Data-Centric, Robust, and Explainable Multimodal Deep Learning for Clinical Decision Support: A Systematic Review

Md Mazharul Islam[a,*], Abrar Mohammed Tanzim Alam[a], Md. Sadikur Rahman Rony[b] and Susmay Das[c]

[a]Department of Electrical and Computer Engineering, North South University, Dhaka, Bangladesh
[b]Department of Statistics, University of Dhaka, Dhaka, Bangladesh
[c]Department of Computer Science and Engineering, Dhaka International University, Dhaka, Bangladesh

ARTICLE INFO

Keywords:
multimodal deep learning
clinical decision support
data-centric artificial intelligence
explainable artificial intelligence
external validation
systematic review

ABSTRACT

Purpose: Multimodal deep learning is increasingly proposed for clinical decision support (CDS), and a "data-centric" framing–prioritizing label quality, missing-modality robustness, distribution shift, calibration, and explainability–has gained rapid traction. Prior reviews have examined multimodal medical AI, CDS, and data-centric methods separately, but none addresses their intersection, leaving it unclear whether this literature is deployment-ready or remains proof-of-concept. We aimed to map the modalities, fusion strategies, and data-centric and explainability techniques used, to quantify how often each is implemented rather than merely mentioned, and to assess external validation, clinical-outcome measurement, equity, and reporting.

Methods: Following the PRISMA 2020 statement (PROSPERO CRD420261427815), we screened 150 records and included primary, clinical, multimodal studies applying machine or deep learning to a decision-support task. Each study was coded for modalities, fusion, five data-centric techniques, explainability, validation strategy, measured outcomes, equity analysis, and reporting standards, separating techniques that were implemented or evaluated from those only mentioned. Synthesis was narrative.

Results: Thirty-nine studies met inclusion; 38 (97%) were published 2024–2026, with a median of three modalities (range 2–7), most commonly structured EHR (56%) and imaging (49%). Data-centric techniques were reported frequently: label-noise handling (85%), distribution-shift handling (79%), calibration and missing-modality handling (77% each), class-imbalance handling (74%), and equity analysis (62%). However, external validation was reported in only 4/39 studies (10%), a patient, clinician, or system outcome in 3/39 (8%), no study reported routine deployment, and prediction-model reporting standards were essentially absent.

Conclusion: Current evidence establishes proof-of-concept but not proof-of-benefit: the field is rich in robustness and explainability techniques yet largely untested out-of-distribution and against clinical outcomes. It requires a shift toward external multi-site validation, outcome measurement, and adherence to AI reporting standards (e.g., TRIPOD+AI) before deployment can be justified.

1. Introduction

Clinical reasoning integrates heterogeneous evidence—medical imaging, structured electronic health records (EHRs), physiological signals, clinical text, and increasingly genomic and wearable data. Multimodal deep learning, which models these streams jointly, is therefore an intuitively attractive substrate for clinical decision support (CDS), and reported gains over unimodal baselines have driven rapid uptake [1, 2]. A recent scoping review of multimodal medicine maps the same expansion across imaging, text, and structured data [3]. In parallel, the locus of methodological attention has shifted from architecture alone toward a data-centric view, in which label quality, missing or asynchronous modalities, class imbalance, distribution shift, and predictive uncertainty are treated as first-class design problems rather than preprocessing afterthoughts [4, 5]. Distribution shift in particular is a recurring threat to transportability [6]. Explainability has become a near-mandatory companion, motivated by clinician trust, accountability, and emerging regulatory expectations [7, 8].

These three threads–multimodality, data-centric robustness, and explainability–are increasingly invoked together as the headline contribution of a study. Yet the presence of a robustness technique does not establish that a model is reliable in deployment. Internal validation routinely overstates real-world performance, and robustness demonstrated on a single development dataset may not survive the distribution shifts introduced by new sites, devices, and populations [9]. Reporting and appraisal standards for clinical prediction models–TRIPOD+AI, DECIDE-AI, and PROBAST+AI–exist precisely to separate demonstrated benefit from technical promise [10, 11], and are complemented by Good Machine Learning Practice and model-appraisal tools [12, 13]. Recent systematic reviews in adjacent clinical-AI domains have repeatedly found favorable model metrics coexisting with absent external validation and unmeasured patient benefit [14, 15], and the broad multimodal scoping review above reached similar conclusions about the gap between technical and clinical maturity [3].

*Corresponding author.
mazharul.islam1@northsouth.edu (M.M. Islam);
abrar.alam01@northsouth.edu (A.M.T. Alam);
mdsadikurrahman-2022116925@stat.du.ac.bd (Md.S.R. Rony);
susmoydas1000@gmail.com (S. Das)
ORCID(s): 0000-0001-6225-1487 (M.M. Islam)

No systematic review has yet examined the specific intersection of data-centric robustness, explainability, and multimodality in CDS, or asked whether the robustness this literature claims is actually validated. We address that gap. We pose three research questions:

(RQ1) Which modalities, fusion strategies, and data-centric and explainability techniques do primary multimodal CDS studies employ?

(RQ2) How often is each technique implemented or evaluated rather than merely mentioned?

(RQ3) To what extent do these studies provide deployment-relevant evidence–external validation, measured clinical outcomes, equity analysis, and standardized reporting?

To our knowledge, this is the first review to synthesize this intersection and to apply an implementation-versus-mention coding lens, which we show is essential to interpreting the field honestly.

2. Methods

2.1. Protocol and registration

This review was conducted and reported in accordance with the PRISMA 2020 statement [16]. The review was registered with PROSPERO (code ID: CRD420261427815) [17]. The completed PRISMA 2020 checklist is provided as Supplementary Material.

2.2. Eligibility criteria

Studies were eligible if they (i) were primary studies developing or evaluating a model or system, not reviews, surveys, editorials, protocols, or bibliometric papers; (ii) were multimodal, combining two or more distinct data modalities (imaging, structured EHR/tabular, physiological signals/time-series, clinical text, genomics/omics, or wearable/sensor) within a single model or system; (iii) applied machine or deep learning to a clinical decision-support task (diagnosis, prognosis, risk stratification, triage, deterioration, or monitoring) in humans; and (iv) reported at least one quantitative result. Non-clinical applications, single-modality studies, and secondary literature were excluded.

2.3. Information sources and search strategy

A search was assembled across major bibliographic databases indexing clinical and biomedical computing literature (PubMed, Scopus, IEEE Xplore, and Web of Science), combining three concept blocks–multimodal/data-centric deep learning; robustness and explainability (missing modality, label noise, distribution shift, calibration, interpretability); and clinical decision support–using controlled-vocabulary and free-text terms joined with Boolean operators. The full search strings are provided as Supplementary Material. Records were consolidated and de-duplicated, yielding 150 unique records for screening.

2.4. Study selection

Records were screened against the eligibility criteria at title/abstract and, where required, full-text level. Screening decisions and exclusion reasons were recorded for every record to populate the PRISMA flow and are provided as Supplementary Material.

2.5. Data extraction and coding

Included studies were coded on a standardized form capturing bibliographic data, clinical task and setting, modalities and modality count, datasets, AI paradigm, fusion strategy, five data-centric techniques (missing-modality handling, label-noise/data-quality handling, class-imbalance handling, distribution-shift handling, calibration/uncertainty), explainability method, validation strategy, external validation, primary metric, deployment status, measured clinical/provider outcome, equity analysis, and reporting standard. For the data-centric, reliability, and outcome items we applied a strict coding rule: an item was coded present only when the study implemented or empirically evaluated it, and absent when it was only mentioned as a challenge, gap, or future direction. This distinction is necessary because, in an emerging field, naming a problem is frequently mistaken for solving it.

2.6. Evidence-quality appraisal and synthesis

Because designs and outcomes were heterogeneous, we used narrative synthesis with descriptive tabulation; no meta-analysis was performed. We appraised evidence quality through deployment-readiness indicators extracted uniformly across studies—external validation, calibration reporting, measured clinical outcomes, equity analysis, and citation of a prediction-model reporting standard–which together characterize the maturity and transportability of the evidence. Findings are reported as the proportion of included studies addressing each item.

3. Results

3.1. Study selection

Of 150 unique records screened, 111 were excluded: 89 were reviews or other secondary literature, 14 were single-modality, and 8 were non-clinical applications. Thirty-nine primary, clinical, multimodal studies met all inclusion criteria and were synthesized (Fig. 1). All included studies are listed individually, with their clinical area, modality count, AI approach, validation strategy, and headline result, in Table 1; the complete extraction is provided in Supplementary Table S1.

3.2. Characteristics of included studies

The corpus is recent: 38/39 studies (97%) were published in 2024–2026 (Table 2), reflecting the rapid emergence of the data-centric multimodal framing. The median number of modalities was three (range 2–7); 23% combined two modalities and 18% combined five or more. Structured EHR/tabular data (56%) and imaging (49%) were the most frequently used modalities, followed by physiological

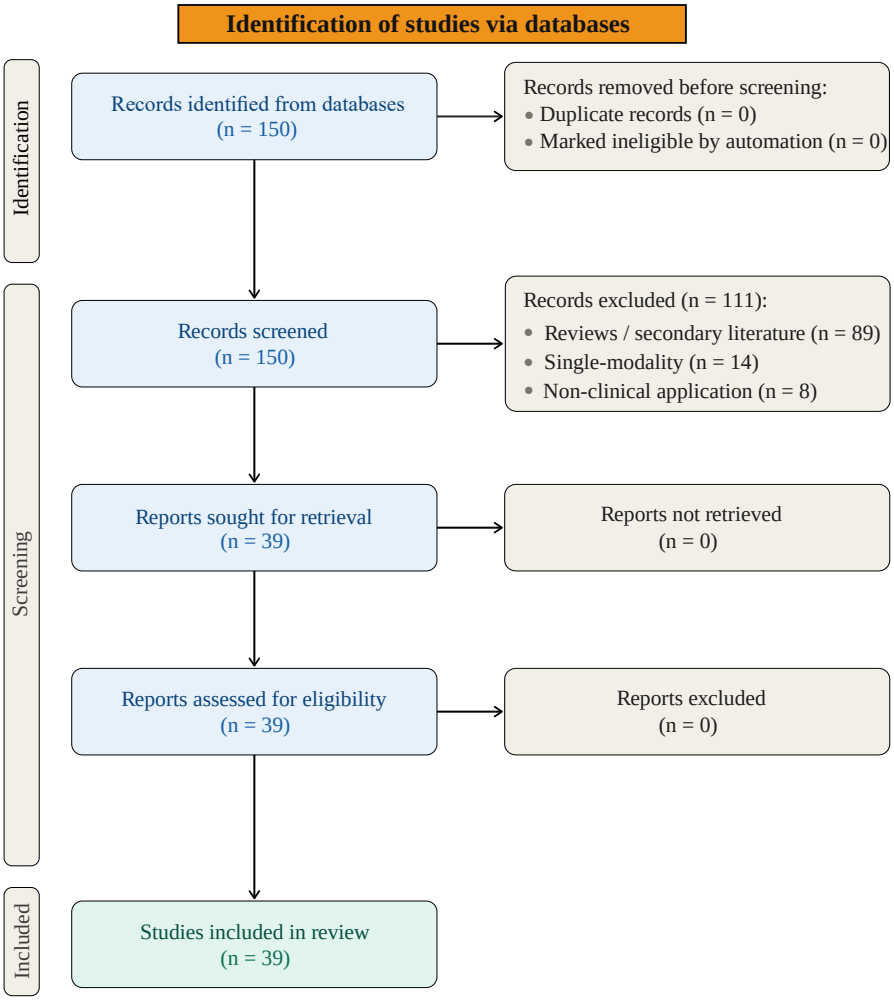

Figure 1: PRISMA 2020 flow diagram of the study selection process.

signals (31%), clinical text and wearable/sensor data (26% each), and genomics/omics (13%). Applications spanned diagnosis, risk stratification, deterioration, and longitudinal monitoring across multiple specialties.

3.3. AI paradigms, fusion, and explainability

Model families were dominated by deep architectures (Table 3). Convolutional networks were the most frequently used component (33%), followed by ensemble or gradient-boosted models (28%), recurrent or long short-term memory networks (23%), and transformer- or attention-based models (21%); large language or generative-AI components and federated-learning frameworks each appeared in 13% of studies, and several studies combined two or more families. Fusion strategy was inconsistently reported: early or feature-level concatenation was most common (33%), with attention- or intermediate-fusion designs in 18% and late or decision-level fusion in 10%; the remainder did not specify a fusion strategy or were conceptual. Explainability was near-universal but overwhelmingly post-hoc: SHAP (56%) and attention visualization (28%) predominated, followed by LIME (23%) and gradient- or saliency-based maps (23%), with integrated gradients (5%) and counterfactual explanations (8%) rare.

3.4. Data-centric and reliability techniques

Data-centric techniques were reported frequently (Table 4): label-noise or data-quality handling was the most common (85%), followed by distribution-shift handling (79%), calibration or uncertainty quantification and missing-modality handling (77% each), and class-imbalance handling (74%). Equity or subgroup analysis was reported in 62%. These high rates must be read in light of the search strategy, which preferentially retrieves studies that foreground such techniques, and of the coding rule above: a present code indicates that a technique was implemented or evaluated, not that robustness was rigorously demonstrated out-of-distribution.

3.5. Explainability

Explainability was near-universal, consistent with a corpus that self-identifies as explainable. Post-hoc attribution dominated—SHAP and LIME, gradient- and saliency-based

Table 1
Characteristics of the 39 included studies. Modality count "Mult." denotes a multimodal pipeline for which a single integer count was not reported. Validation is the highest level reported (External > Internal). Primary results are the headline figures stated by the original authors and are not directly comparable across heterogeneous tasks and reference standards.

Study	Clinical area	Mod. (n)	AI approach	Validation	Primary result
Van Dievoort et al. [18]	Mental health (stress)	3	Generative AI, XAI	Internal (co-design)	Conceptual (dashboard)
Perwej [19]	Multi-domain prognostics	Mult.	Ensemble, deep learning	Internal (CV)	Accuracy 89%
Kovács [20]	Prostate cancer (radiology)	3	CNN (nnU-Net)	External	AUROC 0.82
Sowmya et al. [21]	Multi-disease diagnostics	4	Federated, attention CNN	Internal	Accuracy 95%, AUC 0.96
Weiskirchen [22]	Gastroenterology	5	Multimodal AI (commentary)	None (conceptual)	Conceptual
Tummuri [23]	ICU / monitoring	4	GAN, transformer	Internal (split)	Accuracy 95%, AUC 0.97
Komninos et al. [24]	Sepsis / ICU	3	Deep clustering, LSTM	Internal (split)	AUROC 0.80
K.S.Ranjith et al. [25]	Cardiovascular disease	6	Ensemble, generative AI	Internal	Accuracy 97%
Mohanty et al. [26]	Multi-sector (healthcare)	Mult.	ML, DL, AutoML	Internal	Improved accuracy (NR)
Makesh et al. [27]	General diagnosis	3	Federated, transformer	Internal	Accuracy 94%, AUC 0.96
Ghantasala [28]	Pediatric type-1 diabetes	3	LSTM, LLM copilot	Internal (patient-wise)	MAE 10.9 mg/dL
Hossain [29]	Precision oncology	6	Transformers, GNN	Internal	C-index 0.72–0.75
Abualigah et al. [30]	Translational medicine	4+	GBM, DNN	Internal (CV)	AUROC 0.95–0.96
Noviandy et al. [31]	Oncology (staging)	2	Ensemble (LightGBM)	Internal (split)	Accuracy 86%
Jahan et al. [32]	Alzheimer's disease	3	Federated, random forest	Internal (CV)	Accuracy 99%
Osman [33]	Diabetes	3	1D-CNN	Internal (split)	Accuracy 88%
Demuth et al. [34]	Precision medicine	7+	Digital twins	None (conceptual)	Conceptual
Peggy and Aremu [35]	Value-based care	5	Business intelligence	Internal	Length of stay −15%
Ray and Huma [36]	Oncology / chronic disease	4+	Ensemble (XGBoost, LSTM)	External	Accuracy 91%, AUC 0.94
Naoum et al. [37]	Oral cancer	2	CNN (EfficientNet)	Internal (split)	Accuracy 83%
Azhar et al. [38]	Epilepsy (EEG)	2	Deep learning (BiLSTM, GNN)	Internal (CV)	KLD 0.25
Mukhi et al. [39]	Spine oncology	4	CNN, XGBoost	Internal	Accuracy 0.88, AUC 0.91
Anjani et al. [40]	Neurology	3	Recurrent neural network	Internal (CV)	Accuracy 99%
Abdullah et al. [41]	Endocrinology (stress)	4	CNN, LSTM, Bayesian	Internal (CV)	Macro F1 ∼99% (synthetic)
Singh and Rahman [42]	Autism spectrum disorder	3	Federated, edge AI	Internal (synthetic)	Accuracy 92%
dos Santos [43]	Cardiac surgery / ICU	Mult.	Transformer, CNN	External	AUROC up to 0.89
Mahalle et al. [44]	Multi-disease	Mult.	ML, DL (case studies)	Internal (CV)	Accuracy (varies)
Logan [45]	Retinal disease (OCT)	2	Autoencoders	Internal (split)	Accuracy (varies)
Knowles et al. [46]	Health informatics	4	Ontology, LVLM	None (conceptual)	Conceptual
Gao et al. [47]	Heart failure (ICU)	2	Clinical BERT, tabular NN	External	AUROC 0.77 (external)
Canedo [48]	Computer vision	3	CNN, YOLO	Internal (CV)	Accuracy (varies)
Wang et al. [49]	ESRD / COVID-19	2	ML, LLM (RAG)	Internal (split)	Accuracy 0.80–0.90
Rahman [50]	Autism diagnosis	3	XGBoost–LSTM, federated	Internal	Accuracy 93%
Fiorini et al. [51]	Affective computing (EEG)	2	CNN, transformer	Internal (split)	F1 0.56
Vaddepalli [52]	Medical imaging (MLOps)	4	Data-centric MLOps	Internal	Repro. errors −62%
Battineni et al. [53]	Alzheimer's disease (MRI)	2	Gradient boosting	Internal (CV)	Accuracy 98%, AUROC 0.98
Kumar et al. [54]	Oncology	3	DenseNet, NLP	Internal (split)	Accuracy >90%
KECHABIA et al. [55]	Cross-domain annotation	4	Human-in-the-loop	Qualitative	Qualitative
Darwai et al. [56]	ICU outcomes	2	ML, NLP	Internal (split)	Accuracy 90%, AUC 0.93

Table 2
Summary characteristics of the 39 included studies.

Characteristic	n (%)
Publication year	
2021	1 (3)
2024	13 (33)
2025	14 (36)
2026	11 (28)
Number of modalities	
2	9 (23)
3	13 (33)
4	10 (26)
≥5	7 (18)
Modalities used (studies using each)	
Structured EHR / tabular	22 (56)
Imaging	19 (49)
Physiological signals / time-series	12 (31)
Clinical text	10 (26)
Wearable / sensor	10 (26)
Genomics / omics	5 (13)

methods, and attention visualization—with a minority using integrated gradients, counterfactual, or prototype-based explanations. Formal evaluation of explanation faithfulness or stability was rare; explanations were typically presented as illustrative outputs rather than validated against ground-truth attributions or tested for robustness, limiting their evidentiary value for clinician trust.

3.6. Validation, outcomes, and reporting completeness

In sharp contrast to the technique-reporting rates, indicators of deployment readiness were rare (Table 5). External validation on a separate site, dataset, or cohort was reported in only 4/39 studies (10%): two imaging- or EHR-fusion models validated on independent cohorts [20, 47], and two further multimodal systems validated out-of-sample [36, 43]; the remainder relied on internal validation via held-out splits or cross-validation. A patient, clinician, or health-system outcome was measured in only 3/39 studies (8%): an operational/prognostic gain and a clinician-evaluated copilot [19, 28], together with a translational-medicine framework reporting an outcome proxy [30]. Most studies reported technical metrics alone or described intended benefits (for example, "supports decision-making") without measuring them. Fewer than half of the studies cited any reporting or

Table 3
AI paradigm, fusion strategy, and explainability method (studies using each; n = 39). Categories are not mutually exclusive.

Category	Studies, n (%)
AI paradigm / model family	
Convolutional neural network	13 (33)
Ensemble / gradient boosting	11 (28)
Recurrent / LSTM	9 (23)
Transformer / attention	8 (21)
Large language / generative AI	5 (13)
Federated learning	5 (13)
Fusion strategy	
Early / feature-level	13 (33)
Attention / intermediate	7 (18)
Late / decision-level	4 (10)
Not specified / conceptual	15 (38)
Explainability method	
SHAP	22 (56)
Attention visualization	11 (28)
LIME	9 (23)
Gradient / saliency maps	9 (23)
Counterfactual	3 (8)
Integrated gradients	2 (5)

Table 4
Data-centric and reliability techniques implemented or evaluated (n = 39).

Technique	Studies, n (%)
Label-noise / data-quality handling	33 (85)
Distribution-shift handling	31 (79)
Calibration / uncertainty quantification	30 (77)
Missing-modality handling	30 (77)
Class-imbalance handling	29 (74)
Equity / subgroup analysis	24 (62)

governance standard, and explicit prediction-model reporting guidelines such as TRIPOD+AI were essentially absent; where standards were named they were typically data-protection or governance frameworks rather than model-reporting guidelines. Prospective or in-workflow deployment was not identified in any included study. A curated set of the most methodologically informative studies—those carrying the field's only external validations and measured outcomes—is summarized in Table 6.

Table 5
Deployment-readiness and reporting indicators (n = 39).

Indicator	Studies, n (%)
External / multi-site validation reported	4 (10)
Patient / clinician / system outcome measured	3 (8)
Calibration / uncertainty reported	30 (77)
Equity / subgroup analysis reported	24 (62)
Prediction-model reporting standard (e.g., TRIPOD+AI) cited	<5
Prospective or in-workflow deployment	0 (0)

4. Discussion

4.1. Principal findings

Across 39 primary multimodal CDS studies, almost all published since 2024, data-centric robustness and explainability techniques were reported in roughly three-quarters of studies, yet external validation (10%), measured clinical outcomes (8%), and standardized reporting were nearly absent, and no study reported routine deployment. The field is therefore methodologically enthusiastic but evidentially shallow: it shows that robustness and explainability techniques can be applied to multimodal CDS, not that the resulting systems are reliable or beneficial in practice. This is proof-of-concept, not proof-of-benefit.

4.2. Interpretation

The high technique-reporting rates require caution on two counts. First, because the search targets data-centric, robust, and explainable multimodal work, it preferentially retrieves studies claiming these properties; the rates describe what this slice of the field emphasizes, not the field as a whole. Second, and more fundamentally, a robustness technique reported on a single internal dataset is a claim, not evidence: with external validation at 10%, the out-of-distribution reliability that "robustness" implies is largely untested. The juxtaposition is the key result–strong internal methodology, weak external and outcome evidence–and it is robust to the coding caveats, since the deployment-readiness indicators are objective and uniformly extracted. This pattern mirrors recent systematic reviews in adjacent clinical-AI domains [14, 15] and a broad multimodal-medicine scoping review [3]. It is also consistent with documented optimism and limited transportability in clinical machine-learning prediction models [6, 9], where second-opinion and reproducibility concerns are well documented [57].

4.3. Comparison with existing literature

Prior reviews have examined multimodal medical AI, clinical decision support, and data-centric methods largely in isolation. The Schouten scoping review mapped technical challenges and clinical applications of multimodal medicine and noted that clinical translation lags technical development [3]; reviews of patient-generated-data AI and of wearable interventions found favorable metrics but absent external validation and unmeasured benefit [14, 15]. Our review extends these findings to the specific intersection of data-centric robustness, explainability, and multimodality, and contributes the implementation-versus-mention lens that reveals how often robustness is asserted rather than demonstrated. Against deployment-oriented standards [10, 11], including formal prediction-model appraisal tools [12], none of the included studies met best-practice expectations for external validation and outcome measurement.

4.4. Implications for research and practice

For researchers, continued technique development without external validation and outcome measurement will yield diminishing returns. Four priorities follow directly from the

Table 6
Exemplar studies across modality groups and methodological strengths (curated subset of the 39 included studies). This subset illustrates the breadth and the strongest evidence in the corpus rather than ranking study quality; performance values are point estimates from each study's headline analysis.

Study	Clinical area	AI technique	Mod. (n)	Reason for selection
Kovács [20]	Prostate cancer (MRI)	3D CNN (nnU-Net)	3	Externally validated imaging model (AUROC 0.82)
Gao et al. [47]	Heart failure (ICU)	Clinical BERT + tabular NN	2	External + temporal validation; cites TRIPOD
Ray and Huma [36]	Oncology / chronic disease	Stacked ensemble (XGBoost, LSTM)	4+	External validation with reported outcome and equity analysis
dos Santos [43]	Cardiac surgery / ICU	Transformer (ICU-BERT), CNN	Mult.	Missingness-robust fusion; evaluated in a clinical study
Ghantasala [28]	Pediatric type-1 diabetes	Stacked LSTM + LLM copilot	3	Clinician-evaluated decision support (MAE 10.9 mg/dL)
Perwej [19]	Multi-domain prognostics	Deep ensemble	Mult.	Reports an operational/prognostic outcome

gaps identified: (1) external, multi-site validation under realistic missingness and acquisition heterogeneity, rather than further internal benchmarking; (2) measurement of at least one patient, clinician, or workflow outcome benchmarked against current practice; (3) routine adherence to prediction-model reporting and appraisal standards and formal evaluation of explanation faithfulness rather than illustrative use [10, 12]; and (4) treatment of robustness claims—to missing modalities, label noise, and distribution shift—as hypotheses to be tested out-of-distribution, not as properties established by construction. For healthcare organizations, current evidence supports structured pilot evaluation with prospective monitoring rather than routine deployment.

4.5. Strengths and limitations

Strengths include a focused, novel question at an unaddressed intersection, transparent eligibility criteria, uniform extraction, and an explicit coding rule separating implemented from mentioned techniques. Several limitations should be acknowledged. The search-term selection inflates technique-reporting rates and the corpus is dominated by 2024–2026 work, so frequencies reflect a recent, self-selected slice rather than the whole field; the robust conclusion is the deployment gap, not the absolute technique rates. The eligibility restriction to multimodal primary studies excludes the large secondary literature and single-modality work by design. Heterogeneity in designs, datasets, and outcomes precluded meta-analysis, and inconsistent reporting of fusion strategy and sample size limited some comparisons. Finally, formal item-level risk-of-bias appraisal (PROBAST+AI) and dual-reviewer adjudication would further strengthen the synthesis and are priorities for the full appraisal; broadening the search across additional databases and citation snowballing would enlarge the evidence base. These constraints temper the precision of individual proportions but not the central, repeatedly corroborated finding.

4.6. Conclusion

Data-centric, robust, and explainable multimodal deep learning for clinical decision support is a fast-growing but immature evidence base. Studies widely report missing-modality handling, label-noise and distribution-shift mitigation, calibration, and explainability, yet almost none demonstrate external validity or measured clinical benefit, and few report against AI standards. Closing the gap between technical promise and patient benefit requires external multi-site validation, outcome measurement, and standardized, appraisal-aware reporting. Until then, these systems should be regarded as research artifacts rather than deployable decision support.

CRediT authorship contribution statement

Md. Mazharul Islam: Conceptualization, Methodology, Data curation, Formal analysis, Writing – original draft, Supervision. **Abrar Mohammed Tanzim Alam:** Methodology, Data curation, Validation, Writing – review & editing. **Md. Sadikur Rahman Rony:** Data curation, Validation, Writing – review & editing. **Susmay Das:** Data curation, Validation, Writing – review & editing.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Ethics

Ethical approval was not required, as this review analyzed published, aggregated data and did not involve human participants.

Data availability

The extraction dataset and screening decisions supporting this review are provided as Supplementary Material.

Declaration of generative AI and AI-assisted technologies in the manuscript preparation process

During the preparation of this work the authors used ChatGPT in order to edit the texts (rephrasing, etc). After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

Declaration of competing interest

The authors declare no competing interests.

References

[1] J. N. Acosta, G. J. Falcone, P. Rajpurkar, E. J. Topol, Multimodal biomedical ai, Nature medicine 28 (2022) 1773–1784.

[2] L. R. Soenksen, Y. Ma, C. Zeng, L. Boussioux, K. Villalobos Carballo, L. Na, H. M. Wiberg, M. L. Li, I. Fuentes, D. Bertsimas, Integrated multimodal artificial intelligence framework for healthcare applications, NPJ digital medicine 5 (2022) 149.

[3] D. Schouten, G. Nicoletti, B. Dille, C. Chia, P. Vendittelli, M. Schuurmans, G. Litjens, N. Khalili, Navigating the landscape of multimodal ai in medicine: A scoping review on technical challenges and clinical applications, Medical Image Analysis 105 (2025) 103621.

[4] J. Jakubik, M. Vössing, N. Kühl, J. Walk, G. Satzger, Data-centric artificial intelligence, Business & Information Systems Engineering 66 (2024) 507–515.

[5] Y. Wei, Y. Deng, C. Sun, M. Lin, H. Jiang, Y. Peng, Deep learning with noisy labels in medical prediction problems: a scoping review, Journal of the American Medical Informatics Association 31 (2024) 1596–1607.

[6] L. M. Koch, C. F. Baumgartner, P. Berens, Distribution shift detection for the postmarket surveillance of medical ai algorithms: a retrospective simulation study, NPJ Digital Medicine 7 (2024) 120.

[7] Q. Xu, W. Xie, B. Liao, C. Hu, L. Qin, Z. Yang, H. Xiong, Y. Lyu, Y. Zhou, A. Luo, Interpretability of clinical decision support systems based on artificial intelligence from technological and medical perspective: a systematic review, Journal of healthcare engineering 2023 (2023) 9919269.

[8] A. Pahud de Mortanges, H. Luo, S. Z. Shu, A. Kamath, Y. Suter, M. Shelan, A. Pöllinger, M. Reyes, Orchestrating explainable artificial intelligence for multimodal and longitudinal data in medical imaging, npj Digital Medicine 7 (2024) 195.

[9] Y. Yang, H. Zhang, J. W. Gichoya, D. Katabi, M. Ghassemi, The limits of fair medical imaging ai in real-world generalization, Nature medicine 30 (2024) 2838–2848.

[10] G. S. Collins, K. G. M. Moons, P. Dhiman, R. D. Riley, A. L. Beam, B. Van Calster, M. Ghassemi, X. Liu, J. B. Reitsma, M. van Smeden, et al., Tripod+ai statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods, BMJ 385 (2024) e078378.

[11] B. Vasey, M. Nagendran, B. Campbell, et al., Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: Decide-ai, Nature Medicine 28 (2022) 924–933.

[12] K. G. M. Moons, J. A. A. G. Damen, T. Kaul, et al., Probast+ai: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods, BMJ 388 (2025) e082505.

[13] U.S. Food and Drug Administration, Health Canada, Medicines & Healthcare products Regulatory Agency, Good machine learning practice for medical device development: Guiding principles, Official regulatory guidance, 2021.

[14] T. Warren, W. van der Weegen, R. B. Kool, M. Hoogendoorn, T. Timmers, Artificial intelligence applications using patient-generated health data for pre-care processes in elective healthcare: a systematic review, International Journal of Medical Informatics 218 (2026) 106525.

[15] S. Taborri, E. Renzi, A. Massimi, A. Improta, E. Di Simone, N. Giannetta, N. Panattoni, S. Dionisi, M. Di Muzio, L. Amato, Wearable devices for improving medication adherence in older adults with chronic conditions: A systematic review, International Journal of Medical Informatics 218 (2026) 106515.

[16] M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, E. A. Akl, S. E. Brennan, et al., The prisma 2020 statement: an updated guideline for reporting systematic reviews, bmj 372 (2021).

[17] M. M. Islam, A. M. T. Alam, M. S. R. Rony, S. Das, Data-centric, robust, and explainable multimodal deep learning for clinical decision support: A systematic review, PROSPERO CRD420261427815, 2026. Available from: https://www.crd.york.ac.uk/PROSPERO/view/CRD420261427815.

[18] D. Van Dievoort, R. De Croon, V. Vande Abeele, K. Verbert, Data-centric explanation methods for healthcare professionals, in: International Workshop on Process Mining Applications for Healthcare, Springer, 2025, pp. 66–72.

[19] Y. Perwej, Machine learning driven predictive modelling for intelligent data-centric applications, Machine Learning 36 (2026) 678–683.

[20] B. Kovács, Data-Centric Artificial Intelligence for Enhanced Prostate Cancer Diagnosis on Magnetic Resonance Images, Ph.D. thesis, German Cancer Research Center (DKFZ) / Heidelberg University, 2026.

[21] M. Sowmya, A. Kaliappan, S. Govindaraju, S. Menaka, T. Sangeetha, R. Ranjani, A. Kumar, Interpretable federated learning for privacy-centric genomic diagnostics: A multi-institutional framework, Genetics and Molecular Research (2026).

[22] R. Weiskirchen, Bridging the bench-to-bedside gap with multimodal artificial intelligence in digestive diseases, Livers 6 (2026) 1.

[23] S. S. R. Tummuri, Generative ai for data-centric healthcare with integrated anomaly detection and monitoring, in: 2026 International Conference on Communication, Computing and Emerging Technologies (IC3ET), IEEE, 2026, pp. 520–526.

[24] P. Komninos, T. Kontogiannis, N. Eleftheroglou, D. Zarouchas, A robust generalized deep monotonic feature extraction model for label-free prediction of degenerative phenomena, Data-Centric Engineering 7 (2026) e4.

[25] K.S.Ranjith, A. G. Yasasree, K. Haritha, S. M. Nizamuddin, M. D. Shivaji, N. K. Akula, Ai-driven hybrid decision support framework for heart disease prediction and personalized treatment recommendation, in: 2026 International Conference on Smart Futuristic Technology, 2026, pp. 1–7. doi:10.1109/ICSFT66733.2026.11508018.

[26] M. Mohanty, P. S. Rath, A. G. Mohapatra, A. Mohanty, S. K. Senapati, Ai-enhanced data processing for modeling applications, in: Advances in Computers, volume 142, Elsevier, 2026, pp. 515–540.

[27] N. S. Makesh, M. Naveenkumar, J. Kavitha, K. D. Prasad, P. Badrinath, Transformer-tuned deep learning architecture for real-time medical diagnosis using federated learning, in: 2026 4th International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT), IEEE, 2026, pp. 844–850.

[28] K. K. Ghantasala, Pediatric t1d copilot: Lstm-driven glucose forecasting and conversational decision support from real-world cgm, insulin, and meal streams, in: 2026 5th International Conference on Communication, Computing and Electronics Systems (ICCCES), IEEE, 2026, pp. 1394–1399.

[29] F. Hossain, A multimodal big data and explainable ai framework for personalized cancer care: Extending methods for clinical translation, Journal of Medical and Health Studies 6 (2025) 104–117.

[30] L. Abualigah, S. A. Alomari, M. H. Almomani, R. A. Zitar, K. Saleem, H. Migdady, V. Snasel, A. Smerat, A. E. Ezugwu, Correction: Artificial intelligence-driven translational medicine: a machine learning framework for predicting disease outcomes and optimizing patient-centric care, Journal of Translational Medicine 23 (2025) 908.

[31] T. R. Noviandy, M. Patwekar, F. Patwekar, R. Idroes, An interpretable machine learning framework for predicting advanced tumor stages, Infolitika Journal of Data Science 3 (2025) 61–69.

[32] S. Jahan, M. R. S. Adib, S. M. Huda, M. S. Rahman, M. S. Kaiser, A. S. Hosen, D. Ghimire, M. J. Park, Federated explainable ai-based alzheimer's disease prediction with multimodal data, IEEE Access (2025).

[33] R. A. Osman, Explainable ai-driven 1d-cnn with efficient wireless communication system integration for multimodal diabetes prediction, AI 6 (2025) 243.

[34] S. Demuth, J. De Sèze, G. Edan, T. Ziemssen, F. Simon, P.-A. Gourraud, Digital representation of patients as medical digital twins: Data-centric viewpoint, JMIR Medical Informatics 13 (2025) e53542.

[35] O. O. Peggy, B. K. Aremu, Integrating multimodal patient data and business intelligence for strategic healthcare service optimization and value-based delivery, Int J Adv Res Publ Rev 2 (2025) 422–442.

[36] R. K. Ray, Z. Huma, Intelligent healthcare at scale: Data-driven support through cloud infrastructure and ai for understanding human actions, Multidisciplinary Innovations & Research Analysis 6 (2025) 8–25.

[37] J. Naoum, R. Salama, A. Hamdi, Data-augmented multimodal feature fusion for multiclass visual recognition of oral cancer lesions, arXiv preprint arXiv:2511.21582 (2025).

[38] A. Azhar, A. Mathur, S. Jain, J. Emilian, S. Mandal, N. Shah, Y. J. Zhang, Modeling clinical decision variability in explainable multimodal seizure detection, in: Proceedings of the 4th Machine Learning for Health Symposium, 2025.

[39] S. Mukhi, I. Ali, A. Satyanarayana, S. Fatima, M. A. Hussain, R. D. Kumar, Spine oncology detection net: a data-centric deep neural model for diagnosis, in: 2025 2nd International Conference on Software, Systems and Information Technology (SSITCON), IEEE, 2025, pp. 1–5.

[40] P. Anjani, S. Sujatha, E. Hemavathi, B. Mathivanan, M. Stephen, et al., Iot-enabled brain health monitoring for clinical diagnosis using recurrent neural networks, in: 2025 2nd International Conference on Artificial Intelligence for Innovations in Healthcare Industries (ICAIIHI), IEEE, 2025, pp. 1–6.

[41] Abdullah, Z. Fatima, C. G. Sánchez Mejorada, M. A. Ather, J. L. Oropeza Rodríguez, G. Sidorov, Fair and explainable multitask deep learning on synthetic endocrine trajectories for real-time prediction of stress, performance, and neuroendocrine states, Computers 14 (2025) 515.

[42] A. Singh, S. Rahman, Towards responsible ai in autism care: A multimodal federated-edge framework for real-time behavioral support (2025).

[43] R. B. B. dos Santos, Improving the Robustness of Multimodal AI with Asynchronous and Missing Inputs, Ph.D. thesis, Universidade NOVA de Lisboa (Portugal), 2024.

[44] P. N. Mahalle, N. N. Wasatkar, G. R. Shinde, Data-centric artificial intelligence for multidisciplinary applications, CRC Press, 2024.

[45] Y.-Y. Logan, Data-Centric Approaches for Exploiting Metainformation and Mitigating Model Regression to Aid Neural Networks, Ph.D. thesis, Georgia Institute of Technology, 2024.

[46] P. Knowles, B. Gajderowicz, K. Dugas, Data-centric design: Introducing an informatics domain model and core data ontology for computational systems, arXiv preprint arXiv:2409.19653 (2024).

[47] Z. Gao, X. Liu, Y. Kang, P. Hu, X. Zhang, W. Yan, M. Yan, P. Yu, Q. Zhang, W. Xiao, et al., Improving the prognostic evaluation precision of hospital outcomes for heart failure using admission notes and clinical tabular data: multimodal deep learning model, Journal of medical Internet research 26 (2024) e54363.

[48] D. D. Canedo, Data-Centric Artificial Intelligence in the Context of Computer Vision, Ph.D. thesis, Universidade de Aveiro (Portugal), 2024.

[49] Z. Wang, Y. Zhu, J. Gao, X. Zheng, Y. Zeng, Y. He, B. Jiang, W. Tang, E. M. Harrison, C. Pan, et al., Retcare: Towards interpretable clinical decision making through llm-driven medical knowledge retrieval, in: Artificial Intelligence and Data Science for Healthcare: Bridging Data-Centric AI and People-Centric Healthcare, 2024.

[50] S. Rahman, Trustworthy clinical decision pipeline for autism diagnosis, Frontiers in Computer Science and Artificial Intelligence 3 (2024) 46–51.

[51] L. Fiorini, F. Bossi, F. Di Gruttola, Eeg-based emotional valence and emotion regulation classification: a data-centric and explainable approach, Scientific reports 14 (2024) 24046.

[52] R. K. Vaddepalli, Data versioning for iterative refinement: Adapting ml experiment tracking tools for data-centric ai pipelines, International Journal of AI, BigData, Computational and Management Studies 5 (2024) 129–137.

[53] G. Battineni, M. A. Hossain, N. Chintalapudi, E. Traini, V. R. Dhulipalla, M. Ramasamy, F. Amenta, Improved alzheimer's disease detection by mri using multimodal machine learning algorithms, Diagnostics 11 (2021) 2103.

[54] J. Kumar, A. Chandran, S. Rawal, S. Rehan, A self-adaptive densenet (sad) for enhancing precision oncology in clinical decision support, Journal of Informetrics (ISSN 1875-5879) 20 (2024).

[55] Y.-R. KECHABIA, R. Cyril, M. PETIT, F. LAJONCHERE, P. John, et al., Human-centric annotation of multi-modal data: A framework perspective (2026).

[56] S. Darwai, I. Khan, D. P. Tiwari, K. Patidar, An intelligent framework for machine learning-driven clinical outcome prediction using integrated structured and unstructured patient data (2026).

[57] B. Kompa, J. Snoek, A. L. Beam, Second opinion needed: communicating uncertainty in medical machine learning, npj Digital Medicine 4 (2021) 4.