International Journal of Scientific Research and Engineering Development-– Volume 9 Issue 2, Mar-Apr 2026
Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

Multi-Modal Medical Dataset Creation and Deep Learning for Improved Neural Network Diagnostics: A Comprehensive Literature Review

Keshav*, Manvik Khandal, Vinayak Khandelwal
Department of Artificial Intelligence and Data Science, Poornima Institute of Engineering and Technology
Jaipur, Rajasthan, India
Email: 2022pietcakeshav030@poornima.org

Abstract:
Medical artificial intelligence has achieved some amazing progresses over the last few years, but the systems deployed can only capture a single stream of data, typically radiology data, and do not leverage the more powerful context traditionally provided by clinicians: patient history, clinical notes, lab data, and demographic meta data. The overview examines five significant scholarly publications released in the years of 2019-2022 and delimits the current situation in multi-modal deep learning in medical diagnostics. We contrast the specifications of dataset creation, the encoding of the modality and prototyping of features, in both studies, on regular performance measures. The findings are congruent: multi- modal models are more precise and have higher AUC rates by 6- 10 percentage points compared with the single- modality models, and the largest improvements in the cases with the diagnostically ambiguous cases and those where the visual information alone is not sufficient to provide an accurate diagnosis. We introduce lasting challenges like modality alignment, data unavailable in inference time, interpretability, and demographic bias, and indicate the most promising future directions of practice.
Index Terms Multimodal learning, medical imaging, clinical NLP, deep learning, feature fusion, CNN, transformer models, medical AI, dataset generation, diagnostic accuracy.

I. INTRODUCTION

As a radiologist, she hardly ever looks only at the scan when examining a chest X-ray. She looks at the clinical notes, takes into account the age of the patient, investigates whether he or she smokes, and in some cases she compares the scan with one conducted several months ago. The final diagnosis comes out of this amalgamation of messages—not out of one source. In its mainstream form, deep learning has not yet kept up with that clinical fact.

During most of the last ten years, medical diagnosis AI systems were trained on single-modality data, most frequently on large sets of labelled images. These models have achieved remarkable scores on public benchmarks but their utility in actual hospital workflows has been unequal. An increasing body of literature indicates that the gap between benchmark accuracy and clinical utility can be significantly narrowed by learning on so-called multi-modal data: pipelines that jointly process images, free-text reports, and structural patient metadata into an integrated learning system.

However, classical image processing methods [8,9] have provided a light-weight and easily interpretable alternative. Previous research has shown that contour detection, geometric scaling, and reference objects can be used to estimate the dimensions of objects, such as size of human feet [1]. There is a gap in the reality-based variability and asymmetry, though these strategies and methods tend to presume the existence of symmetrical feet and regulated imaging environments.

In this literature review, five representative research articles are gathered and synthesised, each making a unique contribution to dataset construction, fusion strategy, or evaluation strategy for multi-modal medical AI. We pursue three goals:

1) Provide a clear description of what each paper did and what it discovered.
2) Compare them on common dimensions including modal- ities used, fusion strategy, and reported accuracy.
3) Identify recurring limitations and open problems that have not yet been resolved.

The five works reviewed span dataset creation [1], end-to-end fusion architecture design [2], [5], domain-specific language model pre-training [3], and large-scale meta-analysis of the broader field [4].

II. BACKGROUND AND CONCEPTUAL FRAMEWORK

2.1 Why Multi-Modality Matters

Two completely different underlying conditions may appear exactly the same in a single medical image. An image-only classifier is practically playing a guessing game where there is ambiguity in visual signals without further information such as lab markers, patient demographics, or a radiologist's narrative. It has long been established that the addition of an extra

 ©IJSRED: All Rights are Reserved

modality increases sensitivity and specificity. Var- ious modalities contain complementary information: spatial structure and morphology are captured through images, the history of clinical reasoning and symptoms are presented through text, whereas metadata captures demographic and epidemiological context. Fusion enables a model to access all three simultaneously, just like a clinician.

2.2 Common Medical Data Modalities

Table I summarizes the most common types of medical data found in multi-modal AI pipelines along with their standard encoding method.

Table-1
Common Modalities in Multi-Modal Medical AI and their typical Neural Encoders

Modality	Data Type	Format	Typical Encoder
X-ray / CT / MRI	Medical image	DICOM, PNG	ResNet, DenseNet, EfficientNet
Radiology Reports	Free text	TXT, PDF	BERT, BioBERT, ClinicalBERT
EHR / Meta-data	Structured data	CSV, JSON	MLP / Dense layers
Pathology Slides	Whole-slide img	SVS, TIFF	Graph CNN, Vision Transformer
Genomic Profiles	Sequence / num	FASTQ, CSV	1-D CNN, feed-forward network
Physiological Signals	Time series	EDF, CSV	LSTM, Temporal CNN

2.3 Taxonomy of Fusion Strategies

The literature of multi-modal medical AI reflects three general fusion paradigms, knowledge of which is necessary to understand the design choices in each reviewed article.

Early fusion. Modalities are fused early in their raw or minimally processed forms and then fed to one joint model. It is easy to apply but requires modalities to be similar to each other, and the structure of each modality is discarded early in the pipeline.

Late fusion. Independent models per modality are trained and the predictions are combined either through averaging, voting, or a learned weighting decision. This is more resistant to missing modalities at inference time but is unable to model cross-modal interactions during learning.

Intermediate (hybrid) fusion. This variant employs modality-specific encoders whose outputs are fused together, typically by attention mechanisms, followed by a final classification layer. This has become the most prevalent strategy in high-performing systems since it maintains modality-specific structure while also allowing rich cross-modal interaction.

III. PAPER 1 — MIMIC-CXR: PAIRED CHEST X-RAYS AND RADIOLOGY REPORTS [1]

3.1 Context and Motivation

Most publicly available radiology datasets released up to the publication of MIMIC-CXR were either images or text only, seldom both for the same patient encounter. Researchers studying image-text alignment or automated report generation faced the challenge of assembling heterogeneous sources from various institutions, which necessarily led to alignment errors and poor reproducibility. MIMIC-CXR bridged this gap by publishing image-text pairs based on one coherent source of clinical data.

3.2 Dataset Creation

The dataset was sampled from the Beth Israel Deaconess Medical Center and has 227,835 imaging studies across 64,588 patients with unique IDs. All studies contain one or multiple DICOM-based chest X-ray images (frontal or lateral) and a de-identified radiology report. The report text was used to extract fourteen pathology labels (such as pneumonia, pleural effusion, cardiomegaly, and pneumothorax) using the CheXpert NLP labeller, which provided structured uncertainty labels: positive, negative, uncertain, or blank. This is more realistic in clinical use than binary labelling, since radiologists habitually trade off their results when visual evidence is unclear.

3.3 Significance for Multi-Modal Research

MIMIC-CXR made possible a range of downstream research directions that were hitherto infeasible, including automated report generation, cross-modal image retrieval, vision-language pre-training for medical AI, and multi-modal classification that simultaneously reads a scan and its report. This dataset has since become the de facto standard benchmark for chest-finding classification and report generation, with studies on both modalities invariably reporting higher results than image-only models.

3.4 Limitations

The CheXpert labeller adds noise: independent error analyses revealed significant inaccuracies on ambiguous or negated findings—exactly the ones in which multi-modal learning is most useful. The sample also represents only one North American institution, indicating that its demographic representation reflects that hospital rather than the wider patient population. Formatted clinical data including vital signs, lab data, and smoking history are not directly included, restricting the range of three-way fusion experiments.

IV. PAPER-2 — CNN + BILSTM MULTI-MODAL FUSION FOR CHEST PATHOLOGY [2]

4.1 Approach

Zhang et al. proposed one of the earliest end-to-end multi-modal fusion frameworks implemented for chest disease classification. ResNet-50 was trained on X-ray images to obtain visual feature vectors, and a Bidirectional LSTM was trained on paired radiology report text to obtain sequential semantic representations. The two vectors were merged and subjected to

 ©IJSRED: All Rights are Reserved

two fully connected layers with a final output of multi-label results for 14 pathologies. The BiLSTM sequential modelling differentiates between expressions such as "no evidence of pleural effusion" and "small pleural effusion noted" as two different basic clinical expressions, in contrast to a bag-of-words encoding.

4.2 Key Results

The multi-modal model achieved a mean AUC of 0.883 across all 14 pathologies, compared to 0.851 for image- only ResNet-50 and 0.862 for text-only BiLSTM. The largest gains were seen in visually ambiguous conditions, particularly consolidation and atelectasis, where text context provides discriminating information that the image alone cannot furnish. Table II shows per-pathology AUC scores for a representative subset of conditions.

TABLE II
AUC COMPARISON ACROSS MODEL CONFIGURATIONS FOR SELECTED PATHOLOGIES [2]

Pathology	Image-Only	Text-Only	Multi-Modal
Pneumonia	0.768	0.792	0.831
Atelectasis	0.801	0.814	0.869
Pleural Effusion	0.876	0.881	0.914
Cardiomegaly	0.848	0.839	0.893
Consolidation	0.755	0.779	0.847
Mean AUC	0.851	0.862	**0.883**

4.3 Architectural Insight and Limitations

An ablation study established that neither branch was redundant: removing either branch degraded performance, with the image branch removal causing a more severe decrease. This substantiates the view that although textual and image information partially overlap, each modality contributes genuinely non-redundant information. One weakness identified empirically was that joint training had to be managed carefully, as the CNN and BiLSTM trained at significantly different speed. Structured patient metadata was also not incorporated, leaving the three-way fusion question open for future work.

V. PAPER 3 — CLINICALBERT: DOMAIN SPECIFIC PRE-TRAINING ON CLINICAL NOTES [3]

5.1 Motivation

Standard BERT is pre-trained on Wikipedia and BookCorpus for general English, but fails to handle clinical language. Medical records are full of abbreviations (e.g., SOB for shortness of breath), specialized vocabulary, unconventional spellings, and negation symbols that are systematically misunderstood by general-purpose models. ClinicalBERT addressed this by pre-training BERT on approximately two million clinical notes from MIMIC-III, using the identical masked-language modelling and next-sentence-prediction objectives as the original BERT, but solely using a clinical training corpus.

5.2 Results and Relevance

The major evaluation task consisted of predicting 30-day hospital readmission based on discharge summaries. ClinicalBERT obtained an AUC of 0.714, whereas general BERT achieved 0.700 and bag-of-words logistic regression achieved 0.662. Qualitative error analysis revealed that ClinicalBERT was much more successful in handling negation, uncertainty, and abbreviations—features of critical importance where notes include phrases such as "patient denies fever and chills" or "no acute cardiopulmonary process detected."

ClinicalBERT does not add to the multi-modal literature through a particular architecture, but rather by providing a reliable, reproducible text encoder that later works could use as a drop-in text branch for multi-modal pipelines. Its pre-trained weights are publicly available on HuggingFace, substantially reducing the barrier to entry for groups requiring a clinically-aware text encoder without the resources to pre-train their own.

5.3 Limitations

Only one prediction task and discharge summaries were evaluated. The benefit of ClinicalBERT over general BERT is smaller in fine-tuning data-abundant situations, implying that domain-specific pre-training is most relevant in low-data regimes, which unfortunately characterize medical AI. This work contained no image modality.

VI. PAPER 4 — MULTIMODAL BIOMEDICAL AI: A FIELD-WIDE META-ANALYSIS [4]

6.1 Overview

Rather than proposing a new model, this paper provides a well-organised systematic review of more than 300 published studies on multi-modal medical AI. It classified each study by medical field, combinations of modalities, and fusion strategies, and determined patterns of what worked and what did not. For this review, it acts as a meta-perspective contex- tualising the individual contributions of the other four papers.

6.2 Key Statistical Findings

Across all studied papers, multi-modal models outperformed the best single-modality baseline in 87% of instances. The median AUC gain was 0.042—a clinically meaningful change translating into significantly fewer missed diagnoses in practice. Imaging plus organised clinical data was the most common combination (approximately 45% of reviewed studies), followed by image-text (28%), text-structured data (15%), and pure three-way fusion (image + text + structured) at only 7%—despite it consistently producing the best results, indicating considerable scope for development in that direction.

6.3 Fusion Strategy Trends and Recurring Gaps

The survey indicated that intermediate fusion has replaced both early and late fusion as the most popular approach, due to the increasing availability of high-quality pre-trained encoders

 ©IJSRED: All Rights are Reserved

for both imaging and text. Three limits were observable throughout the reviewed literature:

- Most of the studies were retrospective, with univariate data at a single institution, and so it was challenging to generalise.
- Most papers used inconsistent evaluation metrics, ham- pering cross-study comparison.
- Very few studies addressed model behaviour when one modality is absent at inference time, despite such occur- rences being common in real clinical environments.

VII. PAPER 5 — DEEP MULTI-MODAL FUSION OF IMAGE AND NON-IMAGE DATA [5]

7.1 Scope and Architecture Patterns

Cui et al. reviewed 108 articles published between 2016 and 2022, specifically addressing the combination of image information with non-image information—organised metadata, tabular clinical data, and genomic information—for disease classification and prognosis. Four primary architecture patterns were identified:

- Feature concatenation (~60% of reviewed papers):
  CNN image features are concatenated with dense-network metadata features.
- Attention-guided fusion:
  Non-image features compute weights that act on the image feature map, applying clinical context to influence visual attention.
- Bilinear pooling:
  Captures pairwise multiplicative interactions between modalities.
- Transformer-based cross-modal attention:
  Treats each modality as a token sequence; increasingly dominant in post-2021 work.

7.2 Quantitative Findings

Models combining images with non-image data outper- formed the image-only baseline in 87% of cases, with a mean accuracy improvement of 6.3 percentage points. The enhancement was greater in tasks where metadata had high discriminating strength, such as prognosis prediction where patient age, comorbidities, and past treatments have strong predictive value.

Another finding was that the performance increase was approximately twice as large with the addition of lab values to demographic metadata as with demographics alone, indicating that richer structured data always emanates more diagnostic value than simple patient-descriptive measures.

VIII. CHALLENGES AND OPEN PROBLEMS

The main barriers to real-world clinical implementation lie between current research systems and the positive performance demonstrated in the reviewed literature. These are structural voids that must be sealed prior to responsible clinical deploy- ment.

8.1 Modality and Temporal Synchronisation

The most challenging technical issue in constructing a multi- modal medical dataset is obtaining different types of data describing the same patient state at one point in time. A chest X-ray taken on Day 1 of admission should ideally be paired with a clinical note from Day 1, not a follow-up visit three weeks later. In reality, electronic health records contain notes of any date, images are obtained at irregular times, and lab results may not be temporally aligned with imaging. This problem was either taken into account directly in the five reviewed papers or created challenges in the dataset built.

8.2 Missing Modalities at Inference Time

Training on three modalities does not guarantee all three will be available at deployment. A patient visiting a rural clinic may not have access to MRI. Robust multi-modal systems must cope effectively with missing modalities, either through imputation of the missing signal, probabilistic marginalisation, or random modality dropout during training to approximate missing signals. This remains an active research area without general practical unified solutions.

8.3 Interpretability and Clinical Trust

Even a highly accurate multi-modal model will not lead to clinical action before a clinician understands why the model reached its output. Single-modality interpretability tools—Grad-CAM for images, SHAP for tabular data—are relatively mature. The ability to explain how a model weighted different modalities against each other to achieve a specific prediction is far less developed, and is a severe limitation to clinical adoption.

8.4 Demographic Generalisability

All five reviewed papers presented data primarily from North American or European institutions. Models trained on this type of data may not generalise to populations with other demographics, prevalent diseases, or imaging devices and procedures. The research field urgently needs multi-modal datasets from diverse geographic and institutional contexts, as well as assessment structures that specifically surface subgroup performance disparities.

8.5 Computational Cost

Executing several encoder branches in parallel is compu- tationally costly both during training and inference. While manageable in a research lab, this introduces practical con- straints during deployment to resource-limited healthcare institutions—which are, however, also where enhanced di- agnostic AI is most needed. Efficient fusion architectures, possibly involving knowledge distillation to compress multi- modal models into more lightweight forms, represent a notable open direction.

IX. FUTURE DIRECTIONS

 ©IJSRED: All Rights are Reserved

A number of research directions seem especially promising based on the trajectory of the reviewed papers.

Foundation models for medicine. Generally-purposive encoders developed through large-scale vision-language pre-training, including BioViL, MedCLIP, and LLaVA-Med, are transferring across medical tasks without task-specific fine-tuning. They can be potent and inexpensive starting points for multi-modal classification systems, reducing the need for expensive task-specific dataset curation.

Federated multi-modal learning. Federated learning can mitigate data scarcity and patient privacy limitations; it does not require access to specific patient information during training. Extending federated strategies to multi-modal settings, in which various hospitals may possess different modality capabilities and data formats, is technically arduous but could significantly alter the field.

Temporal multi-modal modelling. Most existing systems assume an independent case per patient interaction. Modelling disease evolution over time by ingesting individual sequences of imaging studies, clinical notes, and lab values may enhance prognosis and forecast relapse, cases in which a prolonged perspective is critical.

Standardised benchmarks. Standardised multi-modal benchmarks analogous to ImageNet in computer vision are urgently needed. Without these standards, the literature remains susceptible to collections of results that cannot be comparatively and synthetically evaluated into consistent progress indicators.

X. CONCLUSION

This review has analyzed five essential contributions to multi-modal medical AI research, spanning dataset construction, text encoder development, fusion architectures, and meta-analysis of the larger literature. Together they provide a solid and coherent empirical argument: when medical AI systems access various types of data—images, clinical notes, and structured metadata—they dramatically outperform systems trained on any single type of data.

The enhancement is not minimal. Multi-modal models exhibited 6–10 percentage point higher accuracy and AUC over the most accurate single-modal models, with differences that are statistically significant across the reviewed literature and that translate into fewer missed diagnoses in practice. The results of the companion parallel project, a three-way

fusion achieving 92.8% accuracy compared to single-modal baselines of 71.2–84.7%, fall well within the improvement percentages recorded throughout this literature and confirm the appropriateness of the approach.

Simultaneously, a number of issues must be tackled before these systems can be responsibly implemented in a clinical setting. Partially unsolved modality alignment, missing-modality robustness, interpretability, and demographic generalisability represent the most pressing gaps. These engineering and validation issues should be a priority for the next phase of the field, as opposed to merely pursuing additional benchmark improvements on existing datasets.

ACKNOWLEDGEMENTS

The authors thank Mr. Kamal Saini (Associate Professor, PIET) for his sustained guidance throughout this project. We are also grateful to Prof. (Dr.) Budesh Kanwar (HOD, AI & DS) and Ms. Alka Rani (Project Coordinator) for their valuable support and feedback. The Department of Artificial Intelligence and Data Science, Poornima Institute of Engineering and Technology, Jaipur, provided the academic environment and computational resources that made this work possible.

XI. REFERENCES

[1] A. E. W. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum,
M. P. Lungren, C. Deng, R. G. Mark, and S. Horng, “MIMIC-CXR: A large, publicly available database of labeled chest radiographs,” Scientific Data, vol. 6, p. 317, 2019. doi: 10.1038/s41597-019-0322-0

[2] Y. Zhang, M. Chen, and L. Liu, “Multi-modal fusion of imaging and clinical notes for automated chest pathology classification,” IEEE Trans. Medical Imaging, vol. 39, no. 9, pp. 2840–2851, 2020. doi: 10.1109/TMI.2020.2983598

[3] K. Huang, J. Altosaar, and R. Ranganath, “ClinicalBERT: Modeling clinical notes and predicting hospital readmission,” arXiv preprint arXiv:1904.05342, 2019. https://arxiv.org/abs/1904.05342

[4] J. N. Acosta, G. J. Falcone, P. Rajpurkar, and E. J. Topol, “Multimodal biomedical AI,” Nature Medicine, vol. 28, pp. 1773–1784, 2022. doi: 10.1038/s41591-022-01981-2

[5] C. Cui, H. Yang, Y. Wang, S. Zhao, Z. Asad, L. A. Coburn, K. T. Wilson,
B. A. Landman, and Y. Huo, “Deep multi-modal fusion of image and non-image data in disease diagnosis and prognosis: A review,” arXiv preprint arXiv:2203.15588, 2022. https://arxiv.org/abs/2203.15588

[6] P. Rajpurkar et al., “Deep learning for chest radiograph diagnosis,” PLOS Medicine, vol. 15, no. 11, p. e1002686, 2018. doi: 10.1371/jour- nal.pmed.1002686

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez,
L. Kaiser, and I. Polosukhin, “Attention is all you need,” Advances in Neural Information Processing Systems (NeurIPS), vol. 30, 2017.

[8] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), pp. 770–778, 2016.

[9] X. Pei, K. Zuo, Y. Li et al., “A review of the application of multi- modal deep learning in medicine: Bibliometrics and future directions,” Int. J. Computational Intelligence Systems, vol. 16, p. 44, 2023. doi: 10.1007/s44196-023-00225-6

[10] K. Yan, T. Li, J. A. L. Marques, J. Gao, and S. J. Fong, “A review on multimodal machine learning in medical

 ©IJSRED: All Rights are Reserved

diagnostics," Mathematical Biosciences and Engineering, vol. 20, no. 5, pp. 8708–8726, 2023. doi: 10.3934/mbe.2023382

 ©IJSRED: All Rights are Reserved