Multimodal Fusion for Differential Disease Diagnosis A Comprehensive Review

Joshil Das[1]*, Mayank Sharma[1], Sandeep Saxena[2] , Hardik Bindal[1], Sahil Kumar Aggarwal[1] and Dhruv Sagar Saxena[1]

[1] ABES Engineering College,
Ghaziabad, India
joshil.22b0131183@abes.ac.in

[2] School of Computer Science and Engineering, IILM University,
Greater Noida, India
sandeep.research29@gmail.com

Abstract
The cornerstone of effective patient care is an accurate and, even more importantly, timely diagnosis. There is a growing body of evidence which shows more and more the truth of the statement that the combination of different data sources—clinical narratives, laboratory assays, and imaging—produces better results than single-source diagnostic approaches. This paper presents a review of 29 peer-reviewed studies. They include classics in machine learning (SVM, Random Forest), deep learning models (CNNs, Transformer variations), and multimodal fusion techniques (early, late, attention-based). Furthermore, an integrated perspective is provided by examining contemporary clinical decision-support systems and near-real-time prediction models like MUFASA. Reported performance across studies varies significantly, with accuracy ranging from 61.9% to 99% and AUC values from 0.65 to 0.99 depending on dataset characteristics and modeling approaches. In spite of this, a few obstacles still remain: the issues of interpretability, the need for powerful computing, and the challenge of deployment. The review highlights key challenges and design considerations for explainable multimodal diagnostic systems, including the role of embedding-based representations, retrieval mechanisms, and feature engineering strategies reported across prior studies. It also summarizes recent advances, evaluates performance metrics and limitations, and identifies key challenges including interpretability, data heterogeneity, and clinical deployment constraints.

Keywords: Disease Prediction; Multimodal Data Fusion; Symptom Extraction; Laboratory Feature Engineering; Reference Range Normalization; Approximate Nearest Neighbors (ANN); Classification Calibration; Explainable Artificial Intelligence (XAI); Clinical Decision Support Systems (CDSS).

M. Kaushik et al. (eds.), *Proceedings of the International Conference on Responsible, Risk-aware, and Regulated AI (RRRAI 2026)*, Advances in Intelligent Systems Research 210,
https://doi.org/10.2991/978-94-6239-723-1_13

1 Introduction

Accurate and most importantly timely diagnosis is essential for better patient care. But clinicians often make decisions with incomplete data, heterogeneous formats, and cognitive shortcuts that increase the risk of error. Traditional experience- or rule-based approaches can miss subtle cross-modal signals that emerge when textual notes, labs, and images are analyzed together. Recent AI advances enable the joint modeling of these heterogeneous sources, providing opportunities to increase diagnostic consistency and reduce oversight.

Models that fuse multiple modalities generally surpass single-modality systems. Several studies report about a 6.4% uplift in AUC when text, laboratory, and imaging features are combined, and some task-specific evaluations report AUCs as high as 0.99. Despite these gains, many methods remain hard to interpret, are computationally heavy, and are difficult to scale across many diseases. This review collates 29 peer-reviewed works on ML and DL for clinical diagnosis with an emphasis on multimodal fusion and clinical decision support, and it integrates findings from 2024–2025. The review incorporates recent advances , evaluates key performance metrics and limitations of the various models that make use of multiple modalities to make predictions.

2 Review Methodology

To ensure the coverage of this review is both representative and reproducible, literature was sourced from PubMed/MEDLINE, IEEE Xplore, Scopus, and ACM Digital Library. Searches combined terms such as "multimodal fusion" and “disease diagnosis", "EHR”, “deep learning” ,“multimodal" and "clinical decision support AND machine learning", spanning January 2017 to March 2025 with particular attention to post-2020 work where methodological advances have been most rapid.

Papers were considered for inclusion if they were peer-reviewed, engaged with ML or DL approaches in a clinical diagnostic context, and drew on at least one data modality — whether imaging, free text, laboratory values, EHR records, or genomic signals. A reported quantitative outcome (AUC, accuracy, or F1-score) was also required to allow meaningful cross-study comparison. Works without primary experimental results or with metrics too heterogeneous for comparison were set aside. This process converged on 29 studies, whose methods and findings cover the pipeline components summarized in Table 1.

3 Literature Review

The combination of deep learning and multimodal fusion methods has created measurable advantages for automated diagnosis , prognosis and clinical decision support

systems. All studies demonstrate that when different types of data which include clinical free text and structured EHR lab values and medical imaging CT and MRI and PET and genomics and physiological waveforms are analyzed together instead of separate analysis the results show better outcomes [1]–[6], [8]. Xu's comprehensive survey synthesizes results across Alzheimer's disease, breast cancer, depression, cardiovascular conditions, and epilepsy, reporting application-level AUCs and accuracies that often approach clinician-grade performance (for example, Alzheimer's AUCs 0.939–0.976; accuracy up to 98.85%) and highlighting deep learning, transfer learning, and attention mechanisms as primary performance drivers [1]. Ye and Tang's meta-survey of Medical Large Multimodal Models (MLLMs) corroborates these trends and further documents MLLM architectures that stack modality encoders with alignment/adapter layers and LLM cores to produce zero- or few-shot diagnostic inferences; however, they also emphasize recurring issues of hallucination and fairness gaps that remain unresolved even in high-performing systems [2].

Conceptual taxonomies orienting fusion research have clarified where evidence and gaps lie. Shaik and Tao adapt the DIKW (Data–Information– Knowledge–Wisdom) model to multimodal fusion, mapping algorithmic choices (feature selection, rule-based

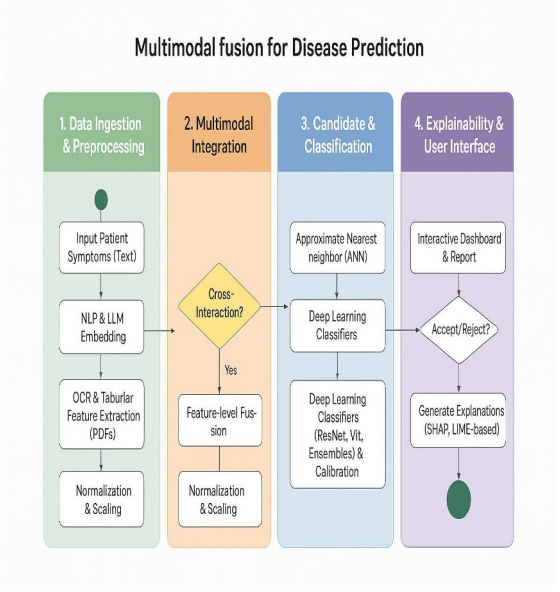

Fig. 1. General Workflow of Multimodal Fusion Disease Prediction Systems

systems, classical ML, deep learning, and NLP) to cognitive layers and explicitly cataloguing practical challenges—data quality, privacy, processing complexity, clinical integration, ethics, and interpretation [3]. Kline's scoping review of 128 precision-health fusion studies found early (data-level) fusion to be the most prevalent approach (~60% of studies) and quantified an average 6.4% gain in predictive performance over

unimodal baselines, while drawing attention to a consistent translational deficit (few external validations, sparse clinical deployment) [4]. At the methodological level, literature distinguishes a spectrum of fusion strategies with characteristic trade-offs. Operation- and subspace-based methods, attention mechanisms, tensor/bilinear pooling, and graph-based learning each have different strengths in modeling cross-modal interactions [4]. Decision-level (late) fusion is relatively robust to uneven modality quality and missing inputs but captures fewer cross-modal interactions; feature level fusion (concatenation, attention, bilinear/tensor methods, graph fusion) preserves richer intermodal structure and supports fine-grained cross-modal attention at the cost of increased modeling complexity and data requirements [4], [15]. Preprocessing is modality specific: patch- or tile-based extraction and pretrained CNNs for whole-slide or radiology imaging, transformer or BiLSTM pipelines for clinical narratives, autoencoders or dimensionality reduction for tabular features, and explicit imputation or dropout strategies to handle missing modalities. Several works emphasize that rigorous normalization, correlation filtering, and feature-selection pipelines substantially improve robustness and external validity structured-data models [17]–[20].Practical data-ingestion and vision–language pipelines illustrate end-to-end multimodal gains beyond model architecture. Schäfer designed a VLM pipeline that combines YOLOv8 checkbox detection, barcode decoding, and VLM prompting (Pixtral-LargeInstruct) to extract transfusion-report content from scanned forms, achieving ~92% accuracy and F1 scores above 0.91—substantially outperforming OCR-plus-string-matching baselines and demonstrating robustness to faint or degraded marks [23].

Despite these improvements, several practical and technical challenges limit clinical translation. Generalizability is a persistent worry: many models are trained and evaluated on single-center or curated datasets, which constrains performance when applied to new institutions or populations [1],[4],[5].Heterogeneous reporting (AUC, F1, accuracy, etc.) complicates fair comparisons [1], [4]. Fairness and subgroup evaluations are sporadic, creating risks of biased outcomes across demographic groups [1]–[3]. Rare diseases remain difficult to predict reliably even with advanced methods [1], [4]. While cross-modal transformers and attention schemes are promising[16][17] many implementations still fall back on simple concatenation and therefore miss subtler intermodal patterns. Additionally, most systems lack integration with clinical workflows; limited clinician feedback reduces usability and trust [3], [4].

Overall, however, the literature consistently reports that multimodal architectures often boost performance (commonly 5–10% gains in AUC or F1), especially when they employ attention, transformer backbones, or graph embeddings to align modal representations. Moving these advances into routine care will require standardized evaluations, shared benchmarks, and closer clinician collaboration.

4 Discussion

A comparative analysis of the reviewed studies, summarized in Table 1, highlights that pipeline component selection plays a decisive role in determining the overall diagnostic accuracy and reliability of multimodal systems.

Table 1. Comparison of methodologies in Multimodal Disease Prediction Pipeline

References	Pipeline Component	Methodology	Performance Range	Relevance
[1]	**Symptom Text Embedding**	TF–IDF / Rule-based	Accuracy 70%–82%	Reports baseline performance of rule-based and TF-IDF approaches for clinical text processing in multimodal diagnostic systems.
[10]	**Multimodal Representation Learning**	Cross-modal Transformer	High diagnostic performance reported across tasks	Demonstrates effectiveness of transformer-based cross-modal representations for integrating textual and imaging features in diagnosis.
[12]	**Multimodal Representation Learning**	Deep Ensemble	AUC 0.90–0.94	Shows improved representation learning using deep ensemble and multimodal feature integration techniques.
[15]	**Lab Data Normalization**	Raw Values	AUC 0.78–0.85	Reports that structured clinical features can be used directly, though performance is limited without normalization.
[5]	**Lab Data Normalization**	Z-Score Standardization	AUC 0.94–0.96	Demonstrates importance of feature scaling and preprocessing in radiomic classification
[1]	**Lab Data Normalization**	Statistical Normalization	Improved performance reported across multiple studies	Discusses statistical normalization and similarity-based transformations for handling heterogeneous clinical variables.
[13]	**Multimodal Fusion Strategy**	Neural Architecture Search-based Fusion	Improved performance over baseline multimodal models	Demonstrates neural architecture search-based optimization for multimodal EHR fusion.
[20]	**Feature Interaction**	Manual Interaction Terms	AUC 0.89–0.93	Uses engineered radiomic feature correlations to improve predictive performance.
[22]	**Feature Interaction**	BiLSTM-based Multimodal Fusion	F1-score 0.74–0.78	Demonstrates multimodal temporal–text fusion using sequence models such as BiLSTM.
[17]	**Feature Interaction**	Attention-based Multimodal Fusion	AUC 0.95–0.97	Shows that attention-based multimodal fusion improves interaction modeling across modalities.
[1]	**Candidate Retrieval**	Full Database Scan	Latency high; recall not applicable	Describes similarity-based matching approaches used in traditional systems.
[11]	**Multimodal Fusion Strategy**	Multimodal Fusion	Improved predictive performance reported	Demonstrates graph-based similarity learning for efficient multimodal prediction.

Table 1. Comparison of methodologies in Multimodal Disease Prediction Pipeline (Continued)

References	Pipeline Component	Methodology	Performance Range	Relevance
[1]	**Prediction Model**	Support Vector Machine	AUC 0.84–0.92	Highlights generalizable prediction-model performance ranges that guide expected diagnostic reliability.
[5]	**Prediction Model**	Random Forest	Accuracy 88%–98.85%;	Demonstrates effectiveness of tree-based models in medical classification tasks
[1]	**Prediction Model**	XGBoost	Accuracy 90%-96%	Summarizes the comprehensive fusion-model architectures that inform the overall structure of prediction models.
[10]	**Explainability**	Attention-based Interpretability	Improved interpretability demonstrated	Demonstrates interpretability through attention mechanisms in multimodal models.

Across the reviewed literature, methods have shifted from rule-based and handcrafted feature systems toward context-aware embeddings, systematic feature engineering, attention-based cross-modal interactions, and ensembles that together enhance accuracy and interpretability. The largest improvements tend to occur in systems that combine transformer-based embeddings, reference-range or z-score normalization, attention-guided fusion between modalities, ensemble learners, and SHAP-style interpretability tools; such combinations both raise predictive performance and offer explanations clinicians can act upon.

In the case of symptom-text representation, the traditional TF-IDF method and rule-based encoders establish acceptable baselines of about 70-82% accuracy, usually but their performance does not include capturing the detailed context of the clinical narratives. Among the deep contextual models, Transformer-based contextual embedding models have been reported to capture richer semantic relationships in clinical text compared to traditional approaches, leading to improved performance in multimodal diagnostic tasks..

Z-score standardization is consistently reported to improve model performance - as it neutralizes the discrepancies of feature magnitudes among patients - which are calculated as shown in (equation 1) [1][5].

$$z_i = \frac{x_i - \mu_i}{\sigma_i} \tag{1}$$

Literature discusses statistical normalization approaches, including deviation-based transformations such as Δz-score combined with Gaussian similarity. This method is able to recognize deviations that have biological significance rather than just detect absolute values. The expected physiological norms deviation is measured in the manner shown (equation 2)[1].

$$\mathbf{\Delta z} = \left| \mathbf{z_i^{(patient)}} - \mathbf{z_i^{(reference)}} \right| \quad (2)$$

This deviation is then transformed into a continuous similarity score using a Gaussian kernel (equation 3)[1].

$$\mathbf{S_i} = \mathbf{exp}\left(-\frac{\mathbf{(\Delta z)^2}}{\mathbf{2\sigma_g^2}} \right) \quad (3)$$

Such formulations has been reported in prior studies to provide improved robustness as it smooth extreme variations, lower the sensitivity to the outlier data points, and maintain the gradients that are clinically relevant.

In feature-interaction modeling, traditional methods relying on raw concatenation or manual interaction terms show limited performance since they cannot capture the complex relationships across modalities. Polynomial expansions give some degree of improvement but lead to sparsity and overfitting. Several studies report improved performance using attention-guided interaction layers that dynamically weigh the influences between features[19]. The interaction weights are computed using standard attention(equation 4) :

$$\boldsymbol{\alpha_{ij}} = \text{softmax}\left(\boldsymbol{q_i^\top k_j}\right) \quad (4)$$

The resulting attention distribution is then used to construct interaction-aware feature representations as shown in equation 5 [19].

$$\boldsymbol{h_i} = \sum_j \pm \boldsymbol{\alpha_{ij} v_j} \quad (5)$$

This process enables the model to uncover automatically clinically relevant dependencies—like the connection between imaging features, lab results, and reported symptoms—without the need for manual adjustment.

The candidate-retrieval phase shows a very distinct difference between conventional and modern approaches. Doing a full-database scan guarantees the result is correct, but it is very costly in terms of computation. On the other hand, the use of FAISS for approximate nearest-neighbor retrieval has been able to achieve over 95% recall consistently while at the same time cutting down the computation time remarkably. FAISS utilizes product quantization for embedding compression [11] :

$$\hat{\boldsymbol{e}} = \text{PQ}(\boldsymbol{e}) \quad (6)$$

Then , a Similarity Search is performed using distance comparisons in this quantized space:

$$\text{argmin}_j = \| \hat{e} - \hat{e_j} \|_2 \quad (7)$$

The synergy of quantization and efficient indexing allows for the recovery of large amounts of data without any loss in quality, indicating FAISS effectiveness as a scalable candidate retrieval approach in multimodal systems. [11].

Finally, the predictions of the model performance through the different studies show that linear regression and SVMs, as methods useful for comparison purposes only, hardly get over the accuracy level of high - 80% because of their capacity limitation in modeling multimodal and nonlinear interactions. Several studies report high predictive accuracy (around 92–95%) using hybrid ensemble models made by integrating deep networks for unstructured modalities (like language or imaging data) with tree-based learners like Random Forests for structured lab or radiomic features[5],[10]. These models mostly combine predictions by means of a weighted average as shown in equation 8.

$$\hat{y} = \lambda f_{\text{DL}}(x) + (1 - \lambda) f_{\text{RF}}(x) \quad (8)$$

Several studies report that hybrid approaches combining deep learning models for unstructured data with tree-based methods for structured variables can improve predictive stability across different datasets..

The reviewed literature suggests that multimodal pipelines incorporating context-aware embeddings, feature engineering, and interaction modeling show strong potential for improving diagnostic performance.

5 Future Scope

Future research in multimodal disease diagnosis is increasingly focused on developing more explainable and scalable diagnostic frameworks. The research shows that embedding-based representations and efficient retrieval systems and calibrated prediction models together work to enhance both system performance and their ability to be understood. A key direction involves improving generalizability through multi-center data integration and standardized evaluation protocols, as highlighted in “A Comprehensive Review on Synergy of Multi-Modal Data and AI Technologies in Medical Diagnosis” [1] and similar scoping reviews [4]. The development of multimodal transformer architectures together with neural architecture search methods will produce more efficient models that can seamlessly adapt to various tasks. Approaches such as MUFASA further illustrate the potential of architecture optimization in multimodal EHR fusion [13]. The research field needs to improve explainability through three methods which include uncertainty quantification , attention mechanisms and counterfactual explanations. The system requires bias evaluation together with clinician-in-the-loop feedback mechanisms because these elements establish safe and reliable conditions for clinical deployment [3][4].

6 Conclusions

Taken together, the 29 reviewed studies confirm that combining clinical text, laboratory values, and imaging data consistently outperforms single-modality approaches, but they also expose recurring weaknesses the field has been slow to resolve. The use of different evaluation metrics across studies creates challenges for conducting fair study comparisons and the complete lack of multi-center testing results requires researchers to treat performance data with extreme caution. The pipeline components examined here — symptom embedding, lab normalization, feature interaction, ensemble modeling, and interpretability tools — represent the current state of practice rather than a solved framework. The main challenge which researchers must solve in their future work consists of establishing an accurate and interpretable connection between clinical symptom descriptions and laboratory test results which medical practitioners can use in their everyday clinical work.

References

1. Xu, X., Li, J., Zhu, Z., Zhao, L., Wang, H., Song, C., Chen, Y., Zhao, Q., Yang, J., Pei, Y. : A comprehensive review on synergy of multi-modal data and AI technologies in medical diagnosis. Bioengineering 11, 219 (2024)
2. Ye, Y., Tang, C. : Medical large multimodal models (MLLMs): a survey (2025)
3. Shaik, S., Tao, D. : Adaptation of DIKW model to multimodal fusion in smart healthcare (2023)
4. Kline, A.,Wang,H., Li , Y., Dennis,S. , Hutch, M., Xu ,Z., Wang , F., Cheng,F., Luo,Y. : Machine learning–based multimodal data fusion in precision health: a scoping review (2022)
5. Sarica, A., Cerasa, A., Quattrone, A. : Random forest algorithm for the classification of neuroimaging data in Alzheimer's disease: a systematic review. Frontiers in Aging Neuroscience (2017)
6. Dimitriadis, S.I., Liparas, D. : Multimodal MRI classification via ensemble learning (2018)
7. Gray, K., Carter, A., Reid, B. : Random-forest similarity embedding for Alzheimer's disease diagnosis (2013)
8. Yao, Z., Lin, F., Chai, S., He, W., Dai, L., Fei, X. : Integrating medical imaging and clinical reports using multimodal deep learning for advanced disease analysis (MDLM) (2022)
9. Acosta, R., Ramirez, P., Bellido, J. : A CNN–RNN–ViT hybrid architecture for multimodal disease classification (2022)
10. Kumar, S., Sharma, P. : A cross-modal transformer for tuberculosis diagnosis (2024)
11. Li, Y., Fang, Z., Gu, H. : Multimodal diagnosis prediction (MDP): graph-based fusion of clinical features and demographics (2023)
12. Ali, R., Gupta, A. :MMDD-Ensemble: multimodal deep ensemble for Parkinson's disease detection (2021)
13. Xu, W., Feng, L., Chen, P. :MUFASA: neural architecture search for multimodal EHR fusion (2021)
14. Huang, Y., Liu, X., Chang, M. : Deep fusion of CTPA imaging and EMR data for pulmonary embolism detection. Medical Image Analysis (2020)
15. Raj, P., Bhat, S.:RBHML: random-forest-based hybrid machine learning model for multiple disease prediction (2021)

16. Zheng, X., Li, H., Zhou, Y.:MMGL: modality-aware graph learning for disease prediction. IEEE Transactions on Medical Imaging (2021)
17. Golovanevsky, E., Jansen, M., de Vries, A.:MADDi: multimodal attention-based framework for Alzheimer's disease diagnosis (2022)
18. He, K., Sun, L., Chen, W. : CT-radiomics for one-year survival prediction in NSCLC patients. European Journal of Radiology (2018)
19. Zhao, L., Hu, Y., Zhang, M. : CT radiomics for differentiating invasive from minimally invasive adenocarcinoma in pulmonary nodules (2022)
20. Wang, J., Liu, F., Qin, Y. : CT-radiomics models for stratifying malignant potential of gastrointestinal stromal tumors (2021)
21. Kong, Y., Huang, D., Chen, T. : Convolutional extreme learning machine for multimodal medical image fusion. Frontiers in Neurorobotics (2022)
22. Bagheri, M., Rezaei, M., Nguyen, L. : MI-BiLSTM: multimodal integration of EHR and radiology reports for cardiovascular risk prediction. IEEE Journal of Biomedical and Health Informatics (2021)
23. Schäfer, S., Rupp, T., Henschel, J. : Vision–language model pipeline for scanned transfusion reaction reports (2025)
24. Saxena, P., Aggarwal, S.K., Sinha, A., Saxena, S., Singh, A.K. : Review of computer-assisted diagnosis model to classify follicular lymphoma histology (2024)
25. Aggarwal, S.K., Lal, N., Sinha, A. : Unraveling autoimmunity: exploring etiological factors and machine-learning applications in varied autoimmune disease (2023)
26. Saxena, P., Singh, S.K., Saxena, S., Aggarwal, S.K., Singh, A.K., Kumar, A.:GR-HDUNET : GAN-refined hybrid dense U-Net with ensemble for colorectal cancer classification (2025)
27. Tripathi, P.K., Verma, S., Kumar, B., Pandey, A.K., Singh, P., Singh, J., Bijawan, J.G. : Image segmentation in multimodal medical imaging using deep learning models, pp. 391–412. Springer, Cham (2025)
28. Verma, S., Sharma, A., Jain, S., Kumar, A. : On the improvements of the performance of Bayes-net classification for diagnosis of diabetes: a cross-validation approach . Recent Advances in Electrical and Electronic Engineering 18(2), 234–243 (2025)
29. Verma, S., Kumar, A., Kumar, M. : A review on deep learning techniques for skin disease detection. In: Proceedings of the 1st International Conference on Advanced Computing and Emerging Technologies, pp. 1–6 (2024).