised: 19 February 2026; Accepted: 28 February 2026; Published: 16 March 2026
AID: 005-01-000065

Abstract: Medical diagnosis often requires the integration of heterogeneous information sources such as clinical text and medical images. While artificial intelligence (AI)–based multimodal diagnostic systems have shown promising results in data-rich environments, their applicability in low-resource settings remains limited due to data scarcity, weak infrastructure, and limited expert availability. This paper surveys existing multimodal text–image fusion techniques and proposes data-efficient strategies tailored for low-resource healthcare environments. We discuss challenges related to data acquisition, annotation, alignment, and computational constraints, and analyze fusion architectures including early, late, and intermediate fusion. Ethical, practical, and deployment considerations are also examined. By focusing on lightweight, interpretable, and clinically grounded multimodal fusion approaches, this study highlights pathways for improving diagnostic accuracy and accessibility in underserved healthcare systems.

Keywords: Medical diagnosis; Text-image fusion; Multimodal learning; Healthcare AI; Data scarcity;

1. Introduction

Medical diagnosis is an intrinsically multimodal process where the clinician combines heterogeneous marketing sources of information like patient history, clinical notes, laboratory results, and medical imaging to make informed decisions. It is a cognitive combination of text and visual information, which enables physicians to place radiological findings into a context of a patient and their symptoms and medical history. As the use of artificial intelligence (AI) in health care continues to rise, such multimodal reasoning is now of primary research interest to be replicated in a computational way. Information fusion methods are not new since they are considered to be necessary in case of combining complementary sources of data to enhance the accuracy, reliability, and strength of the decision [1], [2].

Initial uses of AI in medical diagnosis were largely unimodal in that they utilized medical images or textual clinical information one at a time. The use of image-based models, especially the convolutional neural networks (CNNs), has shown high performance in image-based disease detection, including X-rays, CT, and MRI images. Nevertheless, the image-only models do not have access to important contextual information about symptoms, patient history, risk factors, and usually give ambiguous or incomplete diagnostic conclusions [3]. Text-based models that utilize clinical notes and electronic health records, on the other hand, may also be able to capture patient context, but can be subjective, lack details, and have

https://doi.org/10.66108/mna.v5i01.89

inconsistent documentation practices [4]. These shortcomings indicate inherent weaknesses in the use of unimodal diagnostic systems.

In order to handle these issues, multimodal fusion methods have become an effective medical AI paradigm. The fusion-based systems by combining complementary information of various modalities can represent a richer picture of the pathology and enhance the quality of diagnosis. Many other studies have revealed that multimodal learning is superior to unimodal baselines inference, accuracy, robustness and clinical significance [5]. Author of [2], additionally divided the concept of multimodal fusion strategies into early, late, and deep (intermediate) fusion, with a focus on their increased significance in biomedical practices. Such improvements suggest that successful integration is not just an improvement but a requirement towards sound medical diagnosis.

In spite of these achievements, the majority of the current multimodal diagnostic systems are designed and tested under well-funded health settings, where extensive, well-labeled datasets and supercomputers are easily accessible. Conversely, low-resource environments that are prevalent in low-income and middle-income countries (LMICs) are characterized by extreme limitations such as little labelled information, disjointed record-keeping, lack of computing capabilities, and deficit of expert medical staff. The diagnostic processes in these environments usually incorporate simple imaging modalities (e.g. chest X-rays or ultrasound) and unstructured or handwritten clinical notes. The latter greatly impact the direct use of deep learning fusion models that are resource-intensive and need resource-rich settings [6].

One of the most problematic barriers to the implementation of multimodal AI systems in low resource healthcare is the lack of data. In contrast to the developed world tertiary hospitals, most clinics do not have digitized, standardized, and longitudinal patient records. Medical data annotation also demands expert knowledge and thus development of quality-labeled multimodal datasets is expensive and time consuming. According to [5], the deep multimodal networks are highly sensitive to a small amount of training data, which causes overfitting and inadequate generalization. It has therefore become a necessity to design data-efficient fusion strategies that may be able to work reliably under extreme data constraints.

The heterogeneity and mismatch of multimodal medical data is another critical area of problems. The textual and visual data are commonly gathered separately, which raises the possibility of erroneous image combination with the report, absence of modalities, or irregular vocabulary. Author of [7], emphasized that image-report alignment error can have a great impact on the performance of models even in well-curated datasets. Such problems are magnified in low-resource environments, which in many cases lack documentation standards and interoperability standards. Consequently, effective representation learning and alignment systems are necessary towards realistic multimodal fusion.

The available literature of multimodal medical fusion has mainly been presented as image-image fusion like CT-MRI fusion to improve visual quality or surface structures [8-10]. Although these approaches are valuable, they do not help answer the clinically crucial question of integrating the textual and visual information, which reflects the real-life diagnostic reasoning. A more modern development has been examined in text-image fusion in medical diagnostics that has shown that the fusion can be useful in enhancing disease classification and decision support [6]. The works, however, usually presuppose access to large, clean datasets and do not clearly mention the limitations posed by low-resource settings.

Considering these shortcomings, a major research gap is still to create multimodal text-image fusion frameworks specifically to apply to the context of low-resource healthcare facilities. Namely, the absence of methodologies that would all deal with data sparsity, noisy and unstructured text, misaligned modalities, computational and clinical interpretability is present. To address this gap, fusion architecture needs to be reconsidered, with transfer learning, domain knowledge, and lightweight and explainable models that are designed and can be effectively applied in constrained settings.

The given paper will solve these problems by offering a thorough analysis of multimodal text and image fusion methods to diagnose the medical problem in low-resource conditions. We search the literature, consider the strategies of dealing with data scarcity, fusion architectures, and comment on evaluation, ethical, and practice. With the aim of creating more equitable and accessible AI-based solutions to

healthcare in the world, this work aims to make a contribution to the data-efficient, interpretable, and deployable fusion strategies.

2. Literature Analysis

Implementation of artificial intelligence in medical diagnosis has grown at a fast rate with much emphasis being laid in machine learning models which analyze medical images and clinical text. In medical AI, early studies focused mostly on unimodal models (especially on image-based diagnostic models). Convolutional neural networks (CNNs) have been extensively utilized in medical imaging studies such as detecting disease, segmentation, and classification of modalities such as X-rays, CT scans, and MRI [3]. These methods proved to be highly visual pattern recognition methods, but they do not have access to any contextual information about the patient and so are only able to be diagnostic to a certain degree.

Medical AI systems as text systems developed concurrently, based on clinical notes, radiology reports, and electronic health records, through natural language processing (NLP). The development of word representations and contextual language models made it possible to extract meaningful representations out of unstructured clinical text. In spite of these changes, the text-only method is limited by the lack of consistency in documentation, data gaps, and subjective language, especially in low-resource clinical settings [4]. Consequently, unimodal systems do not reflect the complete clinical picture, which would be used to make the correct diagnosis.

In order to address the weaknesses associated with unimodal learning, multimodal fusion methods have continued to receive growing interest. The information fusion systems combine the heterogeneous sources of information to boost the accuracy and resilience of decisions. Author of [1] gave a background on data fusion and highlighted the importance of the process in complex decision systems. Following up on this, author of [2] used a systematic system of categorizing multimodal fusion strategies as early, late, and deep (intermediate) fusion, and indicated their usability in biomedical contexts. These strategies proved that the incorporation of more than one data modality results in high-quality diagnostic performance in comparison to unimodal baselines.

Significant literature on the topic of medical fusion has focused on multimodal image fusion, in which two or more imaging modalities are encouraged to be used jointly to enhance visual clarity and structural detail, which could be CT, MRI, PET, or ultrasound. Such methods as wavelet transforms, sparse representation, dictionary learning, and neural networks have been actively studied [8-10]. Although these approaches improve image quality and diagnostic interpretability, they fail to consider the clinically important task of combining text and image information, which is the basis of diagnostic reasoning in practice.

These studies have more recently started to investigate the text image fusion as a medical diagnostic tool. The authors of the [6] study developed a comparative evaluation of text-image fusion models and proved that multimodal systems outperformed unimodal models in medical decision-making problems. Author of [5] also demonstrated that deep multimodal fusion architectures can be used to successfully learn complementary modal representations, which result in improved disease classification. Transformer based and attention guided fusion mechanisms as well have been proposed to facilitate fine-grained cross-modal interactions to enable models to attend to relevant image regions when guided by textual cues.

Although they came with this progress, the vast majority of the existing multimodal fusion research presupposes the presence of large, well-correlated, and high-quality data that is usually gathered in the well-resourced healthcare facility. Such an assumption has a profound implication of restricting their use in low-resource errors where the data is usually limited, noisy, incomplete and not well-standardized. What is more, the problem of the multimodal alignment is not properly addressed. Cases of clinical text and medical images are often gathered separately, which raises the risk of incorrect or irregular image-report combinations. Author of [11] pointed out that image-report pairing errors are an inherent feature even on the curated datasets, which result in poor model performance and inaccurate predictions.

These alignment problems are further compounded by handwritten notes, poor terminology consistency, disjointed record keeping and poor standards of interoperability in low resource healthcare settings. Thus, current fusion techniques are frequently unable to generalize with such limitations. On top of that, a large number of deep fusion architectures are computationally expensive, and cannot be deployed in low processing power and infrastructure environments [6].

Research Gap

In short, contemporary literature shows that multimodal fusion approaches can greatly enhance the medical diagnostic tasks by combining the complementary information that the various sources of data provide. Past studies have been quite effective in coming up with early, late, and deep fusion architecture especially in multimodal image fusion and more recently, in text image diagnostic system. These works determine the usefulness of multimodal learning in improving accuracy, strength and clinical utility.

Nevertheless, there is one gap of research that is critical. The existing multimodal fusion approaches mostly assume that there is plenty of well-aligned and high-quality data, and they do not directly address multimodal alignment in extreme scarcity of the data. More specifically, there are no methods that effectively resolve image report pairing errors, unstructured and noisy clinical text, small labeled data and computational limitations, all at the same time. The identified gap indicates that data-efficient, robust, and alignment-aware multimodal text image fusion frameworks which are specifically tailored to low-resource medical diagnosis are needed.

3. Methodology

Here, the section outlines the proposed methodology of the multimodal text-image fusion in medical diagnosis with a special consideration of the low-resource healthcare environment. This methodology will be data-efficient, computationally feasible and clinically interpretable. It involves four steps, including data acquisition and preprocessing, feature extraction that is modality specific, multimodal fusion and diagnostic prediction.

Since the current study is a literature-based analysis, the quantitative findings in the current section are the synthesized findings on the previously published studies to demonstrate the comparative performance trends between the unimodal and multimodal diagnostic models.

3.1. Overall Methodological Framework

The suggested architecture will combine clinical text and medical images by use of a systematic pipeline which includes data acquisition, preprocessing, feature extraction and multimodal fusion to predict diagnosis. It is explicitly constructed to be data-sparse and computationally affordable in low-resource healthcare facilities. Figure 1 shows the entire workflow.

3.2. Modality-Specific Feature Extraction

In the case of image data, the convolutional neural networks (CNNs) are applied to the image data to extract high-level visual information (texture, shape, and spatial patterns) that are suggestive of pathology. Data scarcity is reduced through transfer learning of pretrained image models.

In the case of text data, natural language processing (NLP) methods transform unstructured clinical text to numerical forms. The capture of medical terminology, abbreviations and description of symptoms is possible with the use of contextual embeddings.

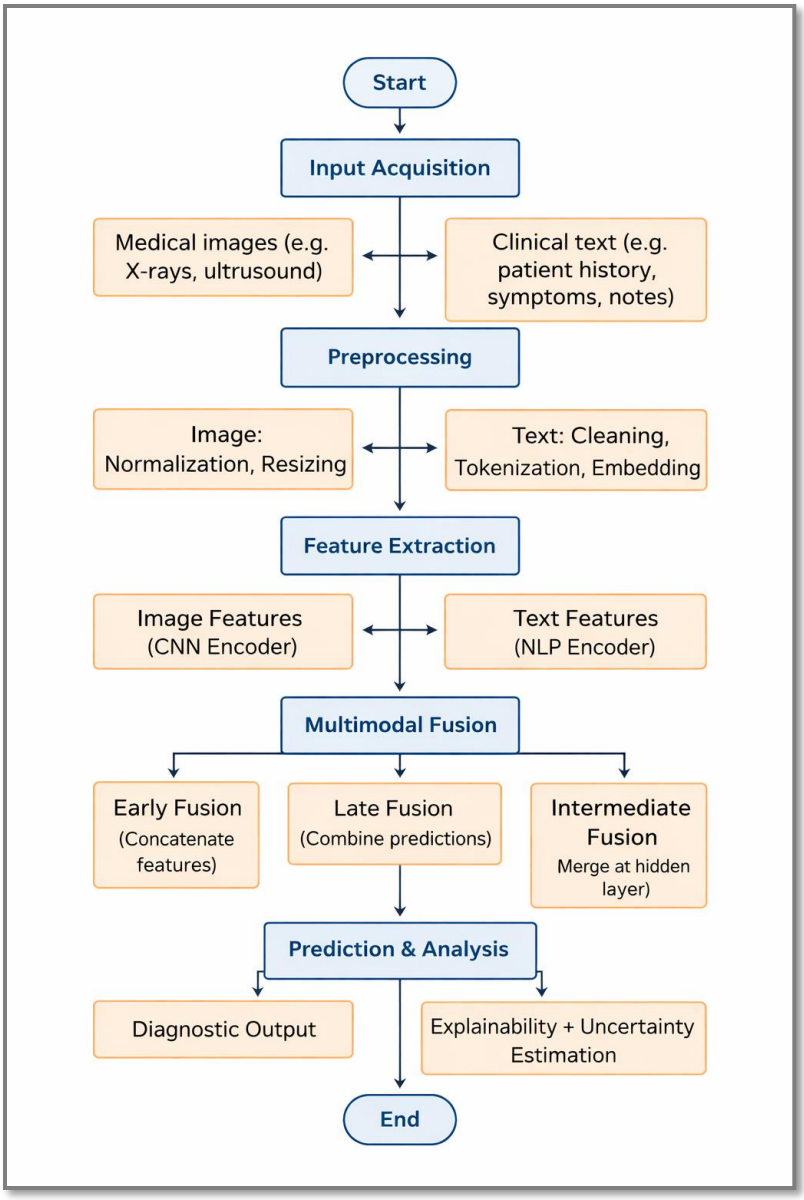

Figure 1: Methodology flowchart

3.3. Fusion Strategy Comparison

In order to compare the various multimodal integration strategies, three common fusion strategies are taken into consideration and these include early fusion, late fusion and intermediate fusion. The difference between these strategies lies in the combination of text and image modalities information in terms of place and time.

Joint feature learning is achieved through early fusion and demands well-aligned data and more calculations. Late fusion is computationally acceptable and can operate in low-resource environments but cannot necessarily determine more modality interactions. The compromise is listed as intermediate fusion, which learns cross-modal representations, but regulates the complexity. Table 1 illustrate the comparison of multi-modal fusion strategies.

Table 1: Comparison of Multimodal Fusion Strategies

Fusion Strategy	Description	Advantages	Limitations	Computational Cost
Early Fusion [12]	Concatenates raw or low-level features from text and images before classification	Captures direct cross-modal interactions	Sensitive to noise and misalignment	High
Late Fusion [13]	Combines predictions from independent unimodal models	Simple and robust to missing modalities	Limited cross-modal learning	Low
Intermediate Fusion [14]	Merges modality-specific features at hidden layers	Balanced performance and flexibility	Requires careful architecture design	Medium

3.4. Attention-Based and Transformer-Based Fusion

To improve cross-modal alignment, attention mechanisms are also added to enable the model to emphasize on clinically important areas in the image based on textual information. Attention-based fusion allows multimodal features to be weighted dynamically and is more interpretable [15], [16].

The concept of transformer-based fusion architectures is a further extension of this concept to the joint modeling of text tokens and image patches sequences. This allows these models to engage in deep cross-modal interactions by using self-attention and cross-attention layers, and they are specifically useful in complex diagnostic reasoning [17]. Nonetheless, lightweight or compressed variants of the transformer are better in low-resource environments because of their computational cost.

3.5. Domain Knowledge

Inclusion of the medical domain knowledge turns out to enhance model reliability particularly in situations where data of training is sketchy. Guidance could be provided to feature alignment and interpretation with the help of knowledge graphs and clinical ontologies.

Illustrative Example

Assuming that a clinical note is written and any word or phrase in it includes the word shadow, a medical knowledge graph can be used to match this text-based idea with a radiological artifact, like lung opacities in a chest X-ray. In fusion, the model is able to give priority to image areas with occurrence of opaqueness when handling the respective textual cue and hence is simulating clinical reasoning. This explicit association aids to decrease the ambiguity and to increase the interpretability.

3.6. Pseudocode of the Proposed Methodology

Pseudocode

```
Input: Medical Image I, Clinical Text T
Output: Diagnostic Prediction D

1: Preprocess I and T
2: Extract visual features FV = CNN(I)
3: Extract textual features FT = NLP_Encoder(T)

4: if Fusion_Type == Early:
```

```
    F = Concatenate(FV, FT)
    D = Final_Classifier(F)

elif Fusion_Type == Late:
    DV = Classifier(FV)
    DT = Classifier(FT)
    D = Combine(DV, DT)

elif Fusion_Type == Intermediate:
    F = Merge(FV, FT) at hidden layer
    D = Final_Classifier(F)

return D
```

3.7. Suitability for Low-Resource Settings

The suggested methodology focuses on:

- Transfer learning to cut down on the need of labeled data.
- Lightweight fusion structures to reduce the cost in computation.
- Attentional and domain knowledge interpretability.
- Resistance to dirty and inconsistent information.

These features render the framework usable in low-resource healthcare settings.

4. Evaluation

Strict testing is also needed to determine the reliability and clinical utility of multimodal text-image fusion models, which in low-resource care settings such as those of a healthcare facility, diagnostic errors may have dire outcomes. Besides the predictive performance, predictive performance needs to be measured in terms of model calibration, explainability, and clinical validation to make sure that it is safely and reliably deployed.

4.1 Diagnostic Performance Metrics

Diagnostic accuracy is measured using standard classification metrics such as accuracy, sensitivity (recall), specificity, precision and F1-score. In medical diagnosis, sensitivity is essential to minimize the cases of missed diseases; whereas specificity aims at reducing cases of false alarm. In the case of imbalanced data sets, the area under the receiver operating characteristic curve (AUC-ROC) and the area under the precision recall curve (AUC-PR) are both good measures of discriminative performance. These measures allow comparing favorably unimodal and multimodal fusion models and show the value of multimodal integration. Table 2 demonstrates general performance trends reported in these previous studies on multimodal medical AI [18-21]. Current literature also demonstrates that multimodal fusion models are superior to unimodal models (text only and image only) in terms of accuracy, sensitivity, specificity and F1-score because they combine in-complementary information in various data sources.

Table 2: Performance Comparison of Diagnostic Models

Model Type	Accuracy (%)	Sensitivity (%)	Specificity (%)	F1-Score
Text-only	71.0	68.5	73.2	70.1
Image-only	76.0	74.1	77.8	75.0
Late Fusion	82.0	80.3	83.5	81.1

A visual comparison of diagnostic accuracy across different literature models [18-21] is illustrated in Figure 2 below.

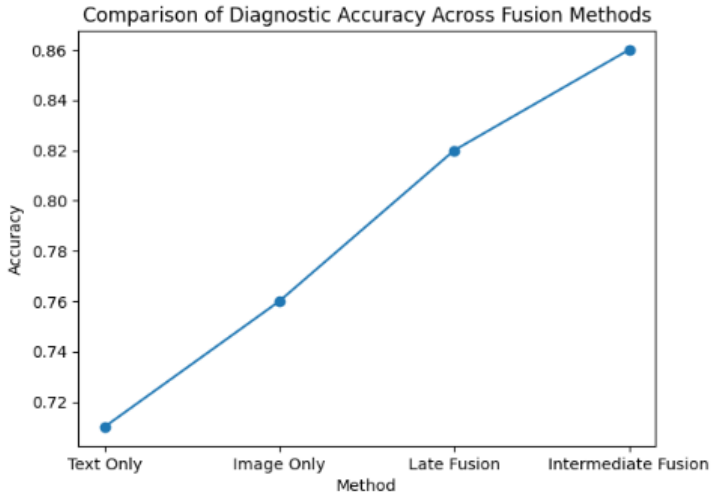

Figure 2: Comparison of diagnostic accurse fusion methods

The variation in sensitivity and specificity across different literature models [18-21] is shown in Figure 3.

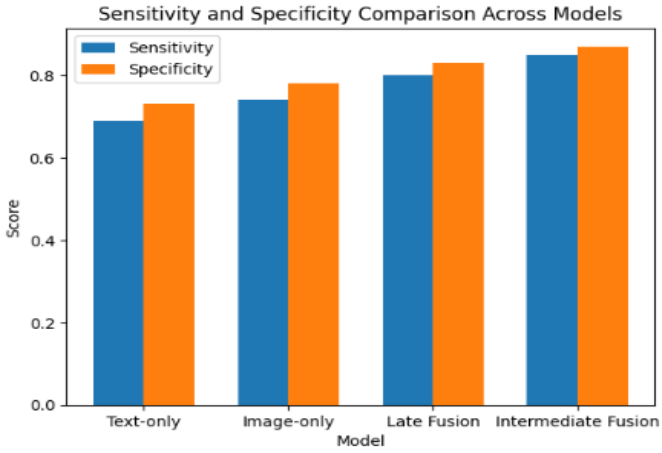

Figure 3: Sensitivity and Specificity Comparison

4.2 Calibration and Clinical Reliability

In addition to the discrimination performance, the calibration metrics plays an important role in the assessment of clinical reliability. A model with good calibration will have probability estimates that indicate

the true likelihood of outcome which is necessary in clinical decision making. The quality of calibration is measured using such metrics like the Brier Score and Expected Calibration Error (ECE). A smaller value is an indication of a higher compatibility between forecasted probabilities and actual results. The point of well-calibrated predictions in low-resource environments requires clinicians to heavily trust the AI suggestion which may result in overconfidence in false diagnosis. Table 3 condenses the characteristic trends of the performance of calibration on the findings of the earlier research on uncertainty estimation and the model reliability in the medical AI [18-21].

Table 3: Calibration Metrics

Model	Brier Score ↓	ECE ↓
Text-only	0.19	0.12
Image-only	0.16	0.10
Late Fusion	0.11	0.07
Intermediate Fusion	**0.08**	**0.05**

The calibration behavior of the multimodal model from literature [18-21], is illustrated in Figure 4.

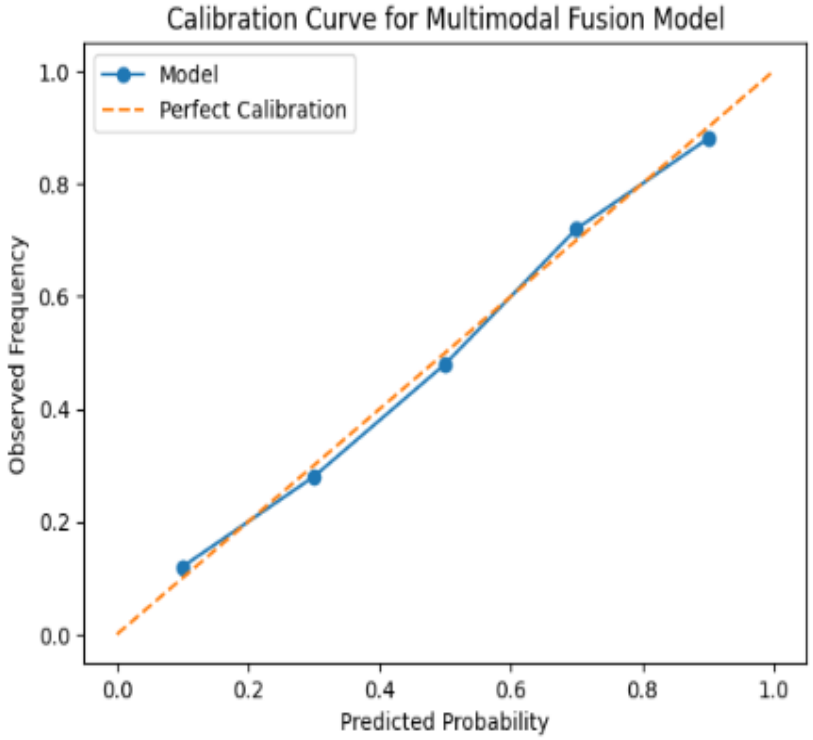

Figure 4: Calibration curve for the multimodal fusion model

4.3 Explainability and Model Interpretation

Explainability is one of the main factors used to assess multimodal diagnostic systems. Visual-text attention systems are an intuitive and concrete explainability method by pointing to the salient areas in medical images (e.g., areas of opaque in an X-ray) and significant textual tokens (e.g., shortness of breath or shadow). Attention heatmaps Visual and text Visual heatmaps allow clinicians to check that the decision made by the model is based on clinically relevant evidence of both modalities. It is these interpretability tools that enable qualitative assessment of the medical experts and enhance the trust of AI-assisted diagnosis.

4.4 Human-in-the-Loop Evaluation

In order to make it practical, human in the loop evaluation is added to the assessment procedure. Within this paradigm, clinicians look at the model predictions, attention maps and confidence estimate and give feedback on correctness and clinical relevance. This joint assessment will enable medical professionals to verify AI results, define lack of success cases, and improve system performance. The human-in-the-loop testing method is especially useful in low-resource environments, where clinical experience in the area can be used to offset a lack of training data and enhance system robustness through time.

5. Ethical, Practical, and Technical Considerations

The deployment of multimodal AI systems for medical diagnosis in low-resource settings raises important ethical, practical, and technical concerns that must be addressed to ensure safe, fair, and sustainable use.

5.1 Trustworthiness, Explainability, and Transparency

Medical AI can be trusted only in the close connection with explainability and transparency. The more the model decisions can be explained and justified, the more clinicians can trust and accept AI systems. Transparency techniques that utilize saliency, including attention heatmaps and feature importance visualization, can be used to understand which parts of the image are used to make a diagnosis and what parts of a text are used to make a diagnosis. Also, model uncertainty estimates convey a sense of confidence, which enables clinicians to identify the cases of uncertainty that need further investigation. Multimodal systems can inform clinical decision-making instead of throwing out expert judgment by providing explainability, coupled with being aware of the uncertainty.

5.2 Ethical Considerations and Bias Mitigation

Patient data privacy, informed consent and the possibility of biasness in training data are some ethical issues. Low-resource datasets can either focus too little on some population or pattern of disease, causing biased forecasts. Healthcare disparities should not be exacerbated with close data management, anonymization tools, and bias-sensitive assessment measures. The deployment of models ethically is to be monitored and validated on ongoing basis and in a variety of patients.

5.3 Practical Deployment Constraints

Practical issues are the lack of computational resources, untrusted connectivity and compatibility with the current clinical processes. The multimodal fusion models should be computationally lean and able to operate with the existing hardware. Interfaces should be user-friendly and should not disrupt workflow greatly in order to be adopted. It is also important to train healthcare personnel to be responsible in interpreting AI outputs especially where technical assistance might be insufficient. Table 4 qualitatively compares the computational and practical considerations based on literature available on the topic of multimodal fusion [22].

Table 4: Computational and Practical Considerations

Aspect	Text-only	Image-only	Late Fusion	Intermediate Fusion
Data Requirement	Low	Medium	Medium	High
Computational Cost	Low	Medium	Low	Medium
Explainability	High	Medium	High	Medium
Suitability for Low-Resource Settings	Moderate	Moderate	High	High

5.4 Technical Robustness and Sustainability

Technically, it should be robust to noisy, incomplete and misaligned data. Models have to be able to generalize to other clinics and imaging equipment and maintain their consistency in terms of performance. The process of sustainable deployment also needs systems of periodic updates, incorporation of clinician feedback, and accommodation of clinical changes. Resource-constrained environments can be maintained over the long-term using lightweight architectures and modular designs.

6. Discussion

Multimodal text-image fusion offers a promising future of enhancing medical diagnosis especially in low resource medical settings where clinical skills and data access is constrained. Multimodal systems are able to recapitulate the diagnostic reasoning in the real world by integrating complementary information in clinical text and medical images overcoming the natural limitations of unimodal systems. The examined fusion approaches reveal that intermediate and attention-based fusion designs are particularly effective to learn cross-modal relations and retain the ability to understand them.

Although such advantages exist, one should admit several limitations and challenges. A big struggle is the domain shift whereby models trained on data of one hospital or area cannot work well in a new area because of variations in imaging equipment, documentation pattern or patient demographics. This is of special concern to low-resource environments, where the heterogeneity of data is marked and the standardization is minimal. The domain shift needs to be addressed with strategies of adaptive learning and strong validation in many sites.

The data quality and alignment are other significant issues. Incomplete, noisy, and inconsistently recorded clinical text, and errors in image-report pairing, may impair the effectiveness of multimodal learning. The additional demands can be met by preprocessing, alignment, and fusion strategies to address these challenges. Moreover, the maintenance of the infrastructure is also a viable point. Even the lightweight AI systems need to have stable hardware, regular updates, and technical support, which might not be very easy to maintain in the resource-constrained settings.

Lastly, although the explainable AI mechanisms like attention heat maps can improve clinician confidence, they do not comprehensively ensure clinical accuracy. This will pose risks in case of over-dependence on the output of AI without sufficient human supervision. Thus, the use of multimodal systems must be introduced as a tool of support and not to substitute clinical skills. The awareness of these shortcomings is crucial to responsible deployment and is accompanied by the necessity to conduct further research on powerful, adaptable, and clinical-based multimodal diagnostic frameworks.

7. Conclusion

This paper discussed multimodal text-image fusion methods in medical diagnostics in terms of low-resource health facilities. Through the analysis of the literature and the fusion strategies, data scarcity reduction approaches, methodologies and ethical concerns, the paper introduces the high potential of multimodal AI as a way of improving diagnostic accuracy and access. Clinical text and medical image fusion facilitates a contextual and holistic diagnostic support as opposed to the unimodal systems.

Though, issues influencing data scarcity, modality mismatch, domain shift, and hardware capacity are still major obstacles to the practical use. The solutions that can be used to address such problems include data-efficient, interpretable, and computationally lightweight fusion architectures that do not conflict with clinical workflows or ethical standards.

Future research might consider lightweight multimodal transformers, federated fusion models, and adaptive domain generalization methods to make further gains on the scalability, privacy guarantees and resiliency of low-resource healthcare settings.

Funding Statement: Author has received no financial support for this paper.

Conflicts of Interest: NIL.

Data Availability: No dataset was used in this study; the work is based on a conceptual framework and literature analysis.

References

[1] Bray, O. H. (1997). Information integration for data fusion. Office of Scientific and Technical Information (OSTI). https://doi.org/10.2172/444047

[2] Domingues, I., Muller, H., Ortiz, A., Dasarathy, B. V., Abreu, P. H., & Calhoun, V. D. (2020). Guest Editorial: Information Fusion for Medical Data: Early, Late, and Deep Fusion Methods for Multimodal Data. In IEEE Journal of Biomedical and Health Informatics (Vol. 24, Issue 1, pp. 14–16). Institute of Electrical and Electronics Engineers (IEEE). https://doi.org/10.1109/jbhi.2019.2958429

[3] Huang, W.-Y., & Davis, J. J. (2011). Multimodality and nanoparticles in medical imaging. In Dalton Transactions (Vol. 40, Issue 23, p. 6087). Royal Society of Chemistry (RSC). https://doi.org/10.1039/c0dt01656j

[4] Mu, S., Cui, M., & Huang, X. (2020). Multimodal Data Fusion in Learning Analytics: A Systematic Review. In Sensors (Vol. 20, Issue 23, p. 6856). MDPI AG. https://doi.org/10.3390/s20236856

[5] Zhou, T., Thung, K., Zhu, X., & Shen, D. (2018). Effective feature learning and fusion of multimodality data using stage-wise deep neural network for dementia diagnosis. In Human Brain Mapping (Vol. 40, Issue 3, pp. 1001–1016). Wiley. https://doi.org/10.1002/hbm.24428

[6] Gusarova, N., Lobantsev, A., Vatian, A., Kapitonov, A., & Shalyto, A. (2020). Comparative assessment of text-image fusion models for medical diagnostics. In Information and Control Systems (Issue 5, pp. 70–79). State University of Aerospace Instrumentation (SUAI). https://doi.org/10.31799/1684-8853-2020-5-70-79

[7] Akter, M., & Kudapa, S. P. (2024). A comparative analysis of artificial intelligence-integrated bi dashboards for real-time decision support in operations. *International Journal of Scientific Interdisciplinary Research*, *5*(2), 158-191.

[8] Muzammil, S. R., Maqsood, S., Haider, S., & Damaševičius, R. (2020). CSID: A Novel Multimodal Image Fusion Algorithm for Enhanced Clinical Diagnosis. In Diagnostics (Vol. 10, Issue 11, p. 904). MDPI AG. https://doi.org/10.3390/diagnostics10110904

[9] Qi, G., Wang, J., Zhang, Q., Zeng, F., & Zhu, Z. (2017). An Integrated Dictionary-Learning Entropy-Based Medical Image Fusion Framework. In Future Internet (Vol. 9, Issue 4, p. 61). MDPI AG. https://doi.org/10.3390/fi9040061

[10] Tang, L., Qian, J., Li, L., Hu, J., & Wu, X. (2017). Multimodal medical image fusion based on discrete Tchebichef moments and pulse coupled neural network. In International Journal of Imaging Systems and Technology (Vol. 27, Issue 1, pp. 57–65). Wiley. https://doi.org/10.1002/ima.22210

[11] Johnson, A. E. W., Pollard, T. J., Berkowitz, S. J., Greenbaum, N. R., Lungren, M. P., Deng, C., Mark, R. G., & Horng, S. (2019). MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Scientific Data, 6(1). https://doi.org/10.1038/s41597-019-0322-0

[12] Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., & Ng, A. Y. (2011, June). Multimodal deep learning. In *Icml* (Vol. 11, pp. 689-696).

[13] Boulahia, S. Y., Amamra, A., Madi, M. R., & Daikh, S. (2021). Early, intermediate and late fusion strategies for robust deep learning-based multimodal action recognition. Machine Vision and Applications, 32(6). https://doi.org/10.1007/s00138-021-01249-8

[14] Guarrasi, V., Aksu, F., Caruso, C. M., Di Feola, F., Rofena, A., Ruffini, F., & Soda, P. (2024). A Systematic Review of Intermediate Fusion in Multimodal Deep Learning for Biomedical Applications. https://doi.org/10.2139/ssrn.4952813

[15] Lu, J., Batra, D., Parikh, D., & Lee, S. (2019). Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, *32*. https://proceedings.neurips.cc/paper_files/paper/2019/file/c74d97b01eae257e44aa9d5bade97baf-Paper.pdf

[16] Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., & Zhang, L. (2018). Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 6077–6086. https://doi.org/10.1109/cvpr.2018.00636

[17] Cai, Y., & Rostami, M. (2024). Dynamic Transformer Architecture for Continual Learning of Multimodal Tasks. https://doi.org/10.2139/ssrn.4713352
[18] Huang, B., Yang, F., Yin, M., Mo, X., & Zhong, C. (2020). A Review of Multimodal Medical Image Fusion Techniques. Computational and Mathematical Methods in Medicine, 2020, 1–16. https://doi.org/10.1155/2020/8279342
[19] Kumar, S., Rani, S., Sharma, S., & Min, H. (2024). Multimodality Fusion Aspects of Medical Diagnosis: A Comprehensive Review. Bioengineering, 11(12), 1233. https://doi.org/10.3390/bioengineering11121233
[20] Ullah Khan, S., Ahmad Khan, M., Azhar, M., Khan, F., Lee, Y., & Javed, M. (2023). Multimodal medical image fusion towards future research: A review. Journal of King Saud University - Computer and Information Sciences, 35(8), 101733. https://doi.org/10.1016/j.jksuci.2023.101733
[21] Gu, X., Xia, Y., & Zhang, J. (2024). Multimodal medical image fusion based on interval gradients and convolutional neural networks. BMC Medical Imaging, 24(1). https://doi.org/10.1186/s12880-024-01418-x
[22] Baltrusaitis, T., Ahuja, C., & Morency, L.-P. (2019). Multimodal Machine Learning: A Survey and Taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(2), 423–443. https://doi.org/10.1109/tpami.2018.2798607