Volume 10
Issue 2
Summer 2026

Westcliff International Journal of Applied Research

WIJAR

10TH ANNIVERSARY EDITION

JUNE 2026
ISSN 2572-7176
wjournal@westcliff.edu

Westcliff International
Journal of Applied Research

Volume 10 - Issue 2 - Summer 2026

OPEN-ACCESS • MULTIDISCIPLINARY

IN THIS ISSUE

ARTICLES

ABOUT WIJAR

Westcliff International Journal of Applied Research (WIJAR) is a multidisciplinary, double-blind, peer-reviewed, open-access journal published by the *LITE Center from the Office of Faculty Affairs* in partnership with the *Doctoral Affairs and Academic Resources Department* at Westcliff University. The journal was founded in 2017 and provides an opportunity for academics, industry professionals, and students to publish innovative research that offers insight into practical implementation. In order to widely disseminate new knowledge and scholarship, WIJAR advocates for all submissions to be written in a style that is accessible/available to a broad audience or readership, including those readers who may not be familiar with either research or the topic studied. The journal aligns with Westcliff University's mission to educate, inspire, and empower individuals through its dedication to supporting authors in the review and revision process to produce the highest quality content possible.

Distinguishing this journal from others similar is the strong support offered to contributors, especially first-time authors who may need additional writing or structural assistance. All contributors have access to the Westcliff University Online Writing Center, where dedicated research/writing specialists can offer support and suggestions.

LETTER FROM THE EDITOR-IN-CHIEF

June 2026

I am pleased to share this issue with our readers on behalf of the Editorial Board of the *Westcliff International Journal of Applied Research* (WIJAR). As we celebrate the *10th anniversary* of our journal, this issue holds a special meaning, marking a decade of commitment to academic excellence and applied research. We wish to honor the vision of our founders and the dedication of everyone who has shaped WIJAR into the respected platform it is today.

This issue brings together four articles selected from submissions received in late 2025 and finalized in early 2026. Its publication is the result of the tireless work and dedication behind the review and editorial process. I would like to take this opportunity to thank those who supported its development, ensuring that each contribution reflects the rigor our community expects.

The journal continues to grow thanks to the devotion of its community. Our editorial board, reviewers, and authors all play an essential role in maintaining the quality and relevance of the work we publish. Their time, care, and expertise are deeply valued.

This work reflects our strong commitment to quality. Each article has undergone a careful peer-review and revision process to ensure it meets high academic standards, including clarity, rigor, and strict adherence to APA guidelines.

I invite our readers, faculty, staff, students, and researchers to explore this issue and engage with the ideas it presents. We hope it not only informs your work but also inspires you to consider contributing your own research to WIJAR.

Sincerely,

Mary Allegra
Editor-in-Chief

ACKNOWLEDGEMENTS

The publication of the *Westcliff International Journal of Applied Research* (WIJAR) relies on the contributions of dedicated individuals. We extend our appreciation on behalf of the journal to:

- Dr. Anthony Lee for his unwavering support and strong belief in the journal's significant value for Westcliff University and the wider academic community.
- Dr. Laura Sliwinski and Dr. Diana Siganoff for their institutional support of the journal.
- Members of the editorial, internal, and external review boards of WIJAR for participating in the review process to assess and select high-quality research articles.
- Every author who has dedicated their time and energy to presenting their thoughts and perspectives in this publication.
- The Marketing Department of Westcliff University for their overall participation and contributions to the journal's marketing, the development of the journal's website, and their significant role in the publication's success.
- The Westcliff University Writing Center for their assistance and collaboration with the authors throughout the revision and review process.

Thank you all!

Smarter Search for Law Enforcement: Evaluating Context-Aware and Traditional Approaches

Archan Dutta
Westcliff University

Thaís D. Poi
San Joaquin Delta College

Abstract

Despite widespread and rapid advancements in natural language processing, most law enforcement case and report management systems still rely on keyword-based search engines. Information retrieval within law enforcement agencies depends on obtaining vast amounts of unstructured text data, including case reports, incident reports, and written statements. Investigators must sift through and analyze this data to find and discover meaningful leads. Keyword-based search engines face challenges with synonyms, equivalent terms, abbreviations, slang, paraphrases, and narrative fluctuation often found in police reports and case files. As a result, investigators may overlook potentially relevant cases unless they manually craft multiple keyword searches. This quantitative study evaluated, using a synthetic dataset generated for this study, whether context-aware models (also known as semantic models or embedding models) reach higher retrieval accuracy than traditional keyword-based search models on law enforcement reports. Results of the study showed that semantic models have significantly better performance than traditional keyword-based search, an improvement of 56 percentage points. Besides investigators, inspectors, and analysts, smarter and enhanced search may benefit other key parties like patrol officers, crime analysts, and prosecutors by allowing faster discovery of related incidents, improved case connection, and richer text analysis. These results highlight the potential to improve operational efficiency, reduce administrative and management burden, and strengthen public safety outcomes.

Keywords: semantic search, information retrieval, law enforcement, natural language processing, embedding models, crime reports, retrieval accuracy

Introduction

Law enforcement departments produce and manage massive volumes of unstructured text-based data every day, from incident reports and witness statements to investigation notes. In the majority of record-management and case processing systems, traditional keyword search continues to be the dominant search technology and may not consistently retrieve all relevant documents. Some examples of record-management systems are municipal file management systems, state repositories, and federal databases such as the National Incident-Based Reporting System and the National Crime Information Center. Traditional keyword search has limitations in terms of accurate retrieval. These limitations arise from the characteristics of

natural language, such as English, which has many narrative variations, paraphrases, interpretations, acronyms, synonyms, polysemy, slang, and other informal expressions. Reports describing the same event may vary widely in vocabulary and phrasing. For example, “car stolen from home garage,” “vehicle taken last night,” or “motor vehicle theft.”

The traditional keyword-based search models like Okapi BM25 and Term Frequency–Inverse Document Frequency (TF-IDF) are not “context-sensitive” and are based on literal and explicit word overlap between a query and a document (Robertson & Zaragoza, 2009). Keyword search models treat these as unrelated, missing vital connections between similar cases. These limitations impede investigators, supervisors, attorneys, and intelligence units from generating accurate insights from historical text data. Current advances in natural-language processing (NLP) provide an innovative alternative. Semantic search methods represent text as numerical vectors (also known as embeddings) that capture the meaning of the sentences. By measuring the similarity between these vectors, systems can retrieve conceptually related reports even when they share a small number of or no common terms. Among these models, the Sentence-BERT (SBERT) class of models has shown strong retrieval performance in the open domain, indicating potential for specialized fields such as law enforcement.

We benchmarked traditional keyword-based retrieval models (BM25, TF-IDF) against two SBERT versions (MPNet, MiniLM) on a document dataset of law enforcement reports, primarily focusing on retrieval accuracy. Through observation-based analysis of semantic and traditional search side by side, this research revealed the possibility of improving investigative workflows, intelligence analysis, and content access across all tiers of policing using context-sensitive Artificial Intelligence (AI) models. By improving search retrieval, smarter retrieval tools have the potential to shorten investigative timelines and support more effective and transparent public-safety operations.

Background

This section establishes the crucial role of information retrieval (IR) in the high-risk area of law enforcement, where efficiency and accuracy immediately impact investigations and community safety. We examined the inherent difficulties in traditional search methods, specifically dealing with the lexical gap problem. Subsequently, we detailed the transformative development of semantic search, resolving these limitations by using vector embeddings to capture the contextual meaning and interpretation of texts and queries. This provides the necessary background needed to understand traditional, keyword-based techniques and state-of-the-art, context-aware semantic methods.

Information Retrieval in Law Enforcement

The complexities of criminal analysis necessitate leveraging the intricacies found exclusively in descriptive, narrative-based data (Lovell et al., 2022; Parker, 2025). These free-text narratives consistently contain critical, subtle contextual details such as inferred officer attitudes concerning victim credibility or nuanced descriptions of perpetrator acts, which influence case results (Lovell et al., 2022; Parker, 2025). As a result, law enforcement and judicial systems increasingly depend on efficient and accurate access to immense amounts of unstructured textual data (Xing & Chen, 2024). If the retrieval/search mechanism is inaccurate, then critical qualitative information remains unseen and inaccessible. Therefore, improving retrieval performance is more than just a gradual efficiency gain but a main factor for helping analysis (Xing & Chen, 2024).

Lexical Gap Problem

Because the traditional retrieval methods are based on the exact word overlap between a query and a document, these systems are fundamentally limited by the vocabulary mismatch problem, also known as the lexical gap (Al-Haddad et al., 2025). This gap occurs when the query and the relevant document describe the same concept using different terminology, such

as synonyms, paraphrasing, or misspellings (Al-Haddad et al., 2025). Table 1contains examples that show the limitations of keyword-based models, highlighting that even though the word overlap between the police report and the query is low, they still have similar semantic meaning. The low word overlap means that keyword-based models will struggle to retrieve these documents, despite having similar semantic meaning.

Table 1

Examples of reports and queries showing limitations of a keyword-based search

Police Report	Query	Word Overlap	Semantic Relationship
The suspect broke into a parked vehicle and stole a laptop from the back seat.	Who took the computer from my car?	"stole" ≠ "took"; "laptop" ≠ "computer"; "vehicle" ≠ "car"	Both are about theft
A woman reported receiving threatening calls from an unknown number.	Has anyone complained about harassment over the phone?	"threatening calls" ≠ "harassment"	Both about phone harassment

Emergence of Semantic Search via Embeddings

In a semantic search model, text is mapped into a continuous, higher-dimensional vector space, known as embeddings (Opitz et al., 2025; Schutz & Hupfeld, 2021). In this space, similarity is measured based on semantic meaning rather than exact word form, typically using cosine similarity (Reimers & Gurevych, 2019). The cosine similarity measures semantic similarity; how similar the meanings of two sentences are. A higher cosine similarity score (closer to 1) means that the sentences have a more similar meaning, while a lower score (closer to 0) means they are semantically different than each other. This approach is widely used in natural language processing applications to identify semantic relationships between words, phrases, or entire texts (Reimers & Gurevych, 2019; Zhelezniak et al., 2019; Steck et al., 2024). Some examples of pairs of text (a report and a query) and their corresponding cosine similarity are shown in Table 2. The first example from Table 2 demonstrates that embedding models can identify the semantic similarity between the report and the query (as shown by the high cosine value) even when the word overlap is low.

Table 2

Examples of Reports and Queries, and their Cosine Similarity

Example	Police Report	Query	Cosine Similarity
1	The suspect broke into a parked vehicle and stole a laptop from the back seat.	Who stole the laptop from the car?	High 0.83
2	A woman reported receiving multiple harassing phone calls from an unknown number.	Any cases of identity theft reported this week?	Medium 0.52
3	A pedestrian was struck by a speeding motorcycle near Main Street.	How many missing persons cases are open right now?	Low 0.28

Understanding Keyword-based Models and Embedding Models

The two most popular keyword-based models are TF-IDF and BM25. This model captures TF (how frequently words occur in a document) and IDF (how unique they are across the corpus). TF-IDF is purely statistical and does not capture semantics (meaning) or word order. BM25, developed in 1994, is an improvement over TF-IDF because BM25 has an additional probabilistic model of relevance. BM25 is widely used in search engines and information retrieval systems to rank search results. MiniLM and MPNet are built on BERT to improve semantic and contextual understanding.

Methods and Materials

Dataset Synthesis

Access to comprehensive, real-world law enforcement data presents significant challenges due to data sensitivity, data privacy regulations, and restricted sharing policies (Güss et al., 2020). Redaction is mandatory for personally identifiable information (PII), such as full names, social security numbers, and specific locations, to comply with regulations like the GDPR and local laws (Böhlin, 2024). Due to those challenges, we synthesized two datasets specifically for this study (the data is not derived from real-world law enforcement records). However, the LLM was prompted with actual law enforcement records to ground the synthetic data in real-world law enforcement records. The first dataset consists of 50 crime reports. The crime reports mirror the structure and linguistic content of the National Incident-Based Reporting System (NIBRS). The synthetic corpus was generated using a large language model (LLM), specifically GPT-4o. The second dataset consists of 50 queries. For each query, a ground truth was established. Ground truth represents the "true" or correct answer, obtained through direct observation or measurement, which is used to evaluate the output of a model. Considering the labor intensity of human annotation, the study leveraged a large language model (LLM) for obtaining the ground truth. Table 3 presents sample queries from the synthesized dataset, and Table 4 presents sample crime reports.

Table 3

Samples from the Synthesized Query Datasets

ID	Queries	Expected Report ID
0	Report concerning the removal of electronic components following a forced entry into a business.	0
1	Driver impairment suspected after a crash where a subject failed to obey traffic lights.	1
2	Domestic violence charges filed after a physical altercation over shared finances.	2
3	Damage to an historic municipal monument requiring expensive specialty restoration.	3
4	Seniors tricked by someone posing as a government official demanding money via gift vouchers.	4

Table 4

Samples from the Synthesized Crime Report Dataset

ID	Report
0	On Tuesday, at 02:30 AM, a commercial entryway breach was reported at the 'Digital Depot' electronics store. Two individuals forcibly gained access via a back portal. They quickly targeted the secure storage unit, absconding with high-value computing hardware before leaving the location in a dark-colored utility vehicle. The estimated value of the removed property is significant.
1	A severe multi-car event occurred at the intersection of Highway 101 and Oak Avenue at 5:45 PM on Friday. One operator, showing clear signs of impaired faculties, failed to yield to a signal and struck a large passenger vehicle. The impaired subject resisted sobriety measures and was detained on site.
2	Officers responded to a disturbance involving a heated cohabitation conflict at apartment complex unit 4B. The male subject was preparing to exit the dwelling after physically assaulting the female victim over a debate concerning domestic finances. The suspect was taken into custody and faces charges related to domestic battery.
3	Extensive architectural defacement was discovered early Sunday morning at Central Park’s main pavilion. Large segments of the stone facade were covered in unauthorized painted messages with derogatory content. Restoration costs are projected to be high due to the required specialist cleaning.
4	The cybercrimes unit identified a complex impersonation scheme targeting elderly citizens. The perpetrators pretended to be revenue agents via voice calls, demanding immediate remittance through prepaid cards or crypto assets to avoid legal consequences. The public is urged to verify callers requesting funds.

Evaluation

The performance of both approaches is measured using a metric called Hit@3, which means retrieval accuracy in the top 3 reports. This metric measures the percentage of queries for which the relevant crime report is present in the top 3 results. The experiment step-by-step is recorded below.

- Generated a synthetic dataset of crime reports and queries.
- Obtained the ground truth for every query.
- For each model type, converted the reports and queries into the specific model type format suitable for retrieval.
- For every combination of query, report, and model.
  - Ran the retrieval for each query using cosine similarity.
  - Picked the top three reports based on the highest cosine similarity score.
  - If the ground truth existed, in the top three reports, marked it as a successful hit; otherwise, marked it as missed.
- Aggregated the hits and misses and calculated the retrieval accuracy in the top 3 as *#Hits / (#Hits + #Misses)*.

Results

Table 5 shows the retrieval results for a test query. The embedding-based models (MPNet-Base and MiniLM-L6) ranked the correct document (ID = 0) in the top position. In contrast, the keyword-based models (TF-IDF and BM25) did not retrieve the correct document in their top three results.

Table 5

Sample Retrieval Results for Queries

Queries	Expected Report ID	Model	Cosine Similarity of Expected Report ID	Top 3 Reports
Report concerning the removal of electronic components following a forced entry into a business.	0	MiniLM-L6	0.5613	[0, 8, 26]
Report concerning the removal of electronic components following a forced entry into a business.	0	MPNet-Base	0.6229	[0, 17, 20]
Report concerning the removal of electronic components following a forced entry into a business.	0	TF-IDF	0.0000	[26, 9, 42]
Report concerning the removal of electronic components following a forced entry into a business.	0	BM25	0.0000	[26, 2, 34]

Table 6 summarizes the overall retrieval performance across all queries using the Hit@3 metric. The MPNet-Base model achieved a Hit@3 score of 100% (on the synthesized dataset), while the MiniLM-L6 model achieved 98%. In comparison, the keyword-based models showed lower performance, with BM25 achieving 44% and TF-IDF achieving 28%.

Table 6

Retrieval Accuracy in Top 3 (Hit@3) by Model Type

Model	Type	Retrieval Accuracy in Top 3 (Hit@3)
MPNet-Base	Embedding	100%
MiniLM-L6	Embedding	98%
BM25	Keyword Based	44%
TF-IDF	Keyword Based	28%

Discussion/Implications

Table 5 shows the retrieval results for a test query. The embedding-based models (MPNet-Base and MiniLM-L6) successfully ranked the correct document (ID=0) in the top position. Their success is quantitatively supported by the high cosine similarity scores (0.6229 and 0.5613, respectively), measuring the semantic overlap between the report and the query. These scores indicate that the models correctly identified the meaning of "removal of electronic components" as being highly similar to "computing hardware," even when the exact words did not match. In contrast, the keyword-based models (TF-IDF and BM25) had a much lower retrieval accuracy, sometimes returning irrelevant documents in their top three ranks. Because keyword-based models only look for word overlap and disregard the underlying meaning, they effectively treated the query and the report as completely unrelated documents. Table 6 shows the Hit@3 for the 4 models, where the embedding-based models demonstrated notably better performance. The MPNet-Base model achieved a Hit@3 score of 100% (on the synthesized dataset), meaning every query successfully retrieved its corresponding crime report within the top three positions. The MiniLM-L6 model also performed exceptionally well, scoring 98%. In contrast, the traditional keyword-based models exhibited poor performance, BM25 achieved a Hit@3 of 44%, while TF-IDF was the lowest-performing model at 28%.

The results suggest the value of leveraging semantic search capabilities for information retrieval within complex, narrative-driven domains such as criminal reports. There is a 52-percentage point difference between MPNet-Base (100%) and BM25 (44%), a vast performance disparity between the embedding model with the highest Hit@3 and the keyword-based model with the highest Hit@3. TF-IDF was less accurate with a Hit@3 of 28%. Because TF-IDF and BM25 rely purely on the frequency and exact matching of terms between the query and the report, these models were unable to

recognize the semantic equivalence of the text, resulting in their low retrieval scores. Table 6 highlights the low retrieval accuracy of traditional keyword-based models. The better performance of embedding models, MPNet-Base (100% Hit@3) and MiniLM-L6 (98% Hit@3), can be attributed to their underlying transformer architectures. These models encode the full contextual meaning of the query and the report into a high-dimensional vector space, where similarity was calculated based on semantic proximity measures such as cosine similarity, allowing the models to correctly identify the relevant reports, even when the query used an entirely different terminology.

Academic and Practical Implications

From an academic perspective, this study contributes to several important gaps in the literature on IR, NLP, and policing. First, much of the IR literature has focused on keyword-based and embedding models in general domains (Boubekeur & Azzoug, 2013). Our work extends that line of inquiry from the general domain into the specific domain of law enforcement. Second, our work adds to the literature on the usefulness of machine learning/NLP in law enforcement. NLP and ML tools in policing are growing but still face challenges in adoption, domain-specific evaluation, and bias/ethics (Sarzaeim et al., 2023).

The findings have direct and practical implications for law enforcement agencies, including investigators, patrol officers, crime analysts, and prosecutors in leveraging smarter search capabilities to improve operations, efficiency, and public-safety outcomes. The superior performance of semantic models in our study implies that law enforcement agencies may deploy these systems to retrieve relevant reports even when lexical overlap is low. The use of semantic models may enable faster discovery of related crimes, better case linkage, and earlier identification of patterns. Patrol officers may retrieve relevant past incidents or intelligence while on shift. Similarly, crime analysts may be able to surface latent patterns that keyword systems may miss.

Deployment Considerations

There are several implementation considerations that law enforcement agencies need to consider when adopting these smarter search capabilities. Evaluating their existing case/report management systems to assess how often keyword searches fail because of vocabulary variation. Additionally, the agency may benefit from training employees on embedding models and how they differ from keyword-based models. Finally, a direct comparison of computational cost and latency is needed before deployment. The[16] [17] higher retrieval accuracy of MPNet may come at the cost of high latency, making MiniLM-L6 a more efficient choice for real-time systems.

Ethical Implications of NLP in Law Enforcement

All NLP/ML is based on historical data. If the training data has biases, it might display as inaccurate results after the deployment. Historical law enforcement narrative descriptions often reflect systemic biases related to sensitive features such as race, ethnicity, and socio-economic status (Shrestha & Yang, 2019). NLP/ML in policing raises concerns about fairness and bias in the sources and systems (Dixon & Birks, 2021). Agencies must audit retrieval results and monitor for bias. Additionally, deployment must adhere strictly to legal and regulatory frameworks governing data privacy, including PII and protected health information (Böhlin, 2024). There is a direct tension between maximizing retrieval utility and ensuring legal compliance. The required extensive redaction to remove PII, a legal necessity, inadvertently removes contextual richness from the police narratives.

Limitations

The datasets of crime reports and queries had 50 samples each, explicitly engineered to create a wide lexical-semantic gap. This highlighted the difference in retrieval performance but does not fully represent the scale, noise, and complex internal structure of a

real-world, large-scale police report database. The models were evaluated using only a single, high-stakes metric, Hit@3. While this is a practical metric for assessing a system's ability to retrieve the report, the evaluation may be improved by considering more metrics, such as normalized discounted cumulative gain or mean average precision.

Future Research

The embedding models used were trained on general web and language datasets. Domain-specific fine-tuning, adapted to the unique terminology and semantic relationships found in law enforcement documents, might perform even better (Chalkidis et al., 2020). Additionally, future research may investigate the magnitude of improvement achievable on a larger dataset (millions of crime reports). Future research may also explore hybrid approaches that combine the strengths of BM25 for exact matching with the semantic capabilities of embedding models.

Conclusion

Our comparative analysis suggests that embedding models may offer a measurable advantage over traditional keyword-based models in terms of retrieval accuracy. The MPNet model achieved significantly higher retrieval accuracy (Hit@3 of 100%) than BM25 (Hit@3 of 44%), a traditional keyword-based model, a difference of 56 percentage points. The other traditional keyword-based model, TF-IDF showed the lowest performance among the evaluated models, with a Hit@3 of 28%. This work points to the relatively limited retrieval accuracy of traditional keyword-based models due to vocabulary mismatch in complex law enforcement reports. Moreover, it emphasizes the need for smarter and more advanced search models in law enforcement. Interdisciplinary collaboration among computational linguists, criminologists, and legal experts may help guide the development of more effective retrieval systems that serve the pursuit of justice.

References

Al-Haddad, R., Al-Zubaidy, S., & Al-Taee, M. (2025). A hybrid information retrieval system for regulatory documents using BM25 and fine-tuned sentence transformers. *arXiv*. https://doi.org/10.48550/arXiv.2502.16767

Böhlin, R. (2024). LLM-based PII detection and remediation: Challenges and opportunities. *arXiv preprint*. https://arxiv.org/abs/2501.12465

Boubekeur, F., & Azzoug, W. (2013). Concept-based indexing in text information retrieval. *arXiv*. https://arxiv.org/abs/1303.1703

Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., & Androutsopoulos, I. (2020). Legal-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. https://doi.org/10.18653/v1/2020.findings-emnlp.261

Dixon, A., & Birks, D. (2021). Improving policing with natural language processing. In *Proceedings of the 1st Workshop on NLP for Positive Impact* (pp. 115-124). https://aclanthology.org/2021.nlp4posimpact-1.13.pdf

Güss, C. D., Tuason, M. T., & Devine, A. (2020). Problems with police reports as data sources: A researchers' perspective. Frontiers in Psychology, 11, Article 582428. https://doi.org/10.3389/fpsyg.2020.582428

Lovell, R. E., Klingenstein, J., Du, J., Overman, L., Sabo, D., Flannery, D., & Ye, X. (2022). *Using sentiment analysis and topic modeling in assessing the impact of police signaling on investigative and prosecutorial outcomes in sexual assault reports* (NCJ 306955). U.S. Department of Justice, National Institute of Justice. https://www.ojp.gov/pdffiles1/nij/grants/306955.pdf

Opitz, J., Möller, L., Michail, A., & Clematide, S. (2025). Interpretable text embeddings and text similarity explanation: A primer. *arXiv*. https://arxiv.org/abs/2502.14862

Parker, S. T. (2025). Assessing Supervised Natural Language Processing (NLP) Classification of Violent Death Narratives: Development and Assessment of a Compact Large Language Model (LLM) Approach. *medRxiv*, 2025-01. https://doi.org/10.1101/2025.01.16.25320680

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 3980–3990). https://arxiv.org/abs/1908.10084

Reimers, N., Gurevych, I., & Muennighoff, N. (2025). Sparse encoder—SentenceTransformers documentation. https://sbert.net/docs/package_reference/sparse_encoder/index.html

Robertson, S. E., & Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends® in Information Retrieval*, *3*(4), 333–389. https://doi.org/10.1561/1500000019

Sarzaeim, P., Mahmoud, Q. H., Azim, A., Bauer, G., & Bowles, I. (2023). A systematic review of using machine learning and natural language processing in smart policing. *Computers, 12*(12), 255. https://doi.org/10.3390/computers12120255

Schutz, P., & Hupfeld, T. (2021). BERT fine-tuning strategies for dense retrieval efficiency. *arXiv*. https://arxiv.org/abs/2109.10739

Shrestha, Y. R., & Yang, Y. (2019). Fairness in algorithmic decision-making: Applications in multi-winner voting, machine learning, and recommender systems. *Algorithms*, *12*(9), 199. https://doi.org/10.3390/a12090199

Steck, H., Ekanadham, C., & Kallus, N. (2024). Is cosine similarity of embeddings really about similarity? *arXiv*. https://arxiv.org/abs/2403.05440

Xing, X., & Chen, P. (2024). Entity Extraction of Key Elements in 110 Police Reports Based on Large Language Models. *Applied Sciences, 14*(17), 7819. https://doi.org/10.3390/app14177819

Zhelezniak, V., Parker, J., Smith, S., Saphra, N., Tsvetkov, Y., & Cotterell, R. (2019). Correlation coefficients and semantic textual similarity. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 1000–1010). https://aclanthology.org/N19-1100.pdf

Multimodal and Explainable Artificial Intelligence for Precision Healthcare: Integrating Federated Learning, Governance, and Affective Computing

Venus Sosa Iglesias
Westcliff University

Nand Janakbhai Modi
Westcliff University

Riba Sebastian Purackal
Westcliff University

Khushnood Shoukat
Westcliff University

Alicem Koyun, MSCS
Westcliff University

Maria Ndatiwelao Nangolo
Westcliff University

Kyungchan Park
Westcliff University

Abstract

Healthcare systems face increasing pressure from an aging population, rising rates of chronic disease and comorbidity, workforce shortages, clinician burnout, escalating care costs and fragmented digital infrastructure. Artificial intelligence (AI) has emerged as a transformative enabler to support descriptive, diagnostic, predictive, and prognostic big data analytics for personalized treatment planning, risk stratification, and longitudinal patient monitoring. However, the current AI paradigm is fragmented across clinical domains and is constrained by limited interoperability, insufficient external validation, algorithm opacity, and demographic bias. Governance and regulatory frameworks lag behind technological advancement, impeding AI adoption and eroding stakeholder trust. This narrative review uses a five-pillar framework for AI-enabled precision healthcare composed of (1) multimodal AI that integrates heterogeneous data sources; (2) explainable AI to improve interpretability, clinical accountability, and regulatory transparency; (3) affective computing and human-centered AI to create a therapeutic alliance with patients; (4) privacy-preserving infrastructures including federated learning (FL), differential privacy, and blockchain-enabled auditability to secure interinstitutional collaboration; and (5) adaptive governance systems for equitable, ethical, and sustainable deployment. Peer-reviewed scientific advances published between 2018 and 2026 are examined. The authors in this review argue that responsible and trustworthy AI-enabled precision healthcare should transition from isolated predictive models to complex socio-technical systems. Through the integration of these socio-technical systems into clinical workflows using adaptive ethical governance and privacy-preserving collaborative infrastructures, clinical reasoning and patient-centered care can be improved.

Keywords: Multimodal artificial intelligence (MAI), explainable AI (XAI), federated learning (FL), precision healthcare, human-centered AI, affective computing, adaptive governance

Introduction

In the twenty-first century, the prevalence of chronic illnesses and complex comorbidities has increased significantly, impacting how healthcare systems operate. In the United States alone, the population aged 65 years and older is projected to increase by 40% between 2022 to 2050, thus revealing the paramount need to provide continuous monitoring and individualized care management (Jones & Dolsten, 2024). Cancer, cardiovascular diseases, neurodegenerative disorders, mental health conditions, and diabetes continue to be major chronic diseases significantly contributing to morbidity and mortality worldwide, straining healthcare infrastructure (Hacker, 2024). Treatment strategies are shifting from generalized and population-based toward patient-specific approaches. Precision healthcare incorporates multidimensional, biological, behavioral, environmental, and clinical information to capture the heterogeneity and complex interactions among genetic, physiological, psychological, and social determinants of health.

Many hospitals operate under resource constraints, with outdated legacy infrastructure and persistent clinician shortages. These challenges contribute to sub-optimal patient care, workforce burnout, and escalating healthcare costs, which are projected to reach $47 trillion worldwide by 2030 (Hacker, 2024; Jones & Dolsten, 2024; Sipos et al., 2024). Moreover, emergency department (ED) overcrowding is a widespread phenomenon that dampens the ability to provide quality of care in a timely manner (Sartini et al., 2022). The increased documentation requirements, fragmented health information systems, and growing volumes of patient-generated data have intensified the operational burden of clinical workflows.

In this context, Artificial Intelligence (AI) has emerged as a transformative enabler of precision healthcare, supporting descriptive, diagnostic, predictive, and prognostic big data analytics to improve patient outcomes, care experience, and

healthcare cost reduction (Bajwa et al., 2021). Innovation across several domains has transformed translational research and has made healthcare more sustainable, efficient, and accessible for patients. 3D printing has enabled the creation of customized prosthetics, implants, and anatomical models to improve individualized therapy (Thacharodi et al., 2024). Telemedicine and remote patient monitoring systems such as wearable technology and the Internet of Medical Things (IoT) have made healthcare more accessible, especially in underserved areas, and facilitate real-time patient monitoring for preventive, personalized, and timely care. However, clinicians need support to interpret the vast and complex amount of model outputs generated (Thacharodi et al., 2024). Emerging technologies such as digital twins ranging from cellular to whole body system models, longitudinal risk forecasting, and adaptive monitoring systems provide opportunities for reactive medicine to transform into proactive, predictive, personalized, optimized, and continuously learning healthcare ecosystems (Khoshfekr Rudsari et al., 2025). Effective and efficient disease management will require a significant investment in scalable and innovative technological infrastructure, precision healthcare approaches that integrate longitudinal data streams, pharmaceutical interventions, and prevention strategies (Hacker, 2024).

Machine learning (ML) and deep learning (DL) are key drivers of AI-driven clinical transformation, demonstrating high performance in addressing systemic healthcare inefficiencies. Their applications span diverse clinical domains: radiology, pathology, ophthalmology, oncology, sepsis prediction, cardiovascular monitoring, emergency triage, and clinical decision support (Bajwa et al., 2021). A significant amount of research has focused on unimodal and single-disease prediction rather than integrating data from multiple modalities across time, which would allow holistic and dynamic patient management. Affective computing and emotionally-aware conversational systems have been used to model emotional context, empathy, and communication style to support global mental health needs, virtual care, and increase patient engagement, but many models have yet to create a therapeutic alliance with the patient (Schlicher et al., 2025). Despite advances in the field and high predictive performance, the “black box” problem of algorithm opacity remains a major challenge in understanding how algorithmic predictions are generated and whether such predictions are plausible biologically and clinically relevant (Mahajan & Helbing, 2026). Explainable artificial intelligence (XAI) techniques aim to improve interpretability and clinician trust; however, the explanations are often post hoc approximations that should progress into clinical integration for true mechanistic understanding and useful clinical decision support (K. Zhang et al., 2026).

Bias can arise at different stages in the AI lifecycle: in data features and labels, model development and evaluation, deployment, and publication. Biases include small sample size, data missingness, underrepresentativeness, inconsistent annotation practices, variability in data acquisition practices, measurement error, class imbalance, and cognitive and algorithmic bias. When left unaddressed, bias can result in substandard clinical decisions that perpetuate health disparities and limit generalizability (Cross et al., 2024). The current AI paradigm should be reoriented from short-term optimization toward systemic resilience, with stakeholder participatory design used to foster a symbiotic relationship between clinicians, patients, and AI-driven models.

AI adoption has been uneven across countries and business sectors, reaching approximately 14% by 2024 (since the launch of generative AI in late 2022) in large firms of the EU27 economies and countries from the Organization of Economic Co-operation and Development (OECD countries) (Kergroach & Héritier, 2024). The USA is the frontrunner in AI adoption, closely followed by China, due to strong core AI enablers including digital infrastructure, global supercomputing power, private investments, R&D, and upskilled workforce (Haag, 2025). Advanced foreign economies (e.g., EU, UK, Japan, and Canada) experience AI implementation challenges due to regulatory fragmentation and infrastructural constraints despite evolving governance. Regulatory and governance models lag behind technological advances. The European Union (EU) AI Act of 2024 has set the standard in responsible AI governance along with the FDA Good Machine

Learning Practice (GMLP), the European Health Data Space (EHDS), the Product Liability Directive (PLD), international projects (e.g., AICare@EU, SHAIPED, EU4Health), and global partners (e.g., WHO Europe, OECD, G7, G20) (Aboy et al., 2024; European Commission, 2025). Enforceability of policies, interoperability, transparency, fairness, accountability, and international harmonization remain challenging. Traditional centralized systems are prone to data breaches and single points of failure, whereas Blockchain Federated Learning (BCFL) frameworks provide an alternative for decentralized trust, incentive-driven participation, and privacy-enhancing infrastructures showing great potential in supporting precision healthcare, global health data collaboration, and large-scale AI deployment (Wang et al., 2026).

This review is guided by three research questions:

RQ1: How do multimodal and explainable AI systems improve predictive performance, interpretability, and clinical utility in precision healthcare?

RQ2: How can affective computing and human-centered AI improve patient-centered care, clinician interaction, trust, communication, and healthcare workflow integration?

RQ3: How do federated learning (FL) and adaptive governance frameworks support privacy-preserving collaboration, accountability, equity, interoperability, and sustainable implementation across international healthcare systems?

The authors propose to understand AI-enabled precision healthcare as a complex socio-technical system that integrates a Five-Pillar Framework that goes beyond isolated technical capability and includes (1) multimodal AI, (2) explainable AI, (3) affective computing, (4) privacy-preserving infrastructures, and (5) adaptive governance systems and institutional stewardship. Additionally, the authors evaluate the five-pillar interactions, trade-offs, and translational implications for responsible, equitable, patient-centered and continuously learning real-world clinical intelligence ecosystems.

Methods and Materials

This narrative review provided quantitative tabular evidence mapping for a comprehensive analysis of the AI-enabled developments in precision healthcare, aggregating conceptual, technical, clinical, ethical and governance-related literature to identify recurring patterns, trade-offs, and gaps in research literature. The review was organized using a socio-technical approach around three research questions that examine multimodal and explainable AI, human-centered AI systems, privacy-preserving infrastructures, and adaptive governance to enable international interoperability and regulatory compliance. Literature search was conducted across multiple interdisciplinary databases including ProQuest Central, PubMed, IEEE Xplore, and EBSCO. Literature searches used a combination of keywords including “multimodal AI”, “precision healthcare”, “explainable AI”, “affective computing”, “human-centered AI”, “federated learning”, “privacy-preserving AI”, “global or national AI regulatory frameworks or governance”, “healthcare AI regulations”, “adaptive governance”, and “clinical decision support”.

Studies were eligible for inclusion if they addressed at least one of the five pillars discussed in this review, were relevant to healthcare, included enough methodological detail for comparative analysis, were peer-reviewed sources published in English primarily between 2018 and 2026 to capture the most recent advances in our topic. A few unimodal studies were added when they represented domain benchmarks in precision healthcare relevant to AI robustness and clinical integration. Seminal theoretical and regulatory sources were added when contextualizing foundational concepts in the areas of technology, ethics, regulatory frameworks, and governance. Studies were excluded if they were unrelated to healthcare applications or lacked relevance to addressing our research questions, duplicated findings across studies without substantial added value, lacked methodological transparency, or were not reliable sources of information with adequate verification.

The analytical framework in this review was guided by a five-pillar socio-technical model to

organize multidimensional components in AI-enabled precision healthcare, including: Multimodal AI, Explainable AI and Interpretability, Affective and Human-Centered AI, Federated Learning and Privacy-Preserving Infrastructure, and Governance, Ethics, and Leadership. The comprehensive tabular analysis included information in the clinical domain and application, sample characteristics and preprocessing considerations, modality composition, fusion strategy, AI algorithm, performance metrics, explainability methods, validation strategies, implementation barriers and mitigation strategies, clinical utility and readiness, generalization risk, privacy-preserving methods, governance considerations, and sources of bias. Cross-pillar analysis specifically mapped system-level trade-offs, including accuracy versus interpretability, personalization versus privacy, technical innovation versus regulatory readiness, and automation versus clinical accountability, yielding an integrated understanding of the conditions required for responsible, equitable deployment. Given the heterogeneity of the study modalities, datasets, methodologies, and outcome measures reported, a narrative review was considered appropriate.

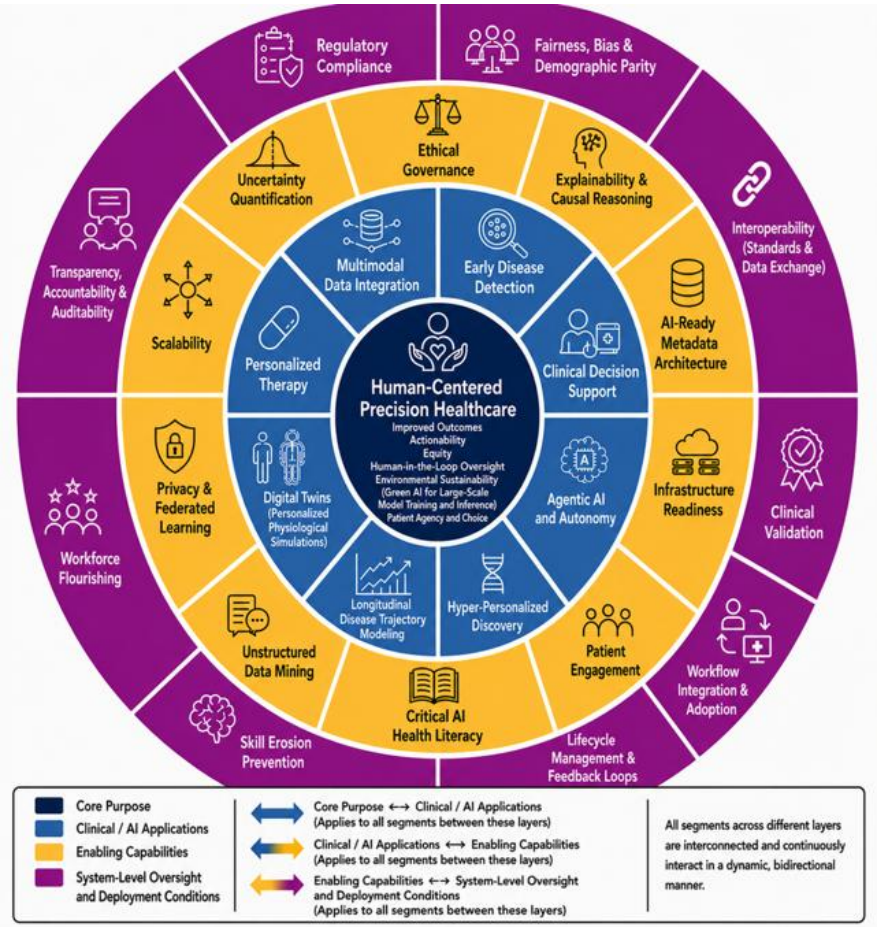
Results

Multimodal AI (MAI) in Precision Healthcare

Current healthcare practices are shifting from a “one-size-fits-all” prescription protocol to a quality-of-life, patient-centric approach, which constitutes the basis of Next-Generation Precision Medicine (ngPM) and MAI deployment (Mohammed et al., 2025). This section aims to evaluate how multimodal AI enhances model performance and clinical utility in precision healthcare (RQ1). Figure 1 provides an organizing framework for the review by illustrating the core purpose of AI, and how clinical applications, enabling capabilities (technical, organizational, and socio-cognitive requirements necessary for responsible implementation), and governance requirements interact in a human-centered precision healthcare ecosystem.

Figure 1

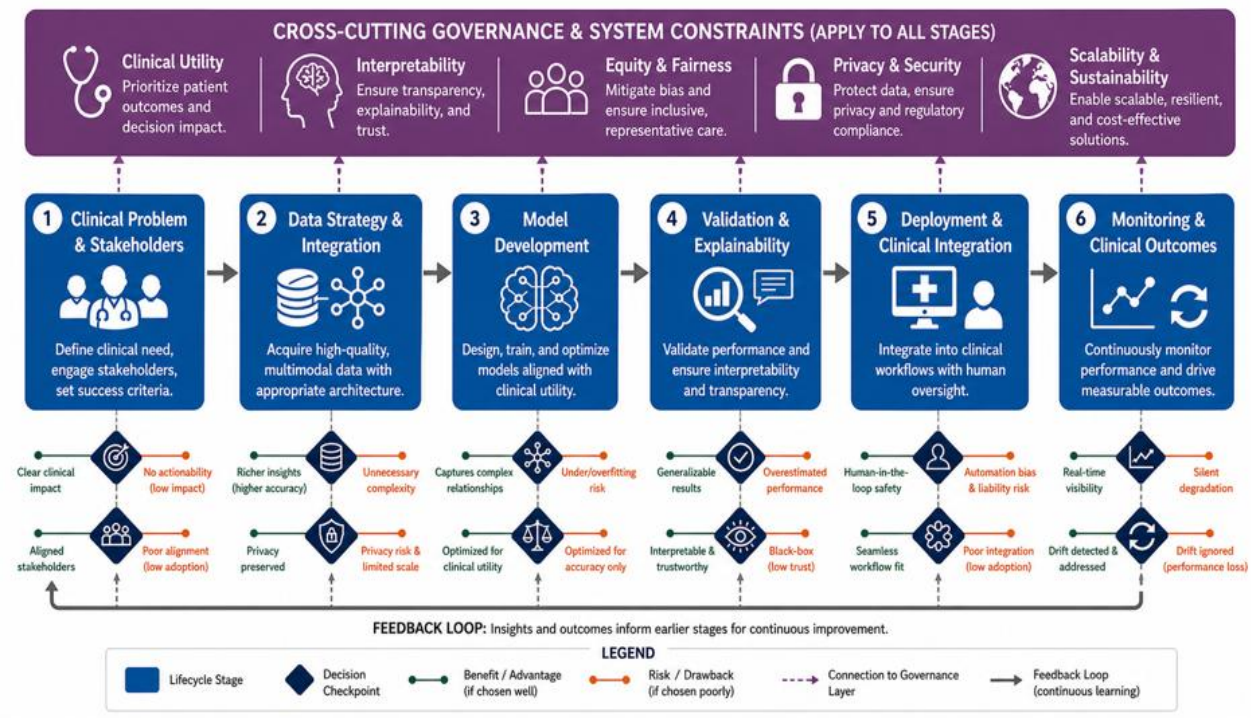
AI-Driven Applications and Implementation Framework in Precision Healthcare



Note. In the "clinical / AI applications" layer, hyper-personalized discovery refers to generative AI approaches to design DNA sequences or protein therapeutics tailored to an individual's molecular profile. In the enabling capabilities layer unstructured data mining using advanced natural language processing can be used to extract phenotypic insights from clinical notes. In the "system-level oversight and deployment conditions" layer, skill erosion prevention is proposed to mitigate overreliance on automated systems and workforce flourishing to emphasize the impact of AI on clinician job satisfaction and burnout. All segments across different ring layers are interconnected and continuously interact and co-evolve in a dynamic, bidirectional manner.

Figure 2

AI in Precision Healthcare: Lifecycle Pipeline, from Problem Definition to Clinical Outcomes, with Key Decision Checkpoints and Cross-Cutting Governance.



Note. This figure presents a streamlined decision-aware lifecycle pipeline for the development and deployment of artificial intelligence (AI) in precision healthcare, organized into six sequential stages. Each stage incorporates targeted decision nodes (diamonds) that highlight key decision checkpoints , with potential benefits (left) and risks associated with each choice (right). A cross-cutting governance layer spans the entire pipeline, underscoring the core considerations that must be addressed continuously. A feedback loop reinforces the iterative nature of clinical AI systems, enabling optimization and refinement of the model.

Data Integration and Fusion Challenges

Three fusion strategies build the architectural design of multimodal systems. Type I or early fusion is based on preprocessing concatenation of features and is sensitive to modality heterogeneity, temporal alignment, overfitting, and the curse of dimensionality. Type II or joint/intermediate fusion is based on co-learning of modalities within neural network layers using transformer model architectures and cross-attention mechanisms and requires well-aligned resources, high-quality datasets, and high-computational resources. Type III or late fusion is based on independent unimodal output ensembling through voting or weighted averaging, and is robust to missing or corrupted data, yet lacks cross-modal interaction (Guo et

al., 2019; Khan et al., 2026; R. Zhang et al., 2026). Multimodal fusion networks, even with synthetic low-quality images, perform better than unimodal high-quality images at certain noise levels, whereby internal fusion within the network (Type I-II) tends to outperform late-stage voting for tasks like tumor segmentation (even when using random forest) (Guo et al., 2019). However, there are several systemic challenges in static and coarse-grained fusion, including the temporal misalignment of longitudinal data, insufficient semantic leveraging, and the curse of dimensionality that may lead to AI-based attribution of scanner-specific artifacts as true pathophysiological signals (Acosta et al., 2022). Applying adaptive gating fusion and multi-head attention captures fine-grained relationships, mitigates bias, improves interoperability, and generalizability (Khan et al., 2026).

Modeling Approaches in MAI

The selection of the model architecture (based on fidelity, resolution, structure) should be aligned with specific clinical decision tasks. Convolutional Neural Networks (CNNs) remain widely used to detect spatial features (e.g., textures, edges, and contours). CNNs are highly effective for anatomical segmentation and delineation, microstructural analysis, image reconstruction, and improved computational radiology (MRI, CT, PET, and X-rays) (Guo et al., 2019; Li et al., 2023; Yoojin et al., 2025). Benchmark datasets for tumor segmentation (e.g., BRATS, KiTS, LiTS), and medical imaging (e.g., MURA, MedPix, NIH Chest X-rays) can be used to meet the CNN training requirement for extensive, high-quality annotated data (Pinto-Coelho, 2023). CNNs have contributed to early detection of disease and health factor analysis (Parvin et al., 2025).

Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) can be leveraged for the analysis of sequential and temporal physiological data, longitudinal electronic health records (EHRs), and social media data which enabled healthcare support during the COVID-19 outbreak and pharmacovigilance adverse event reporting (Awais et al., 2021; Cocos et al., 2017; Li et al., 2023). Autoencoders are unsupervised learning models used for anomaly detection and noise reduction in imaging and have been applied in longitudinal studies to evaluate the progression of age-related Macular Degeneration (Dupont et al., 2020).

Generative Adversarial Networks (GANs) contribute to data augmentation and noise mitigation, allowing image-to-image translation and image quality enhancement (Li et al., 2023; Pinto-Coelho, 2023). GANs have been used to accurately predict the length of stay of patients at emergency departments, improve dermatological skin segmentation beyond the performance of the SMARTSKINS benchmark, and predict drug repurposing for breast cancer (Andrade et al., 2021; Cui et al., 2021; Kadri et al., 2022). However, GANs have several challenges, including quality control of synthetic images, training instability, and the potential introduction of artifacts (Li et al., 2023). Hybrid models leverage complementary strengths, integrating spatial and temporal analysis or reconstruction of heterogeneous data to support the analytical complexity of MAI (Banerjee et al., 2019; Khoshdel et al., 2020; Li et al., 2023).

Transformer-based architectures use attention mechanisms to elucidate dynamic spatial and temporal interactions, but their effectiveness is dependent on the training dataset quality and memory costs (X. Li et al., 2025). To illustrate the diversity in AI applications across clinical domains, Table 1 (expanded in Supplementary Table 1) summarizes representative studies highlighting modality composition, fusion strategies, AI algorithms, validation approaches, performance metrics, deployment limitations, and translational readiness, reporting high predictive performance but facing several challenges for clinical translation.

Table 1

Comparative Evaluation of MAI Systems for AI-Driven Precision Healthcare Across Clinical Domains (2018–2026)

Domain / Representative Study	Clinical Task	Data Configuration	Fusion Strategy	AI / XAI	Validation	Key Performance	Clinical Utility	Main Limitations	Translational Readiness
Cardiology (Demir et al., 2025)	CHD risk prediction	Clinical + laboratory + demographic data (Low burden; n=4,543)	Early fusion	SVM/XGBoost /LSTM/ANN; FI	Internal split	*AUC* 0.93; *F1* 92.8%	Long-term cardiovascular risk stratification	Public benchmark dependence; demographic bias	Validation
Cardiology (Safak et al., 2026)	NSTEMI occlusion prediction	12-lead ECG spectrograms (Medium; n=907)	Intermediate fusion	RAL-CNN; Grad-CAM/SU E	10-fold CV	*Acc* 81.3% vs 65.3% CNN	Rapid invasive-triage support	Single-center; class imbalance	Early validation
Cardiology (Kim et al., 2025)	Sleep-based CVD screening	ECG + airflow + SpO2 (Medium; n=194)	Intermediate fusion	1D-CNN	Internal split	*Acc* 97.6%; *F1* 0.96–0.97	Automated nocturnal CVD screening	Homogeneous cohort	Validation
Cardiology (Revathi et al., 2026)	CAD detection from CCTA	CCTA + texture/deep features (High; n=206)	Intermediate fusion	XAI method used. Mask R-CNN/DeepC onvNet	Internal CV	*Acc* 98.3%; *AUC* 0.98	Coronary segmentation and CAD classification	Annotation burden; limited external testing	Validation
Chronic disease management (Sirapangi & Gopikrishnan, 2024)	IoT risk monitoring	Wearables + environment + demographics (Medium; n=35k)	Early fusion	RNN + FFO/FCM; DeepSHAP	Internal split	*Acc* 94.2%; *AUC* 0.92	Continuous risk monitoring	Noise/privacy concerns	Prototype
Critical care (Nemati et al., 2018)	Sepsis prediction	EMR + bedside vitals (Medium; n≈69k)	Intermediate fusion	AISE; contribution visualization	Internal + external	*AUROC* 0.87	Early ICU sepsis warning	Recall bias; monitoring dependence	Prospective validation

Domain / Representative Study	Clinical Task	Data Configuration	Fusion Strategy	AI / XAI	Validation	Key Performance	Clinical Utility	Main Limitations	Translational Readiness
Critical care (Lundberg et al., 2018)	Hypoxaemia prediction	EHR + intraoperative vitals (Low; n>50k)	Early fusion	XGBoost; SHAP	Held-out test set	*AUC* 0.81 vs 0.66 clinicians	Perioperative desaturation prevention	High feature dimensionality	Validation
Critical care / ED (Lauritsen et al., 2020)	Early warning system	EHR labs + vitals (Low; n=163k)	Early fusion	TCN; DTD	Internal: 5-fold CV	*AUROC* 0.79–0.92	Acute deterioration prediction	Limited ethnic diversity	Validation
Emergency medicine (Russ et al., 2025)	ED workflow assistant	Speech + vitals + history (Low; pilot n=20)	Intermediate fusion	Local LLM; clinician-reviewed explanations	Pilot feasibility	*SUS* 90.6	Triage/documentation support	Small pilot cohort	Prototype
Emergency medicine (Kadri et al., 2022)	Length-of-stay prediction	Retrospective EHR/labs (Low)	Early fusion	GAN-P	Internal cohort	R^2 0.87	Resource allocation and overcrowding prevention	Single-site retrospective data	Validation
Gastroenterology / Oncology (Seza et al., 2025)	Pancreatic cyst differentiation	EUS + CT + MRI + sex (High)	Intermediate fusion	ResNet	Internal	*Acc* 99% vs 81% experts	Surgical decision support	Small cohort; modality burden	Prototype
Hepatology (Peng et al., 2021)	Hepatitis deterioration risk	Clinical + laboratory data (Low; n=155)	Hybrid early + late fusion	RF/XGBoost/SVM; SHAP/LIME/PDP	K-fold CV	*Acc* 91.9%	Personalized deterioration risk prediction	Small benchmark dataset	Validation
Infectious disease (Malik & Rathee, 2024)	COVID-19 screening	Demographics + symptoms + exposure (Low; n>50k)	Early fusion	LightGBM; SHAP	Held-out test set	*Acc* 92.8%	Pandemic triage and screening	Self-reporting bias	Validation

Domain / Representative Study	Clinical Task	Data Configuration	Fusion Strategy	AI / XAI	Validation	Key Performance	Clinical Utility	Main Limitations	Translational Readiness
Laboratory medicine (Hu et al., 2023)	Plasma cell dyscrasia diagnosis	IFE images (High; n=12,703)	Late fusion ensemble	CNN ensemble; Score-CAM	Internal + external	*Acc* 99.8%	Automated laboratory interpretation	Rare-pattern scarcity	Deployment-ready
Longitudinal care / CDSS (Niu et al., 2025)	Longitudinal disease prediction	Clinical notes + label descriptions (Low; n=9,759)	Intermediate fusion	DSSM + attention	Internal multi-dataset	*F1* 0.90; *AUROC* 0.88	Trajectory modeling across visits	Sparse/noisy text	Validation
Mental health (Khan et al., 2026)	Depression detection	Audio + video + text (High)	Intermediate fusion	Transformers; SHAP/attention maps	Subject-independent split	*Acc* 93%; *F1* 91.4%	Continuous mental health screening	Real-world noise; fairness concerns	Validation
Molecular biology (Tunyasuvunakool et al., 2021)	Protein structure prediction	Sequence + structural templates (High)	Intermediate attention fusion	AlphaFold/Evoformer; pLDDT	External benchmark	*AUC* 0.90	Structure-function and drug discovery support	Extreme computational cost	Deployment-ready
Neurology (Yu et al., 2024)	Alzheimer's diagnosis	sMRI + clinical + genomic data (High; n=1,651)	Intermediate fusion; unified transformer	CNN-Transformer; Grad-CAM/SHAP	5-fold CV	*AUC* 0.99	Early AD/MCI diagnosis	Limited racial diversity	Validation
Neurology (Gu et al., 2025)	NMO diagnosis	Fundus images + MRI/labs (High)	Late fusion	Inception-V3; Grad-CAM	Internal CV	*AUC* 0.99	Specialist diagnostic support	Single-center cohort	Validation
Neuropsychiatry (Rahaman et al., 2024)	Schizophrenia classification	sMRI + fMRI + SNPs (High; n=437)	Intermediate attention fusion	BAM multimodal network	Repeated splits	*Acc* 94.1%	Biomarker discovery	High-dimensional noisy data	Research framework
Nutrition (Yan et al., 2025)	Dietary estimation	Food images + nutrient DB (Medium)	Adaptive RAG fusion	GPT-Vision + RAG	Internal datasets	98.9% recognition success	Prec. nutrition + dietary monitoring	Controlled acquisition settings	Prototype/validation

Domain / Representative Study	Clinical Task	Data Configuration	Fusion Strategy	AI / XAI	Validation	Key Performance	Clinical Utility	Main Limitations	Translational Readiness
Oncology (Tabl et al., 2019)	Breast cancer biomarker prediction	Genomics + clinical data (High; n=347)	Early fusion	Hierarchical RF	10-fold CV	*Acc* 80.9–100%	Precision treatment planning	Small imbalanced cohorts	Validation
Oncology (Partin et al., 2023)	Drug response prediction	Gene expression + histology + drug descriptors (High)	Intermediate fusion	MM-Net; CNN	Repeated CV	*AUROC* 0.80	Precision oncology screening	Limited sample size	Prototype
Oncology (Guo et al., 2019)	Sarcoma segmentation	PET + CT + MRI (High; n=50)	Early fusion	CNN/ U-Net	10-fold CV	*Dice* 0.85	Radiotherapy planning	Small datasets; registration errors	Prototype
Oncology (Emegano et al., 2025)	Prostate cancer histopathology	Biopsy WSIs (High; n=1,276)	Unimodal	ResNet50; Grad-CAM	Held-out validation	*Acc/AUC* 0.98	Standardized pathology workflows	Lack of multicenter validation	Validation
Oncology (Abbas et al., 2024)	Federated cancer classification	Histology + CT + MRI (Medium–High)	Adaptive federated fusion	CNN + adaptive FL	Simulated multi-site	*Acc* 90%	Privacy-preserving collaborative learning	Lack of non-IID/domain-shift issues	Prototype
Oncology (Deacon et al., 2025)	Intraoperative methylome profiling	Nanopore genomic sequencing (High; n=50)	Intermediate fusion	Neural networks	Prospective external	76% classified within 1 h	Rapid intraoperative profiling	Workflow complexity	Research-use validation
Oncology (Luo et al., 2019)	GI cancer detection	Endoscopic imaging (Medium–High; >1M images)	Late fusion/cloud inference	GRAIDS DL system	Prospective multicenter	*Acc* 0.92–0.98	Real-time endoscopic cancer detection	Infrastructure dependence	Deployment-ready

Domain / Representative Study	Clinical Task	Data Configuration	Fusion Strategy	AI / XAI	Validation	Key Performance	Clinical Utility	Main Limitations	Translational Readiness
Oncology / Multi-domain (Makarov et al., 2025)	Digital twin prediction	Longitudinal EHR + vitals + LLM inputs (Low; n>52k)	Early fusion	DT-GPT/BioMistral	Multi-dataset benchmarking	*AUC* 0.70; *MAE* 0.55	Predictive clinical decision support	Data sparsity and retrospective bias	Validation
Oncology / Neurology (Shin et al., 2023)	Brain tumor triage	Multiparametric MRI (High; n=877)	Late hierarchical fusion	Hierarchical CNN; LRP	Held-out testing	*AUC* 0.90	Specialist triage/referral support	Single-center retrospective design	Validation
Oncology / Neurology (Hewitt et al., 2023)	Glioma subtype prediction	Whole-slide histology (High; n=2,845)	Sequential late fusion	attMIL DL; heatmaps	External cohorts	*AUROC* 0.95	Molecular subtype prediction	Scanner/domain shift	Advanced validation
Ophthalmology (Ma et al., 2025)	Self-triage chatbot	Text + smartphone/slit-lamp images (Low–Medium; n=15,640)	Late fusion	ChatGPT-3.5 + ResNet50	External multicenter	*Acc* 81.1%	Ophthalmic self-triage	Performance drops with incomplete input	Validation
Pediatrics / Psychiatry (Albahri et al., 2024)	Autism triage	Clinical + sociodemographic criteria (Medium; n=538)	Early fusion	Logistic regression; LIME	10-fold CV	*F1* 0.98; *AUC* 1.00	Early autism prioritization	Imbalanced single-center data	Prototype
Psychiatry (Guleria, 2025)	Medical transcript classification	Clinical text (Low; n=4,998)	Early fusion	BERT + LSTM	Internal validation	*Acc* 94%; *F1* 90%	Automated clinical NLP coding	Single-source dataset	Validation

Domain / Representative Study	Clinical Task	Data Configuration	Fusion Strategy	AI / XAI	Validation	Key Performance	Clinical Utility	Main Limitations	Translational Readiness
Psychiatry (Chen et al., 2024a)	Passive depression assessment	Actigraphy + app usage + voice + NLP (Low; n=183)	Intermediate fusion	ANN	Internal: Leave-one-out CV	*F1* 0.81	Passive mental health monitoring	Small cohorts/privacy concerns	Prototype
Pulmonology (Biswas et al., 2025)	Pneumonia detection	CXR + synthetic augmentation (Medium; n=8,402)	Intermediate fusion	VGG16/MobileNet + SVM; Grad-CAM/LIME	Internal; Federated environment	*Acc* 97.6%; *F1* 98.4%	Privacy-aware automated screening	Lack of data heterogeneity & domain shift	Validation
Pulmonology (Veeramni et al., 2025)	Multiclass lung disease diagnosis	CXR (Medium; n=7,560)	Intermediate fusion of feature maps	Inception-V3 + U-Net; LIME/Grad-CAM	Internal	*Acc* 97.5%; *F1* 0.98	Rapid infectious lung disease screening	Public dataset dependence	Validation

Note. Studies were condensed using standardized descriptors to improve cross-study comparison and publication readability. Fusion approaches were classified as described in Guo et al., 2019. Abbreviations: Acc, accuracy; AD, Alzheimer's disease; AUC/AUROC, area under the receiver operating characteristic curve; BAM, bottleneck attention module; CAD, coronary artery disease; CCTA, coronary computed tomography angiography; CDSS, clinical decision support system; CV, cross-validation; CVD, cardiovascular disease; DSSM, deep state-space model; DTD, Deep Taylor Decomposition; ECG, electrocardiography; EMR, electronic medical record; F1, F1-score; FI, feature importance; FFO, firefly optimizer; GI, gastrointestinal; LIME, local interpretable model-agnostic explanations; LLM, large language model; MRI, magnetic resonance imaging; NMO, neuromyelitis optica; NLP, natural language processing; PET, positron emission tomography; RAG, retrieval-augmented generation; RF, random forest; SHAP, SHapley additive explanations; SMOTE, synthetic minority oversampling technique; SNP, single nucleotide polymorphism; SUS, system usability scale; TCN, temporal convolutional network; WSI, whole-slide imaging.

Interoperability and Data Standardization Challenges

A major challenge to interoperability is fragmented data ecosystems, where hospitals use outdated legacy infrastructure and siloed protected data using formats not conducive for research, inconsistent naming conventions, and semantic mismatch (Jasodanand et al., 2025; Simon et al., 2025). This challenge can be addressed using AI-first frameworks and providing technical alignment using standards set by HL7 Fast Healthcare Interoperability Resources (FHIR), Observational Medical Outcomes Partnership (OMOP), and the Digital Imaging and Communications in Medicine (DICOM) Supplement 145 (Sakaguchi et al., 2025). The harmonization of OMOP Common Data Model (CDM)-on-HL7 FHIR, enhances interoperability of clinical datasets while maintaining semantic consistency (Jayathissa et al., 2025; Papachristou et al., 2024). Additionally, the International Organization for Standardization (ISO), through its Technical Committee 215 (ISO/TC 215) on Health Informatics, has assisted in providing consensus in public and private sectors in 160 countries (Orlova et al., 2017).

Performance Gains vs. Translational Gaps

Multimodal systems with low-quality images show consistently superior performance and noise resilience over high-quality unimodal benchmark systems, suggesting that a redundancy across modalities can compensate for data degradation. However, performance gains are not universal (Guo et al., 2019; X. Li et al., 2025; Marouf et al., 2025). Scaling MAI requires large computing power, storage capacity, and updated high-resource infrastructures with cloud-based Information Management Systems (IMS) to process concurrent modalities or edge computing to reduce latency in critical care tasks, thus creating global inequities, excluding low-resource organizations from accessing ngPM (Papachristou et al., 2024; Zhai et al., 2022). Additionally, global scalability requires aligning early both with regulatory experts and standardization frameworks (FHIR, EHDS, TRIPOD+AI, PROBAST+AI) and following FAIR (Findable, Accessible, Interoperable, Reusable) principles and ROBIN criteria (Kazerooni et al., 2025).

Spurious cross-modal correlations non-existent biologically, modality dominance, and cross-modal bias amplification lead to both inequitable health recommendations and dataset shift that fails to generalize to diverse global populations (Acosta et al., 2022; X. Li et al., 2025; Makarov et al., 2025). Non-Independent and Identically Distributed (non-IID) multimodal clinical data is affected by differences in institutional protocols, equipment manufacturers, and local demographics. There is a translational divide between high-accuracy models trained on clean research datasets and clinical validation on external cohorts due to institutional bias or device variability during workflow integration (Parvin et al., 2025; L. Zhang et al., 2026). Models with robust performance in multicenter external validation and demonstrated safe implementation in prospective clinical trials is limited because many models rely on internal resampling that limits generalizability and are susceptible to overfitting (Sakaguchi et al., 2025).

Bias, Missingness, and Failure Modes

In multimodal deployment, bias is amplified unless addressed through fairness regularization. Bias comes from patient demographics, modality, curated versus real-world data, algorithmic and cross-species (R. Abbas et al., 2025). Clinical data is often imbalanced, underrepresented, and incomplete, where EHRs are extremely sparse, and missingness may lead to exclusion before training, potentially introducing selection bias (Acosta et al., 2022). Missingness and corrupted datasets may result in blind spots and inequitable outcomes. The statistical power of genome-wide association studies can be improved by multiple imputation pervasive preprocessing facilitated by using large reference datasets with deep genotypic coverage. However, imputation can introduce medical noise if not paired with clinical logic (Acosta et al., 2022). Innovative AI techniques deal with missing data natively, through random feature masking modeling, attention-based weighting mechanisms, and uncertainty-aware inference practices (Jasodanand et al., 2025). Models should consider designs with modality-specific dropout or generative components capable of synthesizing plausible data for missing values to avoid a decrease in model performance (Khan et al., 2026).

From Technical Capability to Clinical Utility

Achieving successful clinical utility requires a socio-technical system balancing technical innovation, ethical governance, and patient-centered approach, with human-in-the-loop designs using AI as a Clinical Decision Support (CDS) tool

and regulatory compliance aligned with EU AI Act, FUTURE-AI, and FDA guidelines that require both external validation and post-market monitoring (Dias Cabaço & Rodrigues, 2026; Lekadir et al., 2025; Makarov et al., 2025; Parvin et al., 2025). The computational burden of processing high-resolution multimodal data and the extended duration for AI-based device regulatory validation create significant bottlenecks (Huang et al., 2025). Most models integrating blockchain remain at a proof-of-concept stage validated in controlled laboratory environments, that is, Technology Readiness Level 3 (TRL3), and only a few progress to prototype validation (TRL4–TRL5) in real-world deployment (K. Li et al., 2025). While MAI enhances predictive capability through data integration, the increased model complexity raises concerns in terms of interpretability and clinical accountability, thus driving interest in explainable AI mechanisms to make model output more transparent and enable clinically accountable reasoning.

Explainable AI (XAI) in Precision Healthcare

XAI is essential for precision healthcare because it translates complex model outputs into interpretable evidence useful for clinicians (Gerdes, 2024; Räz et al., 2025). This section examines how XAI supports model performance, clinician confidence, and accountable decision-making in precision healthcare (RQ1). As summarized in Table 2, explainability operates differently across model families and clinical applications, where SHAP, LIME, saliency maps, and related XAI methodologies have been useful in disease prediction, risk stratification, and clinical decision support. Benchmarks included AUC values around or above 0.80 for discrimination tasks, referral accuracy around 70% when benchmarked against radiologists, or accuracy above 90% in constrained imaging or laboratory classification tasks. Additionally, findings required support from external validation, expert comparison, or biologically coherent explanation maps. While clinical applicability has increased in the real world as the use of explainable AI shifts from a conceptual foundation to workflow integration, reproducibility remains limited because 29% of recent studies shared codebases and less than 10% performed cross-dataset reproducibility testing (Q. Abbas et al., 2025). While the use of high-dimensional complex architectures optimizes predictive accuracy, the “black box” nature of DL struggles to explain MAI inter-variable dependencies and is a barrier for clinical adoption, where there is a distinction between plausibility of explanations that make sense to a clinician versus faithfulness of explanations accurately reflecting the model’s logic (Q. Abbas et al., 2025). Table 3 compares different XAI methodologies, their strengths, limitations, and best-fit clinical uses in precision healthcare.

Table 2

Comparative Evidence on Original XAI Studies, Sample and Validation Design, and Clinical Relevance in Precision Healthcare (2021–2025)

Clinical Task	Model and XAI method	Sample and validation design	Key results	Key limitation	Clinical Relevance	Study
AML drug-response stratification and treatment recommendation	Multi-dimensional Module Optimization with a deterministic decision-tree treatment logic	BeatAML cohort of 319 ex vivo tumor samples with 10-fold cross-validation and external validation in CERES, DEMETER2, and GDSC	The pipeline identified biomarker-linked treatment paths and validated drug-response associations across external datasets, including $p = 5.5 \times 10^{-5}$ in CERES, $p = 6.8 \times 10^{-6}$ in DEMETER2, $p = 5.5 \times 10^{-4}$ in GDSC, and a 30-fold lower IC50 in one FLT3-targeted example	Later decision branches involved small biomarker-defined subgroups and depended on *ex vivo* screening availability	Shows how explainable treatment logic can be embedded into precision oncology stratification	Gimeno et al., 2022
Hepatitis deterioration prediction	Random forest, XGBoost, SVM, logistic regression, decision tree, KNN with SHAP, LIME, and PDP	The study used the UCI hepatitis benchmark with 155 instances; the article did not clearly report the exact train-test split in the accessible text	Best-performing random forest reached 91.9% accuracy, and SHAP, LIME, and PDP exposed global and local feature effects	Limited methodological detail in accessible text and no clear external validation	Useful example of interpretable risk prediction in structured data	Peng et al., 2021
Myocardial infarction prediction	XGBoost with SHAP versus logistic regression	UK Biobank cohort of 502,506 participants with 90% training and 10% test split	*ROC* 0.86 (XGBoost) versus 0.77 (Logistic Regression), although accuracy was 0.75 XGBoost) versus 0.77 (Logistic Regression)	Transportability beyond the source cohort remained uncertain	Strong case for explainable cardiovascular risk stratification	Moore & Bell, 2022
Alzheimer's disease classification from FDG-PET	3D CNN with saliency map and LRP	2,552 FDG-PET scans from 836 subjects; subject-level split with 20% test set and the remaining 80% divided 80-20 into training and validation, preserving class distribution across 5 trials	Average test *AUC* was 0.81 for cognitively normal, 0.63 for mild cognitive impairment, and 0.77 for Alzheimer's disease, with LRP mapping anatomy more effectively than saliency maps	No external validation; heatmaps did not equal biologic proof	Supports cautious use of *post hoc* explanation in neuroimaging.	De Santi et al., 2023

Clinical Task	Model and XAI method	Sample and validation design	Key results	Key limitation	Clinical Relevance	Study
DAT-SPECT classification in parkinsonian syndromes	3D CNN with LRP	Final dataset of 1,296 scans split into 864 training, 144 validation, and 288 test cases	Test performance reached 95.8% accuracy, 92.8% sensitivity, 98.7% specificity, 98.5% positive predictive value, and 93.7% negative predictive value, with LRP maps generated in about 3 seconds and highlighting the most affected putamen	Reference standard was expert interpretation rather than etiologic confirmation	Useful as an objective second reader in nuclear medicine	Nazari et al., 2021
Brain lesion referral triage using MRI	Deep learning triage system with LRP	Developed in 747 patients and externally validated in 130 emergency patients	The system gave correct referral suggestions in 94 of 130 patients (72.3%), closely matching radiologists at 72.6%; achieved *AUC* 0.90 and *AUPRC* 0.94 for tumour discrimination; and showed a *Dice* coefficient of 0.77 versus 0.33 for high-relevance overlap in tumours versus non-tumors.	*Post hoc* explanation still required specialist oversight	Clinically relevant because explanation informed referral-level decisions	Shin et al., 2023
Immunofixation electrophoresis interpretation	CNN ensemble with Score-CAM	12,703 expert-annotated images with five-fold cross-validation and generalizability testing	Coarse recognition accuracy reached 99.30%, average fine-pattern accuracy reached 99.82%, and external testing accuracy reached 99.81%, with visually reviewable band localization	Single-center data and limited rare-pattern coverage	Shows XAI value in laboratory diagnostics, not only in imaging	Hu et al., 2023
Multi-omics kidney cancer module detection and survival stratification	Explainable Greedy Decision Forest with TreeSHAP	TCGA multi-omics kidney cancer tasks with repeated 80-20 train-test splits performed 20 times, 500 random walks, and out-of-bag module selection; exact patient counts were not explicitly reported in the article text	The top survival module reached out-of-bag accuracy 0.72 and test accuracy 0.69, whereas the kidney-versus-other-cancers task reached test accuracy 0.78, with biologically coherent modules and TreeSHAP-based attribution	Performance depended on predefined network structure and small held-out clinical subsets, limiting direct bedside generalizability	Useful for biologically traceable multi-omics stratification, but still upstream from direct clinical deployment	Pfeifer et al., 2022

Table 3

Comparison of Explainable AI Methods, Strengths, Key Limitations, and Best-Fit Clinical Uses in Precision Healthcare (2021–2025)

XAI method	Explanation mechanism	Best-fit data/model context	Application	Main limitation/failure mode	Best-fit clinical use	Representative reference
SHAP	Assigns contribution values to features using Shapley-value logic; supports both local and global feature attribution.	Structured EHR-style data, tabular risk models, laboratory variables, multi-omics features, and tree-based models.	Clinicians need to know which variables most influenced a prediction at both the individual-patient and cohort level.	Can appear mathematically precise even when features are correlated, causally ambiguous, or weakly transportable across populations.	Risk prediction, prognosis, structured clinical decision support, and audit-ready feature attribution.	Moore & Bell, 2022; Peng et al., 2021; Pfeifer et al., 2022
LIME	Builds a simplified local surrogate model around one prediction to approximate why the model made that case-level decision.	Heterogeneous tabular or mixed-data models where individual case explanation is needed.	The goal is patient-specific clarification rather than explaining the entire model globally.	Sensitive to perturbation strategy, sampling choices, and local-neighborhood definition; explanations may vary across runs.	Case-level clarification, clinician review of individual risk predictions, and second-opinion decision support.	Peng et al., 2021
LRP	Backpropagates relevance through neural-network layers to identify which input regions contributed to the output.	CNN-based medical imaging models, especially PET, MRI, SPECT, and other neural-network imaging applications.	Clinicians need spatially localized relevance maps connected to image regions or anatomical structures.	Anatomically plausible heatmaps do not guarantee biological or causal faithfulness; outputs depend on architecture and propagation rules.	Radiology, nuclear medicine, neuroimaging, and image-based diagnostic review.	De Santi et al., 2023; Nazari et al., 2022; Shin et al., 2023
CAM-family visual localization *(Score-CAM)*	Uses class-activation maps to highlight image regions that influence a model's classification decision.	CNN-based image classification or laboratory-imaging tasks where visual localization is clinically meaningful.	The clinical task requires visual confirmation of influential regions, bands, lesions, or image patterns.	May remain coarse, architecture-dependent, and insufficient for explaining non-image features or multimodal interactions.	Specialist verification of image, band, lesion, or scan classification.	Hu et al., 2023
Framework-level XAI orchestration / clinician-facing explanation architecture	Combines multiple explanation outputs, uncertainty information, clinical context, and user-facing explanation interfaces.	Multimodal, longitudinal, or multidisciplinary care settings where no single XAI method is sufficient.	Clinical reasoning requires layered explanation, uncertainty handling, follow-up interpretation, or multidisciplinary review.	Evidence remains more conceptual; many frameworks lack standardized validation metrics, usability testing, or prospective clinical evaluation.	Multistep decisions, follow-up care, tumor boards, complex multimodal review, and multidisciplinary interpretation.	Pahud de Mortanges et al., 2024; Räz et al., 2025

Note. SHAP, Shapley additive explanations; LIME, local interpretable model-agnostic explanations; LRP, layer-wise relevance propagation; Score-CAM, score-weighted class activation mapping; PET, positron emission tomography; MRI, magnetic resonance imaging; SPECT, single-photon emission computed tomography.

XAI Applications in Diagnosis, Imaging, and Risk Prediction

In multimodal systems, explainability is especially important because model output predictions may reflect interactions among imaging, genomics, EHRs, wearable signals, and behavioral features. XAI bridges model performance and clinical judgment by enabling clinicians to assess plausibility in informing care (Gerdes, 2024; Räz et al., 2025). In a UK Biobank study, XGBoost predicted myocardial infarction from health data and outperformed logistic regression. SHAP explanations highlighted key risk factors such as waist circumference, blood pressure, and sex, supporting clinician understanding of the prediction process (Moore & Bell, 2022). XAI was also applied to predict the progression of hepatitis, where a highly accurate random forest model used SHAP and LIME to demonstrate which variables increased individual patient risk (Peng et al., 2021). XAI is especially useful in diagnostic imaging procedures, such as MRI, PET, and SPECT. CNN was used to analyze Alzheimer's data, where PET scans were integrated. The model predicted disease patterns, and the layer-wise relevance propagation showed the most correlated brain areas, enabling physicians to get a clear perspective on raised concerns regarding some cases of Alzheimer's disease (De Santi et al., 2023). In another study, XAI was implemented for clinically uncertain Parkinsonian syndromes using SPECT scans of dopamine transporters, where an accurate model produced relevance maps that highlighted affected brain regions and supported diagnostic confirmation (Nazari et al., 2021). Explainability is particularly well suited to medical imaging where clinicians routinely interpret visual evidence; consequently, making AI-generated outputs such as heatmaps and highlighted regions more comprehensible and clinically actionable.

In precision healthcare, AI can analyze large genomic datasets to gain insights on tumor biology and treatment response patterns faster than humans, yet decisions must remain justifiable and trustworthy. In acute myeloid leukemia research, an XAI system was used to link genomic biomarkers with drug reactions and assist with medically viable treatment selections (Gimeno et al., 2022). A renal cancer research study used Greedy Decision Forest and TreeSHAP to identify biologically meaningful cancer classifications and predict survival rates (Pfeifer et al., 2022). Score-CAM in combination with DL models has been used successfully to identify electrophoretic band patterns derived from immunofixation samples, achieving expert-level performance in laboratory medicine (Hu et al., 2023).

Limitations: Illusion of Transparency and Over-Reliance

When healthcare professionals can visualize which symptoms or biomarkers drove a diagnostic conclusion, they feel more confident to question, contextualize, and communicate those results to patients. Despite these benefits, XAI has several challenges. Some explanations may appear convincing while failing to represent the model's underlying decision process, thus creating a fidelity gap, or the explanations might not constitute the actual cause of the prediction, thus creating a plausibility gap. This can create a false sense of confidence in potentially biased or erroneous decisions (Räz et al., 2025). Second, many AI models are tested using data from a single hospital or geography, so findings may not generalize across other populations or health systems, making external validation essential before large-scale clinical use (Shin et al., 2023). Third, combining genetics, images, and reports raises privacy and technical challenges, and decisions based on multiple data sources are harder to explain than those from a single-input model (Pahud de Mortanges et al., 2024). Finally, some explainable methods may be too technical or impractical for busy clinicians. Research into XAI is moving toward systems that are interactive and adaptable, providing plausible justifications for a specific patient's history and current health condition (Noor et al., 2025). Beyond the improved predictive and cognitive transparency capabilities, AI models are expected to interact with humans, patients and clinicians, thus emphasizing the need for improved communication, empathy, and user experience driving the emergence of affective computing and human-centered AI.

Affective Computing and Human-Centered AI

This section addresses how affective

computing and human-centered AI improve patient-centered care, clinician interaction, trust, communication, and healthcare workflow integration (RQ2). The focus is on AI systems designed to recognize, model, and respond to human emotional intelligence traits (i.e., empathy), thus fostering trust and improving user acceptance in healthcare.

Evaluating Empathy and Communication Quality in Chatbots and Physicians

A study utilizing 195 real patient questions taken from a public forum found that chatbot responses were preferred over physician responses 78.6% of the time. Chatbot's responses were perceived to be more empathetic and were generally longer than physician responses (Ayers et al., 2023). In another study, six oncology physicians rated the responses of chatbots as having higher emotional empathy, cognitive empathy, and an overall empathy score better than that of the physicians (Chen et al., 2024). A cross-sectional survey of patients aged 65 years and older found that chatbots responded more empathetically as compared to physicians (Chen et al., 2025).

Studies published in JMIR on healthcare chatbots concluded that chatbots with personas help build trust and empathy. In one chatbot study, four different personas were considered which were Institution (focused on policies), an Expert (provided recommendations), a Peer (friendly and relatable), and a Dialogical Self (reflective). Adults aged 40 years or older preferred formal and distant personas (institutional or expert), while those aged under 40 reported greater comfort and engagement with peer-oriented dialogical personas. Allowing participants to choose a persona increased both affective bond and intention to use (Nißen et al., 2022).

The healthcare chatbot intervention was evaluated using two independent variables: complexity (defined by how technical or nontechnical the language was) and persona. In a study conducted on a small cohort of students, the measured outcomes included high effectiveness defined as an increase in participant knowledge after using the tool perceived usability, and trust, with health literacy evaluated as a baseline user trait. The results indicated that chatbots employing technical language achieved high effectiveness, while higher health literacy predicted greater trust. To interact with the system, participants typed natural language chats and search queries. These findings suggest that user experience (UX) design must be directly informed by the specific application of the AI system (Biro et al., 2023). Furthermore, users preferred text-based chatbots over anthropomorphic AI interfaces (Thunström et al., 2024). Perceived chatbot usefulness emerged as a key predictor of continued adoption (Al Mamun et al., 2025).

Human-Centered Drivers of AI Adoption in Healthcare

The Unified Theory of Acceptance and Use of Technology (UTAUT) is a framework used to evaluate technology adoption, including AI, based on the help it provides, ease of use, social normality, and economic viability. A quantitative survey revealed that users with higher EI were more willing to adopt the use of AI (Ibrahim et al., 2024). An analysis of 517 Reddit posts revealed that AI- centered communities focused on the benefits of the technology, whereas mental health communities expressed that AI was not as conversationally proficient as humans; meanwhile experienced users articulated both benefits and limits (Lee et al., 2025). Virtual reality-based empathy training may help address these concerns and integrate AI into healthcare systems by enhancing participants' empathetic responses (Lin et al., 2024). Ultimately, integration of AI into the healthcare system requires greater support from healthcare professionals and higher EI amongst users. Table 4 illustrates key performance metrics from empirical studies that demonstrate how EI combined with AI can assist mental healthcare, including psychotherapy and behavioral health support for recognizing emotion-based tasks and to provide therapeutic advice. Individual studies examined did not consistently distinguish between the interface type and the underlying model architecture.

Table 4

Key Performance Metrics on Emotional AI in Mental Health

Population/Setting	AI Application	Key Performance Metrics	Limitations/Challenges Noted	Study (Year)
90 young adults with depressive symptoms	AI-delivered internet-based CBT (ChatGPT-powered "Xiao Zhi") vs. human peer counselor-delivered iCBT	- PHQ-9: AI vs. control d = -0.86 (week 2), d = -0.65 (week 4) - BSI-CV (suicidal ideation): baseline→week 2 d = 0.48; decline week 2→4 - No significant AI vs. human difference at week 2 (d = -0.12) or week 4 (d = 0.55)	Initial efficacy comparable to human support up to week 2; plateau effect by week 4 due to limited emotional perception and personalization; concerns about AI's emotional understanding	Liu et al., 2026
Simulated patient interactions (raters evaluating written responses)	Comparison of physicians vs. Claude chatbots (V1, V2, V2+CoT)	-Physicians M = 2.01 [1.88, 2.13]; -Claude V1 M = 3.35 [3.23, 3.48]; -Claude V2 M = 3.72 [3.62, 3.81]; -Claude V2+CoT M = 4.11 [3.99, 4.22]	Ratings based solely on text; may not capture real-time, non-verbal empathy; potential rater bias; no validation in live clinical encounters	Chen et al., 2025
Adults in community clinics	AI platform for behavioral treatment (Eleos Health) providing real-time therapist feedback	- Depression reduction: 34% (AI) vs. 20% (*TAU*); d = 0.82 - Anxiety reduction: 29% (AI) vs. 8% (*TAU*) - 67% more sessions attended (AI M = 5.24, *TAU* M = 3.14) - Therapists submitted notes 55 h earlier (AI group)	No difference in treatment satisfaction/perceived helpfulness; long-term (>3 mo) effects not sustained in some studies	Torous et al., 2023
177 university students (1.5-year study)	AI-driven dynamic psychological assessment (LLM/RAG + WeChat mini-program) correcting SAS, SDS, SCL-90 scales	- SAS (anxiety): dynamic model AUC = 0.95 vs. traditional AUC = 0.86 - SDS (depression): dynamic model AUC = 0.93 vs. traditional AUC = 0.82 - HAM-A reduction 15.2% (p = 0.004); HAM-D reduction 40.0% ($p < 0.001$) - Cognitive voting participation 79%; behavioral check-ins 42% - SCL-90: initial R^2 = 0.34, later MSE = 102.74 vs. traditional MSE = 84.17	Diminished specificity for complex SCL-90 over time (MSE increase); challenges handling intricate, long-term symptom patterns; need for multimodal data integration and ethical safeguards	Tong et al., 2025

Population/Setting	AI Application	Key Performance Metrics	Limitations/Challenges Noted	Study (Year)
Emotion- recognition task	Generative AI (ChatGPT-4, Google Bard) interpreting emotions via RMET & LEAS	- ChatGPT-4 RMET scores 26 & 27 (significantly above random; *z*-tests vs. pop mean 26.2: -0.05, 0.22; $p > .8$) - LEAS total score = 97 ($z = 4.20$ vs. French norms; $p < .001$) - High interjudge agreement on LEAS scoring (>0.9)	Limited to visual emotion-recognition tasks; generalizability to complex, mixed emotional states uncertain; reliance on English-language norms vs. French comparison sample	Nandrino et al., 2024
Licensed mental health experts	AI vs. human expert asynchronous written psychological advice	- AI advice rated more favorable for emotional empathy (OR = 1.79, 95% *CI* [1.1, 2.93], $p = .02$) and motivational empathy (OR = 1.84, 95% *CI* [1.12, 3.04], $p = .02$) - Participants could not reliably distinguish AI vs. expert-authored advice ($p = .27$) - AI advice judged equally or more empathetic and sound; perceived expert authorship bias increased favorability for expert-attributed responses	Biases favoring perceived human authorship may reduce AI's perceived credibility; lack of genuine human relationship and nuances may limit therapeutic effectiveness despite comparable performance	Lai et al., 2025

Note. Reported metrics vary across studies and are not directly comparable. AI systems described as "chatbots" refer to conversational applications.

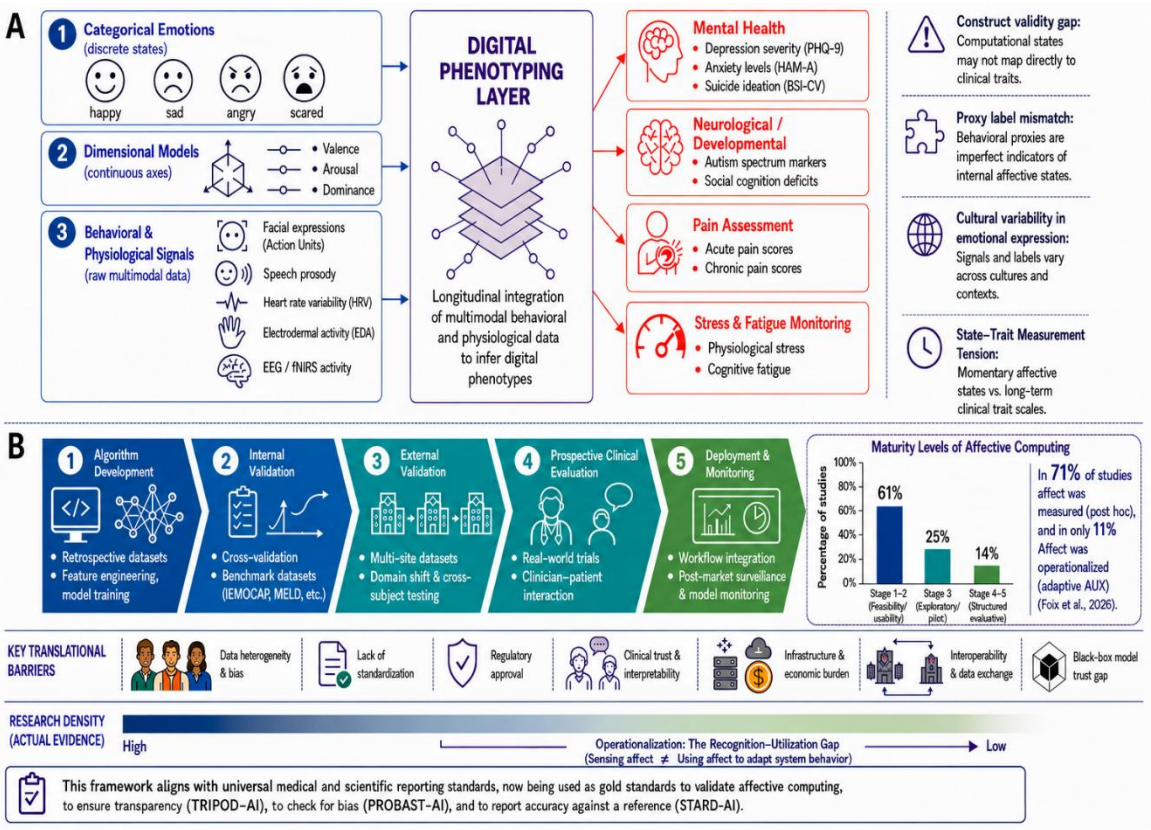
Standardized Metrics and Operationalization

Table 4 demonstrates the various evaluation methods, such as clinical symptom scales (PHQ-9, HAM-D), engagement metrics, and empathy ratings. However, it lacks a standardized metric that creates a roadblock for cross-study examinations and synthesizing evidence (Hicks et al., 2022). There are recent advances that offer empirically validated frameworks that are specifically designed for precision healthcare (Abbasian et al., 2024). The standard proposed for emotionally intelligent AI healthcare agents is based on four dimensions: Emotional Support, Health Literacy, Fairness and Personalization (Abbasian et al., 2024). The Chatbot Compassion Quotient (CCQ) is a method based on healthcare frameworks that quantifies compassion (Muthukumar, 2025). This method helps overcome the limitations highlighted in table 3 and offers a method that can be used in clinical settings and can be useful in the healthcare industry to provide emotionally intelligent processes.

systems (Hicks et al., 2022). Figure 3 synthesizes both the conceptual foundations and translational maturity of affective computing, highlighting that the majority of studies remain at early stages of affect recognition, with relatively few achieving structured clinically actionable validation and deployment and the measurement of affect (post-hoc vs operationalized via adaptive Affective User Experience, AUX), specifically for autism spectrum disorders (Foix et al., 2026). Key translational barriers include data heterogeneity and bias, lack of standardization, regulatory approval, clinical trust and interpretability, infrastructure and economic burden, interoperability and data exchange, and black-box model and trust gap. As AI systems become dependent on large-scale and diverse datasets, concerns regarding privacy, security, and interinstitutional collaboration arise thus driving the emergence of Federated Learning to reduce the need for centralized data sharing in collaborative model development

Figure 3

Conceptual and Translational Frameworks for Affective Computing in Precision Healthcare.



Note. A) Mapping between affective computing constructs: categorical emotions, dimensional models, and behavioral and physiological signals; and clinically actionable endpoints across mental health, neurological and developmental conditions, pain assessment, and stress or fatigue monitoring. A central digital phenotyping layer represents the longitudinal integration of multimodal signals into clinically meaningful constructs highlighting key translational challenges. B) Staged translational readiness framework, depicting the progression from algorithm development to deployment. This framework aligns with universal medical and scientific reporting standards to validate affective computing, to ensure transparency (TRIPOD+AI), to check for bias (PROBAST+AI), and to report accuracy against a reference (STARD-AI).

Federated Learning (FL) and Privacy-Preserving Collaboration

This section addresses RQ3 by critically examining FL as a privacy-preserving infrastructure for cross-institutional AI collaboration. FL enables institutions to collaboratively train AI models by exchanging aggregated model parameters rather than raw patient records (Pati et al., 2024; Zhang et al., 2024), but it is an enabler, not a privacy solution. Its protections are implementation-dependent: FL does not inherently ensure fairness, robustness, or absence of data leakage, nor does it constitute a regulatory compliance mechanism. This section evaluates its principal limitations, including adversarial vulnerabilities, non-IID data heterogeneity, client drift, bias amplification, interoperability constraints, and governance gaps.

Adversarial Vulnerabilities: Gradient Leakage, Model Inversion, and Membership Inference

A critical misconception is that sharing model updates eliminates privacy risk. Gradient leakage, model inversion, and membership inference are not fringe scenarios but systemic risks: adversarial actors can reconstruct training data from shared gradients, potentially re-exposing the medical images and EHR entries the architecture was meant to protect (Kalodanis et al., 2025; Pati et al., 2024). Privacy in a federated system is not a product of decentralization itself, but of the safeguards layered on top of it. Two primary mitigations approaches exist: differential privacy (DP) injects calibrated noise into model updates, while secure multi-party computation (MPC) enables collaborative computation without exposing individual inputs. However, both approaches impose computational overhead that create an inherent direct trade-off between privacy preservation and model performance. In clinical contexts, aggressive noise injection can disproportionately degrade performance on rare conditions where training signals are already weak, a paradox in which privacy preservation actively undermines diagnostic equity (Kalodanis et al., 2025; Pati et al., 2024).

Performance Trade-offs: Non-IID Data, Client Drift, and Bias Amplification

Non-Independently and Identically Distributed (Non-IID) data causes client drift, and the consequences are not evenly distributed across participating institutions. Large hospitals dominate gradient updates simply by virtue of their data volume and annotation quality, while rural and low-resource nodes contribute to updates that carry far less weight in the aggregation process. FedAvg, the canonical algorithm, is vulnerable to this heterogeneity, meaning the resulting global model reflects the priorities of data-rich centers more than the populations served by smaller clinics. This is not a neutral technical outcome; FL may unintentionally amplify institutional inequalities by encoding systemic bias directly into model weights, depriving the very communities that cross-institutional collaboration was meant to serve. FedProx mitigates it via proximal regularization, constraining local update divergence. Hierarchical FL organizes nodes into regional clusters before global aggregation, improving communication efficiency and institutional accommodation (Upreti et al., 2024). These are meaningful corrections, but they require deliberate design choices that do not come as defaults.

FL vs. Centralized Models: Competing Paradigms and Settings

Centralized models pool data for training but require transfer and create single points of failure,

whereas FL preserves data sovereignty at the cost of coordination complexity and model drift. Settings include horizontal FL (shared feature space, e.g., EHR), vertical FL (complementary feature types, e.g., imaging plus genomics), and cross-device FL across IoMT networks. Evidence demonstrates AUC above 0.9 across 15 global sites (Acosta et al., 2022), comparable performance in embryo selection (Lee et al., 2024) and endometrial cancer diagnostics (Yeom et al., 2024), with gaps persisting in data-sparse environments. .The Federated Tumor Segmentation (FeTS) initiative demonstrated that glioma segmentation models trained across geographically distributed institutions using multi-parametric MRI can match or surpass centralized performance benchmarks, though gains remained sensitive to site heterogeneity, data quality disparities, and the choice of aggregation algorithm (Linardos et al., 2025).

Multimodal FL: Aggregation Complexity, Interpretability, and Bias

Extending FL to multimodal data (genomics, imaging, EHR, biosensors) compounds complexity: each modality requires dedicated encoders reconciling heterogeneous representation spaces. When one institution contributes imaging data, another contributes genomic profiles, and a third contributes only structured EHR records, the federation faces representation misalignment: the model must reconcile feature spaces that were never designed to correspond, without ever having access to the raw data needed to learn that correspondence directly. Asynchronous modality availability makes this worse, as nodes that lack a given modality at training time either drop out of relevant aggregation rounds or introduce missingness patterns that distort the global model in ways that are difficult to detect or correct. XAI methods designed for centralized single-modality models do not transfer directly to federated multimodal settings. Modality-specific bias propagation follows naturally from this architecture: modalities contributed by well-resourced institutions disproportionately shape shared representations, systematically disadvantageous underrepresented populations whose data profiles do not match the dominant modality mix. FL-based ocular disease detection without sharing retinal scans illustrates clinical utility (Gulati et al., 2025), yet cross-modal interpretability and aggregation consistency remain open and underaddressed challenges.

Governance, Liability, Data Sovereignty, and Regulatory Alignment

FL liability is diffused across model coordinators, node contributors (hospitals, clinics), and developers, creating a structural governance gap that challenges medical liability frameworks built around identifiable custodians and traceable decision chains. HIPAA, GDPR, and the EU AI Act were designed around centralized data architectures and face significant implementation challenges in FL environments. The GDPR's right to erasure is difficult to operationalize once patient data has contributed to model parameters distributed across multiple nodes. Similarly, auditability is hindered by limited provenance tracking while cross-border parameter exchange creates unresolved jurisdictional ambiguity. These issues highlight broader challenges between data sovereignty, the ability of institutions to maintain control over how and which data leaves local infrastructure, and model sovereignty, which concerns ownership, oversight, and auditability of the shared global model weights shaped by distributed contributions. Although blockchain integration supports tamper-proof provenance tracking and auditability, standardized global governance protocols for FL remain underdeveloped (Ning et al., 2024; Zhang et al., 2024).

FL Applicability, Interoperability, and Systemic Limitations

FL demonstrates optimal performance in contexts where data cannot be legally centralized, where institutional diversity adds collaborative learning value, and where participating nodes maintain sufficiently robust infrastructure to support federated coordination. FL is most effective in high-resource, well-coordinated institutional networks where participating organizations have sufficient technical infrastructure, data quality, and governance capacity. Performance is diminished in fragmented environments with uneven participation nodes, variable data quality with highly non-IID data, limited computational

resources, or aggressive privacy-preserving noise injection that reduces model utility below clinically acceptable thresholds. Furthermore, interoperability between FL platforms (PySyft, FATE, TensorFlow Federated) and institutional health IT systems remains a persistent structural barrier, compounded by unclear accountability mechanisms when models malfunction across distributed networks (Xu et al., 2025).

FL should therefore be understood as a context-dependent tool with limitations, rather than a universally applicable solution. Responsible deployment requires recognizing these operational boundaries, including severe data heterogeneity that impedes coherent aggregation, insufficient participation to achieve adequate statistical power, privacy constraints that substantially reduce utility, and governance structures too fragmented to assign accountability. Technical safeguards are insufficient to ensure safe and equitable deployment of AI models, a governance structure that addresses accountability, oversight, fairness and long-term sustainability is critical.

Governance, Ethics, and Leadership in AI-Enabled Healthcare

Governance shouldn't be viewed merely as a compliance layer applied after AI systems are developed; rather as an institutional capability that shapes whether AI can be deployed safely, equitably, and sustainably within precision healthcare. Because AI-enabled diagnosis, risk stratification, treatment selection, and patient communication can directly influence clinical outcomes, governance must be embedded across the full lifecycle of model design, validation, deployment, monitoring, and revision.

Multilevel Governance Frameworks (EU, USA, Global)

AI governance in precision healthcare operates across interconnected supranational, national, and institutional systems, and its scope extends well beyond regulatory compliance. Accountability, institutional readiness, clinical stewardship, and ethical leadership are all implicated particularly because precision healthcare depends on sensitive patient data, predictive analytics, and clinical decision-support systems whose outputs bear directly on diagnosis, treatment selection, and patient outcomes.

The EU AI Act represents a decisive shift toward enforceable, risk-based governance (European Commission, 2024). Classifying clinical decision-support and diagnostic systems as high-risk, the Act imposes requirements for transparency, human oversight, documentation, and post-market monitoring (European Commission, 2024; Kalodanis et al., 2025). This shifts governance from voluntary ethical guidance toward legally enforceable accountability, which is its principal strength, although its complexity may slow innovation or prove difficult to operationalize across diverse healthcare settings.

The United States operates through a sector-specific and adaptive model. Acting jointly, the USA Food and Drug Administration (FDA), Health Canada (HC), and the UK Medicines and Healthcare products Regulatory Agency (MHRA) developed Good Machine Learning Practice (GMLP) principles addressing dataset representativeness, lifecycle monitoring, risk management, and real-world performance (FDA et al., 2021). This flexibility supports innovation, yet oversight remains distributed across existing regulatory structures rather than consolidated under a single AI statute, which risks fragmented accountability (Bajwa et al., 2021; Matheny et al., 2023).

At the global level, the World Health Organization and the OECD offer normative frameworks centered on transparency, fairness, accountability, and human-centered values (OECD, 2019; World Health Organization, 2021). Their value lies in ethical alignment, but limited enforceability constrains real-world impact, especially where regulatory capacity and digital infrastructure remain weak (Kergroach & Héritier, 2024). No single model fully reconciles the competing demands of accountability, innovation, and equity (Matheny et al., 2023). As shown in Table 5, the effectiveness of governance frameworks differ substantially in scope, enforceability, analytical strength, and implementation limitations.

Table 5

Comparative Multilevel Governance Frameworks in Precision Healthcare

Governance level	Framework	Scope	Approach	Enforceability	Analytical strength	Structural limitation	Sources
Supranational	EU AI Act	Cross-sector; includes healthcare as high-risk	Risk-based, precautionary	High; legally binding	Strong accountability and regulatory clarity	May constrain innovation; complex implementation	European Commission, 2024; Kalodanis et al., 2025
National	FDA GMLP (U.S. A.)	Healthcare-specific medical device and clinical AI	Adaptive, lifecycle-based	Moderate-high through device oversight	Supports innovation and real-world monitoring	Fragmented accountability across agencies	Bajwa et al., 2021; FDA et al., 2021; Matheny et al., 2023
Global	WHO AI ethics guidance	Healthcare-focused	Normative, principles- based	Low; non-binding	Strong ethical alignment	Dependent on national implementation	World Health Organization, 2021
Global	OECD AI principles	Cross-sector	Soft law	Low; voluntary adoption	Human-centered responsible AI	Limited healthcare specificity	Kergroach & Héritier, 2024; OECD, 2019
Institutional	Clinical governance	Healthcare-specific organizational level	Operational oversight	Variable; policy and leadership dependent	Direct implementation and stewardship	Uneven resources and capacity	Abbas et al., 2024; Allen, 2024; Cresswell & Sheikh, 2013

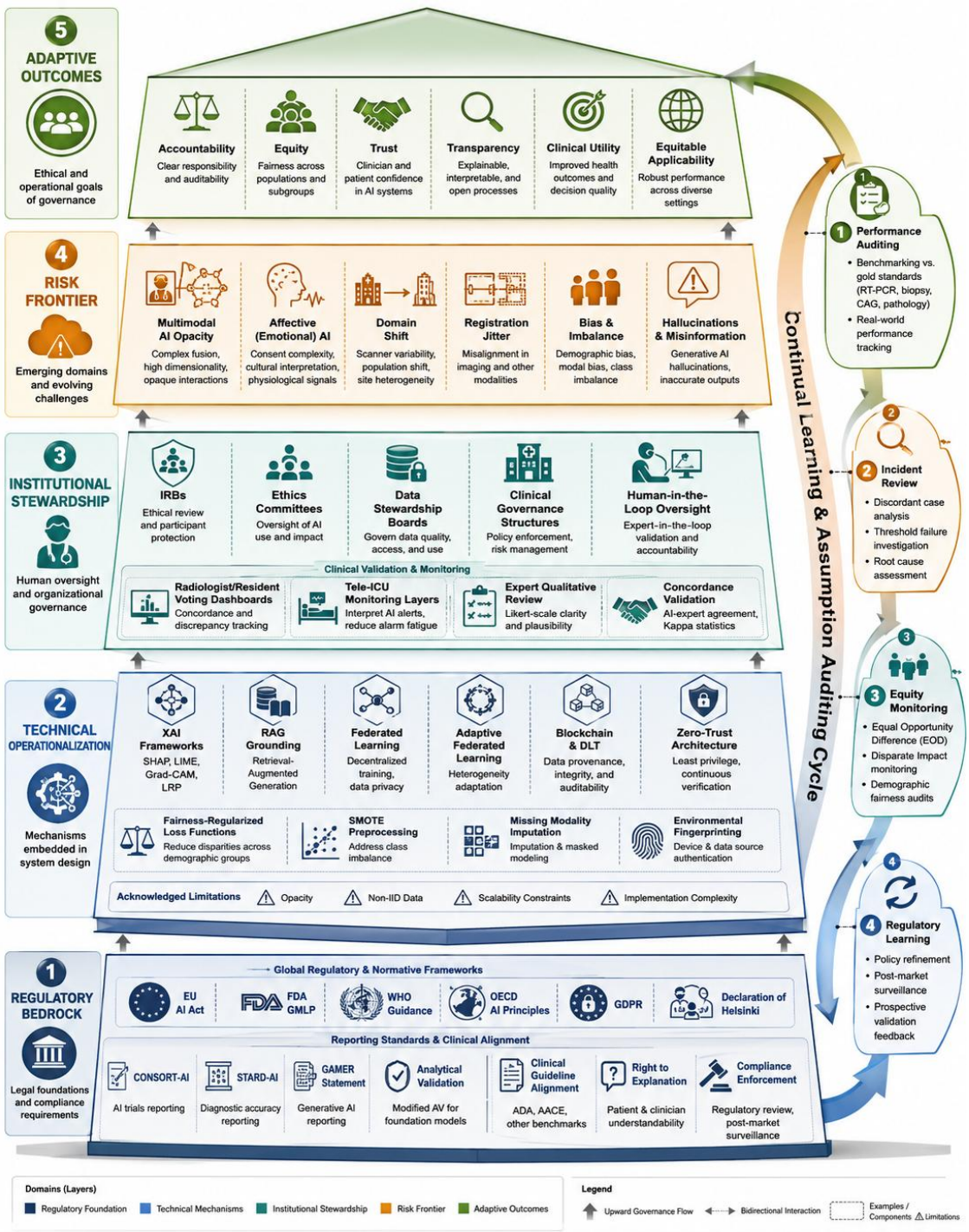
Technical Mechanisms as Governance Tools (XAI, FL, Blockchain)

Effective governance must be embedded within system design rather than imposed from outside it. Explainable AI (XAI) is often positioned as a mechanism for strengthening clinical accountability by making algorithmic outputs interpretable to clinicians (Allen, 2024). In practice, however, most XAI methods operate through post hoc approximation rather than genuine model transparency, a structural tension that technical tools alone cannot resolve. Clinical and institutional responsibility must remain with humans.

FL supports privacy and data minimization by enabling model training across distributed sites without transferring raw patient data (Rieke et al., 2020; Sheller et al., 2020). Yet non-identically distributed data, uneven institutional participation, and absent agreed protocols constrain generalizability and complicate model aggregation (Almogadwy & Alqarafi, 2025). Blockchain improves auditability and consent tracking but faces scalability limits and tension with the GDPR right to erasure (Agbo et al., 2019; Hasselgren et al., 2020). Zero-trust architecture provides continuous authentication and context-sensitive access control, though deployment is infrastructure-intensive and may disrupt workflows where access delays affect urgent care (Al-Hammuri et al., 2024). What these mechanisms share is that they redistribute governance risk rather than eliminate it. Figure 4 presents the proposed Stratified Adaptive Governance (SAG) model, structured across five interdependent layers integrating regulatory, technical, institutional, and clinical oversight mechanisms into a unified governance framework.

Figure 4

Adaptive AI Governance in Precision Healthcare: The Stratified Adaptive Governance (SAG) Model in Five Layers.



Note. Layer 1 establishes the regulatory bedrock. Layer 2 embeds technical operationalization in system design. Layer 3 represents institutional stewardship with human oversight and organizational governance. Layer 4 maps the risk frontier. Layer 5 defines adaptive outcomes. A continual learning and auditing cycle connects all layers.

Clinical Oversight and Institutional Stewardship

Clinical oversight must reside within identifiable governance bodies, Institutional Review Boards, hospital ethics committees, clinical governance committees, data stewardship boards, AI safety committees, and regulatory agencies such as the FDA and European national competent authorities. Their core functions, validating use cases, monitoring safety, auditing for bias, and preserving clinical accountability, become increasingly critical as aging populations place greater demands on healthcare systems (Jones & Dolsten, 2024). Effectiveness, however, depends on adequate resources, technical expertise, and organizational maturity, which is why adaptive governance must be treated as an institutional capability rather than a static policy artefact (Cresswell & Sheikh, 2013; Matheny et al., 2023; World Health Organization, 2021). Table 6 demonstrates that governance has several functions including ensuring accountability, privacy, transparency, security, compliance and equity which require complementary technical, and organizational considerations for their implementation.

Table 6

Governance Functions, Supporting Mechanisms, and Implementation Realities

Function	Mechanism	Role	Outcome	Implementation challenges	Limitations	Source
Accountability	Explainable AI	Interprets model outputs	Clinical trust and decision justification	Workflow integration; clinician training; variable interpretation	*Post hoc* approximation does not provide full model transparency	Abbas et al., 2024; Allen, 2024
Privacy	Federated learning	Decentralized training without raw data sharing	Data minimization and patient data protection	Non-IID data; uneven participation; coordination complexity	Reduced generalizability and performance degradation	Rieke et al., 2020; Sheller et al., 2020; Almogadwy & Alqarafi, 2025
Transparency	Blockchain	Immutable audit trails	Traceability and consent management	High computational cost; legacy system integration	Scalability constraints; GDPR right-to-erasure tension	Agbo et al., 2019; Hasselgren et al., 2020
Security	Zero-trust architecture	Continuous verification and access control	Risk mitigation and breach prevention	Infrastructure demands; interoperability; workflow disruption	Latency; difficult in low-resource settings; access delays	Al-Hammuri et al., 2024
Compliance	Risk-tier frameworks	Monitoring and reporting	Regulatory alignment	Regulatory fragmentation; evolving standards	May lag behind technological development	Kalodanis et al., 2025
Equity	Interoperability systems	Data exchange across systems and regions	Broader access and reduced disparities	Infrastructure gaps; lack of standardized formats	Digital divide and uneven global adoption	Kergroach & Héritier, 2024

Note. Implementation challenges refer to practical barriers to use, while limitations refer to structural or inherent constraints of the mechanism.

Adoption, Trust, and Decision-Science Perspectives

Regulatory and technical readiness alone do not determine whether AI systems are adopted in clinical settings. Whether governance mechanisms translate into actual practice depends on clinician perceptions of usefulness, workflow compatibility, and trustworthiness. Output uncertainty, variable performance, and limited explainability can erode confidence and reduce uptake (Cresswell & Sheikh, 2013; Matheny et al., 2023). Governance that delivers credible accountability, transparent validation, and meaningful privacy safeguards builds the appropriate trust on which adoption depends.

Emerging Challenges: Multimodal AI and Affective Computing

Multimodal AI, which draws on imaging, genomics, electronic health records, and wearable sensor data simultaneously, may improve diagnostic precision and personalization, but it also amplifies opacity, consent complexity, and cross-modal bias. Affective or emotional AI raises further concerns: interpreting facial expressions, vocal patterns, or behavioral cues in clinical contexts is culturally variable and may affect triage or risk stratification in ways that existing frameworks have not fully addressed (Bajwa et al., 2021; Matheny et al., 2023; World Health Organization, 2021). The gap between regulatory intent and operational reality is likely to widen as these technologies diffuse into clinical environments.

Global Inequities and Implementation Gaps

Global inequities compound these governance challenges. In sub-Saharan Africa, fragmented health information systems and limited digital infrastructure constrain even basic implementation. Across South Asia, digital health expansion has repeatedly outrun regulatory development. In Latin America, persistent interoperability gaps and uneven institutional capacity create divergent trajectories across urban and rural systems. Governance frameworks developed in high-income settings cannot be exported without contextual adaptation (Kergroach & Héritier, 2024; World Health Organization, 2021). There is also a structural risk of data colonialism, whereby low-resource regions may contribute training data while exercising limited influence over model design, governance standards, or the distribution of resulting benefits.

Sustainable governance in precision healthcare ultimately depends on the alignment of law, ethics, technology, and leadership. The EU model offers stronger enforceability; the USA model offers adaptability. Neither fully resolves opacity, bias, or uneven implementation. Governance must therefore function as a dynamic capability and adaptive organizational capacity integrating regulatory alignment, technical understanding, clinical oversight, and ethical decision-making under conditions of incomplete interpretability. The central question is not whether AI can support precision healthcare, but whether healthcare systems can govern it responsibly, equitably, and adaptively across diverse real-world contexts. That capacity is built not through regulation alone but through the sustained alignment of legal frameworks, technical architectures, and institutional accountability structures.

Discussion/Implications

Cross-Pillar Synthesis

This review synthesizes an interdependent five-pillar framework forAI-enabled precision healthcare that supports a transition from isolated unimodal algorithmic development toward an integrated, holistic, trustworthy and clinically actionable socio-technical model of care. The evidence provided in support of RQ1 suggests that multimodal and explainable AI can improve diagnostic and prognostic performance over many unimodal models (Acosta et al., 2022; Parvin et al., 2025; R. Zhang et al., 2026), but should be viewed as extensions, and not replacements, of high-performing unimodal architectures that leverage model-specific strengths in comprehensive decision-aware pipelines to enhance clinical relevance (Figure 2;

Table 1; Table 2; Suppl. Table 1). Within a systems-level paradigm, the foundational data used in MAI systems can be grouped into three broad categories: 1) structural and anatomical information extracted from imaging, 2) physiological and longitudinal dynamic data from sensors and time-series monitoring, and 3) contextual and semantic data extracted from EHR, clinical unstructured text and large language models. Most high performing systems incorporate at least two of these categories. Emerging architectures incorporate all three categories to enable deep phenotyping, longitudinal monitoring, and personalized treatment planning (Table 1, Suppl. Table 1).

Despite architectural complexity, multimodal gains can plateau when data is imbalanced, quality is poor, integration introduces noise or cross-modal correlations are spurious. Consequently, marginal performance gains don't necessarily translate into improved clinical utility, patient outcomes or workflow efficiency (Hicks et al., 2022; Shamszare & Choudhury, 2023). Clinical usefulness requires going beyond sophisticated model performance and addressing calibration, external validation, workflow integration, interpretability, and cost-efficiency challenges (Acosta et al., 2022; Dantone et al., 2026).

The integration of MAI systems into the analysis requires balancing critical technical trade-offs, including the fusion strategy, and AI algorithm. An early fusion strategy maximizes feature interaction but is sensitive to missingness and temporal misalignment. An intermediate fusion strategy improves cross-modal dependencies (with the use of transformers and cross-attention mechanisms) but increases computational burden and opacity (R. Zhang et al., 2026), whereas a late fusion strategy treats missingness and heterogeneity at the expense of deep inter-modal reasoning. Different AI algorithms across CNNs, RNNs, LSTMs, GANs, transformers, autoencoders, and hybrid architectures provide richer and more complex analyses at the expense of significant infrastructure requirements and energy costs, interpretability, and scalability in real-world clinical settings. High-quality data availability is one of the major bottlenecks in improving model performance, and hence many algorithms are trained in small, region-specific (mostly high-income regions in the USA, EU, or China), inadequately validated, highly curated or synthetic datasets that are not representative of the heterogeneous real-world clinical landscape, which in turn increases bias (unless fairness is integrated) and limits generalizability.

The literature suggests that XAI is evolving into a core evidentiary mechanism, critical for clinical reasoning, pathophysiological plausibility, and regulatory compliance, for trustworthy clinical deployment and accountability (Q. Abbas et al., 2025; Allen, 2024; Gerdes, 2024; K. Zhang et al., 2026). In this regard, there is a trade-off between high performance of complex models (such as transformer-based systems and hybrid architectures), and level of interpretability through methods such as SHAP, LIME, Grad-CAM, Score-CAM, saliency maps, and layer-wise relevance propagation. As observed in our table analysis, true mechanistic explainability is limited as most methods provide post-hoc approximations (Table 1; Table 2; Table 3; Suppl. Table 1). An ensemble of complementary explainability approaches is needed to support three stakeholder levels: locally at the patient level for clinical decision-making, and globally for researchers to improve the mechanism, and for regulators to ensure auditability, fairness, reproducibility and compliance.

Our literature search addressing RQ2 suggests that affective computing and human-centered AI are emerging and comparatively less mature fields of research where conversational AI systems have primarily been evaluated based on their anthropomorphic interface design, emotion-recognition performance, and perceived empathetic, and supportive responses. However, there are other metrics that are not often evaluated or achieved, including therapeutic alliance, clinical usability, health-literacy alignment, patient adherence, workflow compatibility, psychological safety, and long-term patient outcomes (Biro et al., 2023). The evaluation of highly sensitive physiological, vocal, facial, and behavioral signals introduces an ethical and governance challenge due to the re-identification risk based on stable psychological traits. HIPAA and GDPR were not originally designed to regulate emotional inference and behavioral monitoring systems, thus leaving open

a regulatory gap regarding patient consent and model accountability (Bouderhem, 2024).

To address RQ3, existing research suggests that FL provides a privacy-preserving decentralized infrastructure enabling interoperability across institutions but has limitations in guaranteeing privacy and security including gradient leakage, membership inference, model inversion attacks, client drift, and non-IID data heterogeneity. Adaptive FL systems can improve local personalization but increase the privacy leakage risk. Mitigation strategies for these vulnerabilities include differential privacy, secure multi-party computation, blockchain integration, and zero-trust architectures (K. Li et al., 2025; Ning et al., 2024; Wang et al., 2026). Fairness-aware architectures with adaptive governance models are needed to mitigate institutional inequalities and data disparities but they introduce several challenges: computational overhead, communication burden, and governance complexity. International governance models differ significantly where EU models provide strict enforceable oversight and accountability, USA models emphasize sector-specific adaptive innovation, and global ethical frameworks provide normative guidance but lack enforceability. Governance extends beyond regulation to cover institutional stewardship, clinical oversight, and socio-technical alignment. AI adoption requires models that are useful, interpretable, workflow-compatible, institutionally supported, compliant and adequately governed (Cresswell & Sheikh, 2013; Venkatesh et al., 2003; Venkatesh et al., 2012).

This narrative review has quantitative comparability limitations as it covers a broad scope that includes the synthesis of heterogeneous studies with diverse methodologies, data modalities, validation procedures, outcome metrics, and reporting standards spanning across multiple clinical domains.

Implications

The practical implication is that healthcare AI should be evaluated not only by technical performance but also by whether clinicians perceive it as useful, interpretable, workflow-compatible, institutionally supported, and appropriately governed (Dingel et al., 2024; Shamszare & Choudhury, 2023). For developers, this means prioritizing external validation, explanation usability, subgroup performance, uncertainty reporting, and lifecycle monitoring. For healthcare institutions, it means building governance structures that can review safety, bias, privacy, workflow fit, and patient impact before and after deployment. For policymakers, it means designing frameworks that address both technical risk and implementation capacity across diverse health systems (Aboy et al., 2024).

Limitations

This narrative review has quantitative comparability limitations as it covers a broad scope that includes the synthesis of heterogeneous studies with diverse methodologies, data modalities, validation procedures, performance metrics, and reporting standards spanning across multiple clinical domains. Because it is a narrative review with structured evidence mapping rather than a systematic review or meta-analysis, the findings should be interpreted as integrative synthesis rather than pooled effect estimation. Many studies relied on retrospective datasets, single-institution cohorts, simulated interactions, short-term outcomes, or highly curated data, which may overstate real-world readiness. Additional limitations include potential publication bias arising from publication date restrictions specified by the journal and overrepresentation of studies conducted in high-income settings. Moreover, because affective computing research remains comparatively immature, the evaluation on this section was challenging.

Conclusion

Artificial intelligence is increasingly enabling precision healthcare through the integration of multimodal data, advanced predictive analytics, and clinical decision-support systems. This review proposes a five-pillar framework, consisting of multimodal and explainable AI, affective computing, privacy-preserving collaboration, and adaptive governance, enabling an analytical shift from isolated unimodal data analysis to holistic, human-centered, trustworthy,

and continuously learning actionable clinical ecosystems.

Our findings demonstrate that predictive accuracy alone is insufficient for meaningful clinical translation. Model performance is highly affected by the presence of data heterogeneity, modality imbalance, missingness, institutional and demographic bias, weak external validation, computational burden, and poor workflow integration. While explainable AI has improved transparency, accountability, clinician trust, and regulatory alignment; addressing clinical plausibility in model output remains a challenge. Hence, ML and DL AI systems must be used as clinical decision support tools integrated within human-in-the-loop frameworks for improved clinical translation. Although methodologically immature and insufficiently validated, affective computing and human-centered AI demonstrate emerging promise in introducing emotionally intelligent interfaces in conversational systems for mental health support, chronic disease management, and patient-centered communication. Future work in this field should go beyond technical performance, and prioritize achieving therapeutic alliance, emotional safety, accessibility, and health-literacy alignment.

Future research should prioritize mechanism-aware models, cloud-based multimodal health information systems, adaptive lifecycle monitoring, fairness-aware and bias-sensitive model learning strategies, standardized explainability and human-centeredness metrics, energy and resource-efficient architectures for more generalized sustainable deployment across resource-rich and -limited settings, and robust post-market surveillance frameworks. Post-market surveillance involves performing equity audits, continuous model drift assessment, multicenter external validation, lifecycle performance assessment, privacy-preserving systems with clear liability rules, and longitudinal prospective clinical trial outcome monitoring. Other frontiers for research include causal digital twins development, genome editing, geroscience, wearable ecosystems, and smart-home monitoring. and. There is a need for regulatory bodies to remain agile in providing standardization efforts , interoperability and reporting protocols for improved reproducibility and comparability. Evidence-based and human-centered AI precision healthcare systems will become more effective through the implementation of a sophisticated socio-technical framework ensuring a holistic patient assessment, clinical interpretability and plausibility, compassionate care aligned with human values, physician accountability, patient trust, and ethical and equitable governance.

Acknowledgments

The authors gratefully acknowledge Adj. Professor Shelley Spessard (College of Education, Westcliff University), Adj. Professor Julio Zelaya (College of Education, Westcliff University), and writing specialist Stephanie Mojica (Writing Center, Westcliff University) for providing feedback and/or editorial guidance for the review article. NotebookLM and Perplexity were used to support literature exploration and research synthesis, Grammarly was used for grammar review, and ChatGPT was used to assist with figure generation under human oversight and iterative revision.

References

Abbas, Q., Jeong, W., & Lee, S. W. (2025). Explainable AI in clinical decision support systems: A meta-analysis of methods, applications, and usability challenges. *Healthcare, 13*(17), 2154. https://doi.org/10.3390/healthcare13172154

Abbas, S. R., Seol, H., Abbas, Z., & Lee, S. W. (2025). Exploring the role of artificial intelligence in smart healthcare: A capability and function-oriented review. *Healthcare, 13*(14), 1642. https://doi.org/10.3390/healthcare13141642

Abbas, T., Fatima, A., Shahzad, T., Alharbi, M., Khan, M. A., & Ahmed, A. (2024). Multidisciplinary cancer disease classification using adaptive federated learning in healthcare industry 5.0. *Scientific Reports, 14*(1), 18643. https://doi.org/10.1038/s41598-024-68919-1

Abbasian, M., Khatibi, E., Azimi, I., Oniani, D., Abad, Z. S. H., Thieme, A., Sriram, R., Yang, Z., Wang, Y., Lin, B., Gevaert, O., Li, L.-J., Jain, R., & Rahmani, A. M. (2024).

Foundation metrics for evaluating effectiveness of healthcare conversations powered by generative AI. *npj Digital Medicine,* *7*(1), 82. https://doi.org/10.1038/s41746-024-01074-z

Aboy, M., Minssen, T., & Vayena, E. (2024). Navigating the EU AI Act: Implications for regulated digital medical products. *npj Digital Medicine,* *7*(1), 237. https://doi.org/10.1038/s41746-024-01232-3

Acosta, J. N., Falcone, G. J., Rajpurkar, P., & Topol, E. J. (2022). Multimodal biomedical AI. *Nature Medicine,* *28*(9), 1773–1784. https://doi.org/10.1038/s41591-022-01981-2

Agbo, C. C., Mahmoud, Q. H., & Eklund, J. M. (2019). Blockchain technology in healthcare: A systematic review. *Healthcare,* *7*(2), Article 56. https://doi.org/10.3390/healthcare7020056

Al Mamun, M. R. A., Ntsweng, O., David, A., Baah-Peprah, P., & Prybutok, V. (2025). What drives user intention to continue using conversational AI? How functional and emotional values influence continuance intention. *AIS Transactions on Human-Computer Interaction,* *17*(1), 1–34. https://doi.org/10.17705/1thci.00216

Al-Hammuri, K., Gebali, F., & Kanan, A. (2024). ZTCloudGuard: Zero trust context-aware access management framework to avoid medical errors in the era of generative AI and cloud-based health information ecosystems. *AI,* *5*(3), 1111–1131. https://doi.org/10.3390/ai5030055

Albahri, A. S., Joudar, S. S., Hamid, R. A., Zahid, I. A., Alqaysi, M. E., Albahri, O. S., Alamoodi, A. H., Kou, G., & Sharaf, I. M. (2024). Explainable artificial intelligence multimodal of autism triage levels using fuzzy approach-based multi-criteria decision-making and LIME. *International Journal of Fuzzy Systems,* *26*(1), 274–303. https://doi.org/10.1007/s40815-023-01597-9

Allen, B. (2024). The promise of explainable AI in digital health for precision medicine: A systematic review. *Journal of Personalized Medicine,* *14*(3), Article 277. https://doi.org/10.3390/jpm14030277

Almogadwy, B., & Alqarafi, A. (2025). Fused federated learning framework for secure and decentralized patient monitoring in healthcare 5.0 using IoMT. *Scientific Reports,* *15*, Article 24263. https://doi.org/10.1038/s41598-025-06574-w

Andrade, C., Teixeira, L. F., Vasconcelos, M. J. M., & Rosado, L. (2021). Data augmentation using adversarial image-to-image translation for the segmentation of mobile-acquired dermatological images. *Journal of Imaging,* *7*(1), 2. https://doi.org/10.3390/jimaging7010002

Awais, M., Raza, M., Singh, N., Bashir, K., Manzoor, U., Islam, S. U., & Rodrigues, J. J. P. C. (2021). LSTM-based emotion detection using physiological signals: IoT framework for healthcare and distance learning in COVID-19. *IEEE Internet of Things Journal,* *8*(23), 16863–16871. https://doi.org/10.1109/JIOT.2020.3044031

Ayers, J. W., Poliak, A., Dredze, M., Leas, E. C., Zhu, Z., Kelley, J. B., Faix, D. J., Goodman, A. M., Longhurst, C. A., Hogarth, M., & Smith, D. M. (2023). Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. *JAMA Internal Medicine,* *183*(6), 589–596. https://doi.org/10.1001/jamainternmed.2023.1838

Bajwa, J., Munir, U., Nori, A., & Williams, B. (2021). Artificial intelligence in healthcare: Transforming the practice of medicine. *Future Healthcare Journal,* *8*(2), e188–e194. https://doi.org/10.7861/fhj.2021-0095

Banerjee, I., Ling, Y., Chen, M. C., Hasan, S. A., Langlotz, C. P., Moradzadeh, N., Chapman, B., Amrhein, T., Mong, D., Rubin, D. L., Farri, O., & Lungren, M. P. (2019). Comparative effectiveness of convolutional neural network (CNN) and recurrent neural network (RNN) architectures for radiology text report classification. *Artificial Intelligence in Medicine, 97*, 79–88. https://doi.org/10.1016/j.artmed.2018.11.004

Bhushan, A., & Misra, P. (2025). Unlocking the potential: Multimodal AI in biotechnology and digital medicine—economic impact and ethical challenges. *npj Digital Medicine, 8*(1), 619. https://doi.org/10.1038/s41746-025-01992-6

Biro, J., Linder, C., & Neyens, D. (2023). The effects of a healthcare chatbot's complexity and persona on user trust, perceived usability, and effectiveness: Mixed methods study. *JMIR Human Factors, 10*, e41017. https://doi.org/10.2196/41017

Biswas, S., Mostafiz, R., Uddin, M. S., & Uddin, M. S. (2025). FLPneXAINet: Federated deep learning and explainable AI for improved pneumonia prediction utilizing GAN-augmented chest X-ray data. *PLoS ONE, 20*(7), e0324957. https://doi.org/10.1371/journal.pone.0324957

Bouderhem, R. (2024). Shaping the future of AI in healthcare through ethics and governance. *Humanities and Social Sciences Communications, 11*(1), 416. https://doi.org/10.1057/s41599-024-02894-w

Chen, D., Parsa, R., Hope, A., Hannon, B., Mak, E., Eng, L., Liu, F.-F., Fallah-Rad, N., Heesters, A. M., & Raman, S. (2024). Physician and artificial intelligence chatbot responses to cancer questions from social media. *JAMA Oncology, 10*(7), 956–960. https://doi.org/10.1001/jamaoncol.2024.0836

Chen, D., Chauhan, K., Parsa, R., Liu, Z. A., Liu, F.-F., Mak, E., Eng, L., Hannon, B. L., Croke, J., Hope, A., Fallah-Rad, N., Wong, P., & Raman, S. (2025). Patient perceptions of empathy in physician and artificial intelligence chatbot responses to patient questions about cancer. *npj Digital Medicine, 8*(1), 275. https://doi.org/10.1038/s41746-025-01671-6

Chen, J., Chan, N. Y., Li, C.-T., Chan, J. W. Y., Liu, Y., Li, S. X., Chau, S. W. H., Leung, K. S., Heng, P.-A., Lee, T. M. C., Li, T. M. H., & Wing, Y.-K. (2024a). Multimodal digital assessment of depression with actigraphy and app in Hong Kong Chinese. *Translational Psychiatry, 14*(1), 150. https://doi.org/10.1038/s41398-024-02873-4

Chen, Y., Xin, Z., Yang, D., Song, X., Zhong, J., Weng, J., Zhang, Y., Liu, D., Wang, M., Kang, L., & Yuan, J. (2025). Application of artificial intelligence software to identify emotions of lung cancer patients in preoperative health education: A cross-sectional study. *Journal of Nursing Scholarship, 57*(4), 546–556. https://doi.org/10.1111/jnu.70001

Chettri, S. K., Deka, R. K., & Saikia, M. J. (2025). Bridging the gap in the adoption of trustworthy AI in Indian healthcare: Challenges and opportunities. *AI, 6*(1), 10. https://doi.org/10.3390/ai6010010

Cocos, A., Fiks, A. G., & Masino, A. J. (2017). Deep learning for pharmacovigilance: Recurrent neural network architectures for labeling adverse drug reactions in Twitter posts. *Journal of the American Medical Informatics Association, 24*(4), 813–821. https://doi.org/10.1093/jamia/ocw180

Cresswell, K., & Sheikh, A. (2013). Organizational issues in the implementation and adoption of health information technology innovations: An interpretative review. *International Journal of Medical Informatics, 82*(5), e73–e86. https://doi.org/10.1016/j.ijmedinf.2012.10.007

Cross, J. L., Choma, M. A., & Onofrey, J. A. (2024). Bias in medical AI: Implications for clinical decision-making. *PLOS Digital Health, 3*(11), e0000651. https://doi.org/10.1371/journal.pdig.0000651

Cui, C., Ding, X., Wang, D., Chen, L., Xiao, F., Xu, T., Zheng, M., Luo, X., Jiang, H., & Chen, K. (2021). Drug repurposing against breast cancer by integrating drug-exposure expression profiles and drug–drug links based on graph neural network. *Bioinformatics, 37*(18), 2930–2937. https://doi.org/10.1093/bioinformatics/btab191

Dantone, M., Lacsamana, M., Zeng, K. G., Kenny, P. A., Geras, K. J., & Witowski, J. (2026). Analytical validation of multimodal AI test predicting breast cancer recurrence risk (Ataraxis Breast RISK). *Diagnostics, 16*(7), 1023. https://doi.org/10.3390/diagnostics16071023

Deacon, S., Cahyani, I., Holmes, N., Fox, G., Munro, R., Wibowo, S., Murray, T., Mason, H., Housley, M., Martin, D., Sharif, A., Patel, A., Goldspring, R., Brandner, S., Sahm, F., Smith, S., Paine, S., & Loose, M. (2025). ROBIN: A unified nanopore-based assay integrating intraoperative methylome classification and next-day comprehensive profiling for ultra-rapid tumor diagnosis. *Neuro-Oncology, 27*(8), 2035–2046. https://doi.org/10.1093/neuonc/noaf103

Demir, S., Selvitopi, H., & Selvitopi, Z. (2025). An early and accurate diagnosis and detection of coronary heart disease using deep learning and machine learning algorithms. *Journal of Big Data, 12*(1), 228. https://doi.org/10.1186/s40537-025-01283-7

De Santi, L. A., Pasini, E., Santarelli, M. F., Genovesi, D., & Positano, V. (2023). An explainable convolutional neural network for the early diagnosis of Alzheimer's disease from 18F-FDG PET. *Journal of Digital Imaging, 36*(1), 189–203. https://doi.org/10.1007/s10278-022-00719-3

Dias Cabaço, G., & Rodrigues, L. (2026). Artificial intelligence in pediatric inflammatory bowel disease: Applications in diagnosis, monitoring, and therapeutic decision-making. *Children, 13*(2), 260. https://doi.org/10.3390/children13020260

Dingel, J., Kleine, A.-K., Cecil, J., Sigl, A. L., Lermer, E., & Gaube, S. (2024). Predictors of health care practitioners' intention to use AI-enabled clinical decision support systems: Meta-analysis based on the unified theory of acceptance and use of technology. *Journal of Medical Internet Research, 26*, e57224. https://doi.org/10.2196/57224

Dupont, G., Kalinicheva, E., Sublime, J., Rossant, F., & Pâques, M. (2020). Analyzing age-related macular degeneration progression in patients with geographic atrophy using joint autoencoders for unsupervised change detection. *Journal of Imaging, 6*(7), 57. https://doi.org/10.3390/jimaging6070057

Elyoseph, Z., Hadar-Shoval, D., Asraf, K., & Lvovsky, M. (2023). ChatGPT outperforms humans in emotional awareness evaluations. *Frontiers in Psychology, 14*, 1199058. https://doi.org/10.3389/fpsyg.2023.1199058

Elyoseph, Z., Refoua, E., Asraf, K., Lvovsky, M., Shimoni, Y., & Hadar-Shoval, D. (2024). Capacity of generative AI to interpret human emotions from visual and textual data: Pilot evaluation study. *JMIR Mental Health, 11*, e54369. https://doi.org/10.2196/54369

Emegano, D. I., Mustapha, M. T., Ozsahin, D. U., Ozsahin, I., & Uzun, B. (2025). Histopathology-based prostate cancer classification using ResNet: A comprehensive deep learning analysis. *Journal of Imaging Informatics in Medicine, 39*(1), 604–619. https://doi.org/10.1007/s10278-025-01543-1

European Commission. (2024). *Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act).* European Commission. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689

European Commission. (2025). *Artificial intelligence in healthcare. Directorate-General for Health and Food Safety.* European Commission. https://health.ec.europa.eu/ehealth-digital-health-and-care/artificial-intelligence-healthcare_en

Foix, A. M., Cano, S., Osega, J., & Moreira, F. (2026). Affective user experience (AUX) in immersive environments: A systematic review of affective computing in immersive environments for individuals with Autism Spectrum Disorder (ASD). *Applied Sciences, 16*(3), 1528. https://doi.org/10.3390/app16031528

FDA, HC, & MHRA. (2021). *Good machine learning practice for medical device development: Guiding principles.* USA Food and Drug Administration (FDA), Health Canada (HC), & Medicines and Healthcare products Regulatory Agency (MHRA). https://www.fda.gov/medical-devices/software-medical-device-samd/good-machine-learning-practice-medical-device-development

Gerdes, A. (2024). The role of explainability in AI-supported medical decision-making. *Discover Artificial Intelligence, 4,* Article 29. https://doi.org/10.1007/s44163-024-00119-2

Gimeno, M., San José-Enériz, E., Villar, S., Agirre, X., Prosper, F., Rubio, A., & Carazo, F. (2022). Explainable artificial intelligence for precision medicine in acute myeloid leukemia. *Frontiers in Immunology, 13,* 977358. https://doi.org/10.3389/fimmu.2022.977358

Gu, S., Bao, T., Wang, T., Yuan, Q., Yu, W., Lin, J., Zhu, H., Cui, S., Sun, Y., Jia, X., Huang, L., & Ling, S. (2025). Multimodal AI diagnostic system for neuromyelitis optica based on ultrawide-field fundus photography. *Frontiers in Medicine, 12,* 1555380. https://doi.org/10.3389/fmed.2025.1555380

Gulati, S., Guleria, K., & Goyal, N. (2025). Privacy-preserving and collaborative federated learning model for the detection of ocular diseases. *International Journal of Mathematical, Engineering and Management Sciences, 10*(1), 218–248. https://doi.org/10.33889/IJMEMS.2025.10.1.013

Guleria, P. (2025). NLP-based clinical text classification and sentiment analyses of complex medical transcripts using transformer model and machine learning classifiers. *Neural Computing and Applications, 37*(1), 341–366. https://doi.org/10.1007/s00521-024-10482-x

Guo, Z., Li, X., Huang, H., Guo, N., & Li, Q. (2019). Deep learning-based image segmentation on multimodal medical imaging. *IEEE Transactions on Radiation and Plasma Medical Sciences, 3*(2), 162–169. https://doi.org/10.1109/TRPMS.2018.2890359

Haag, A. (2025). *The state of AI competition in advanced economies.* FEDS Notes. Board of Governors of the Federal Reserve System. https://www.federalreserve.gov/econres/notes/feds-notes/the-state-of-ai-competition-in-advanced-economies-20251006.html

Hacker, K. (2024). The burden of chronic disease. *Mayo Clinic Proceedings: Innovations, Quality & Outcomes, 8*(1), 112–119. https://doi.org/10.1016/j.mayocpiqo.2023.08.005

Hähnel, M. (2025). Ethical challenges and solutions in AI-driven medical data management: A focus on distributed machine learning. *Discover Artificial Intelligence, 5*, 53. https://doi.org/10.1007/s44163-025-00266-0

Hasselgren, A., Kralevska, K., Gligoroski, D., Pedersen, S. A., & Faxvaag, A. (2020). Blockchain in healthcare and health sciences: A scoping review. *International Journal of Medical Informatics, 134,* Article 104040. https://doi.org/10.1016/j.ijmedinf.2019.104040

Hewitt, K. J., Löffler, C. M. L., Muti, H. S., Berghoff, A. S., Eisenlöffel, C., van Treeck, M., Carrero, Z. I., El Nahhas, O. S. M., Veldhuizen, G. P., Weil, S., Saldanha, O. L., Bejan, L., Millner, T. O., Brandner, S., Brückmann, S., & Kather, J. N. (2023). Direct image to subtype prediction for brain tumors using deep learning. *Neuro-Oncology Advances,* *5*(1), 1–11. https://doi.org/10.1093/noajnl/vdad139

Hicks, S. A., Strümke, I., Thambawita, V., Hammou, M., Riegler, M. A., Halvorsen, P., & Parasa, S. (2022). On evaluation metrics for medical applications of artificial intelligence. *Scientific Reports, 12,* 5979. https://doi.org/10.1038/s41598-022-09954-8

Hidayat, D., Pangaribuan, C. H., Putra, O. P. B., & Irawan, I. (2021). Contemporary studies of task–technology fit: A review of the literature. *In 2021 International Conference on Information Management and Technology (ICIMTech) (pp. 309–313). IEEE.* https://doi.org/10.1109/ICIMTech53080.2021.9535028

Hu, H., Xu, W., Jiang, T., Cheng, Y., Tao, X., Liu, W., Jian, M., Li, K., & Wang, G. (2023). Expert-level immunofixation electrophoresis image recognition based on explainable and generalizable deep learning. *Clinical Chemistry,* *69*(2), 130–139. https://doi.org/10.1093/clinchem/hvac190

Huang, G., Chen, X., & Liao, C. (2025). AI-driven wearable bioelectronics in digital healthcare. *Biosensors,* *15*(7), 410. https://doi.org/10.3390/bios15070410

Ibrahim, M. M. A., Sumari, P., Keikhosrokiani, P., Almashagba, L. A. G., & Theeb, A. A. (2024). Exploring emotional intelligence in Jordan's artificial intelligence (AI) healthcare adoption: A UTAUT framework. *Journal of Electrical Systems,* *20*(10s), 502–541. https://doi.org/10.52783/jes.5143

Jasodanand, V. H., Bellitti, M., & Kolachalama, V. B. (2025). An AI-first framework for multimodal data in Alzheimer's disease and related dementias. *Alzheimer's & Dementia, 21*(9), e70719. https://doi.org/10.1002/alz.70719

Jayathissa, P., Rohatsch, L., Sauermann, S., & Hussein, R. (2025). OMOP-on-FHIR: Integrating the clinical data through the FHIR bundle to OMOP CDM. *Studies in Health Technology and Informatics, 327*, 667–671. https://doi.org/10.3233/SHTI250432

Jones, C. H., & Dolsten, M. (2024). Healthcare on the brink: Navigating the challenges of an aging society in the United States. *npj Aging, 10*(1), 22. https://doi.org/10.1038/s41514-024-00148-2

Kadri, F., Dairi, A., Harrou, F., & Sun, Y. (2022). Towards accurate prediction of patient length of stay at the emergency department: A GAN-driven deep learning framework. *Journal of Ambient Intelligence and Humanized Computing, 14*(9), 11481–11495. https://doi.org/10.1007/s12652-022-03717-z

Kalodanis, K., Feretzakis, G., Anastasiou, A., Rizomiliotis, P., Anagnostopoulos, D., & Koumpouros, Y. (2025). A privacy-preserving and attack-aware AI approach for high-risk healthcare systems under the EU AI Act. *Electronics,* *14*(7), 1385. https://doi.org/10.3390/electronics14071385

Kawamura, H., Miura, T., Maeda, Y., Okada, Y., & Zempo, K. (2025). Framework for emotion recognition using cross-modal transformers with non-contact multimodal signals aiming clinical service support. *IEEE Access, 13*, 99490–99502. https://doi.org/10.1109/ACCESS.2025.3573648

Kazerooni, A. F., Familiar, A. M., Aboian, M., Brüningk, S. C., Vossough, A., Linguraru, M. G., Huang, R. Y., Hargrave, D., Peet, A. C., Resnick, A. C., Storm, P. B., Mirsky, D., Yeom, K. W., Weller, M., Prados, M., Chang, S. M., Mueller, S., Villanueva-Meyer, J. E., Bakas, S., Fangusaro, J., Kann, B. H., & Nabavizadeh, A. (2025). Artificial intelligence for response assessment in pediatric neuro-oncology (AI-RAPNO), part 2: Challenges, opportunities, and recommendations for clinical translation. *The Lancet Oncology, 26*(11), e607–e618. https://doi.org/10.1016/S1470-2045(25)00489-9

Kergroach, S., & Héritier, J. (2024). *Emerging divides in the transition to artificial intelligence.* (OECD Regional Development Papers No. 147). OECD Publishing. https://doi.org/10.1787/7376c776-en

Khan, L., Khan, M. Z., & Aljubayri, I. (2026). A comprehensive framework for multi-modal depression detection: Integrating adaptive fusion, fairness regularization, and explainable AI. *Mathematics, 14*(4), 711. https://doi.org/10.3390/math14040711

Khoshdel, V., Asefi, M., Ashraf, A., & LoVetri, J. (2020). Full 3D microwave breast imaging using a deep-learning technique. *Journal of Imaging, 6*(8), 80. https://doi.org/10.3390/jimaging6080080

Khoshfekr Rudsari, H., Tseng, B., Zhu, H., Song, L., Gu, C., Roy, A., Irajizad, E., Butner, J., Long, J., & Do, K. A. (2025). Digital twins in healthcare: A comprehensive review and future directions. *Frontiers in Digital Health, 7,* 1633539. https://doi.org/10.3389/fdgth.2025.1633539

Kim, Y., Jang, T. G., Park, S. Y., Park, H. Y., Lee, J. A., Oyun-Erdene, T., Kim, S.-H., Park, Y. J., Cho, S. P., Park, J., Kang, D., & Urtnasan, E. (2025). Multimodal AI approach for the automatic screening of cardiovascular diseases based on nocturnal physiological signals. *npj Cardiovascular Health, 2*(1), 15. https://doi.org/10.1038/s44325-025-00051-z

Kumar, R., Singh, A., Kassar, A. S. A., Humaida, M. I., Joshi, S., & Sharma, M. (2025). Leveraging AI to achieve sustainable public healthcare services in Saudi Arabia: A systematic literature review of critical success factors. *Computer Modeling in Engineering & Sciences, 142*(2), 1289–1349. https://doi.org/10.32604/cmes.2025.059152

Lauritsen, S. M., Kristensen, M., Olsen, M. V., Larsen, M. S., Lauritsen, K. M., Jørgensen, M. J., Lange, J., & Thiesson, B. (2020). Explainable artificial intelligence model to predict acute critical illness from electronic health records. *Nature Communications, 11*(1), 3852. https://doi.org/10.1038/s41467-020-17431-x

Lee, C., Tzeng, C., Li, M., Lai, H., Chen, C., Huang, Y., Chang, T. A., Chen, C., Huang, C., Lee, M., & Liu, M. (2024). Leveraging federated learning for boosting data privacy and performance in IVF embryo selection. *Journal of Assisted Reproduction and Genetics, 41*(7), 1811–1820. https://doi.org/10.1007/s10815-024-03148-z

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors, 46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

Lee, S., Rheu, M., & Zhuang, J. (2025). The ChatGPT effect: Investigating shifting discourse patterns, sentiment, and benefit–challenge framing in AI mental health support. *Behavioral Sciences, 15*(9), 1172. https://doi.org/10.3390/bs15091172

Lekadir, K., Frangi, A. F., Porras, A. R., Glocker, B., Cintas, C., Langlotz, C. P., Weicken, E., Asselbergs, F. W., Prior, F., Collins, G. S., Kaissis, G., Tsakou, G., Buvat, I., Kalpathy-Cramer, J., Mongan, J., Schnabel, J. A., Kushibar, K., Riklund, K., Marias, K., Amugongo, L. M., … Starmans, M. P. A. (2025). FUTURE-AI: International consensus guideline for trustworthy and deployable artificial intelligence in healthcare. *BMJ, 388,* e081554. https://doi.org/10.1136/bmj-2024-081554

Li, J. G., Xu, K., Xiao, B., Li, J., Zhu, Y.-C., Jin, H., Qi, Q., Wang, L., Zhao, L., Wu, Z., Zhao, S., Zhang, T. J., & Wu, N. (2026). Addressing the diagnostic gap through deep phenotyping. *Human Genomics, 20*(1), 59. https://doi.org/10.1186/s40246-026-00925-y

Li, K., Lohachab, A., Dumontier, M., & Urovi, V. (2025). Privacy preservation in blockchain-based healthcare data sharing: A systematic review. *Peer-to-Peer Networking and Applications, 18*, 302. https://doi.org/10.1007/s12083-025-02148-9

Li, M., Jiang, Y., Zhang, Y., & Zhu, H. (2023). Medical image analysis using deep learning algorithms. *Frontiers in Public Health, 11,* 1273253. https://doi.org/10.3389/fpubh.2023.1273253

Li, X., Peng, L., Wang, Y.-P., & Zhang, W. (2025). Open challenges and opportunities in federated foundation models towards biomedical healthcare. *BioData Mining, 18,* 2. https://doi.org/10.1186/s13040-024-00414-9

Lin, H.-L., Wang, Y.-C., Huang, M.-L., Yu, N.-W., Tang, I., Hsu, Y.-C., & Huang, Y.-S. (2024). Can virtual reality technology be used for empathy education in medical students: A randomized case-control study. *BMC Medical Education, 24*(1), 1254. https://doi.org/10.1186/s12909-024-06009-6

Linardos, A., Pati, S., Baid, U., Edwards, B., Foley, P., Ta, K., Chung, V., Sheller, M., Khan, M. I., Jafaritadi, M., Kontio, E., Khan, S., Mächler, L., Ezhov, I., Shit, S., Paetzold, J. C., Grimberg, G., Nickel, M. A., Naccache, D., . . . Bakas, S. (2025). The MICCAI Federated Tumor Segmentation (FeTS) Challenge 2024: Efficient and robust aggregation methods for federated learning. *Machine Learning for Biomedical Imaging, 4,* 2025:033. https://doi.org/10.59275/j.melba.2025-5242

Lundberg, S. M., Nair, B., Vavilala, M. S., Horibe, M., Eisses, M. J., Adams, T., Liston, D. E., Low, D. K., Newman, S.-F., Kim, J., & Lee, S.-I. (2018). Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature Biomedical Engineering, 2*(10), 749–760. https://doi.org/10.1038/s41551-018-0304-0

Luo, H., Xu, G., Li, C., He, L., Luo, L., Wang, Z., Jing, B., Deng, Y., Jin, Y., Li, Y., Li, B., Tan, W., He, C., Seeruttun, S. R., Wu, Q., Huang, J., Huang, D.-W., Chen, B., Lin, S.-B., Chen, Q.-M., Yuan, C.-M., Chen, H.-X., Pu, H.-Y., Zhou, F., He, Y., & Xu, R.-H. (2019). Real-time artificial intelligence for detection of upper gastrointestinal cancer by endoscopy: A multicentre, case-control, diagnostic study. *The Lancet Oncology, 20*(12), 1645–1654. https://doi.org/10.1016/S1470-2045(19)30637-0

Ma, R., Cheng, Q., Yao, J., Peng, Z., Yan, M., Lu, J., Liao, J., Tian, L., Shu, W., Zhang, Y., Wang, J., Jiang, P., Xia, W., Li, X., Gan, L., Zhao, Y., Zhu, J., Qin, B., Jiang, Q., Wang, X., Lin, X., Chen, H., Zhu, W., Xiang, D., Nie, B., Wang, J., Guo, J., Xue, K., Cui, H., Cheng, J., Zhu, X., Hong, J., Shi, F., Zhang, R., Chen, X., & Zhao, C. (2025). Multimodal machine learning enables AI chatbot to diagnose ophthalmic diseases and provide high-quality medical responses. *npj Digital Medicine, 8*(1), 64. https://doi.org/10.1038/s41746-025-01461-0

Mahajan, S., & Helbing, D. (2026). Revisiting big data optimism: Risks of data-driven black box algorithms for society. *Ethics and Information Technology, 28*, 13. https://doi.org/10.1007/s10676-026-09888-z

Makarov, N., Bordukova, M., Quengdaeng, P., Garger, D., Rodriguez-Esteban, R., Schmich, F., & Menden, M. P. (2025). Large language models forecast patient health trajectories enabling digital twins. *npj Digital Medicine, 8*(1), 588. https://doi.org/10.1038/s41746-025-02004-3

Malik, S., & Rathee, P. (2024). Enhancing COVID-19 diagnosis accuracy and transparency with explainable artificial intelligence (XAI) techniques. *SN Computer Science, 5*(7), 806. https://doi.org/10.1007/s42979-024-03103-w

Marouf, A. A., Rokne, J. G., & Reda, R. (2025). Integrating multi-omics and medical imaging in artificial intelligence-based cancer research: An umbrella review of fusion strategies and applications. *Cancers, 17*(22), 3638. https://doi.org/10.3390/cancers17223638

Matheny, M. E., Thadaney Israni, S., Ahmed, M., & Whicher, D. (Eds.). (2023). Artificial intelligence in health care: The hope, the hype, the promise, the peril. *National Academies Press.* https://doi.org/10.17226/27111

Mohammed, S. G. A. A., Qoronfleh, M. W., Acar, A., & Al-Dewik, N. I. (2025). Holistic precision wellness: Paving the way for next-generation precision medicine (ngPM) with AI, biomedical informatics, and clinical medicine. *FASEB BioAdvances, 7*(4), e70005. https://doi.org/10.1096/fba.2024-00198

Moore, A., & Bell, M. (2022). XGBoost, a novel explainable AI technique, in the prediction of myocardial infarction: A UK Biobank cohort study. *Clinical Medicine Insights: Cardiology, 16,* 11795468221133611. https://doi.org/10.1177/11795468221133611

Muthukumar, K. (2025). Empathy AI in healthcare. *Frontiers in Psychology, 16,* 1680552. https://doi.org/10.3389/fpsyg.2025.1680552

Nazari, M., Kluge, A., Apostolova, I., Klutmann, S., Kimiaei, S., Schroeder, M., & Buchert, R. (2021). Explainable AI to improve acceptance of convolutional neural networks for automatic classification of dopamine transporter SPECT in the diagnosis of clinically uncertain parkinsonian syndromes. *European Journal of Nuclear Medicine and Molecular Imaging, 49,* 1176–1186. https://doi.org/10.1007/s00259-021-05569-9

Nemati, S., Holder, A., Razmi, F., Stanley, M. D., Clifford, G. D., & Buchman, T. G. (2018). An interpretable machine learning model for accurate prediction of sepsis in the ICU. *Critical Care Medicine, 46*(4), 547–553. https://doi.org/10.1097/CCM.0000000000002936

Ning, W., Zhu, Y., Song, C., Li, H., Zhu, L., Xie, J., Chen, T., Xu, T., Xu, X., & Gao, J. (2024). Blockchain-based federated learning: A survey and new perspectives. *Applied Sciences, 14*(20), 9459. https://doi.org/10.3390/app14209459

Nißen, M., Rüegger, D., Stieger, M., Flückiger, C., Allemand, M., Wangenheim, F. v., & Kowatsch, T. (2022). The effects of health care chatbot personas with different social roles on the client-chatbot bond and usage intentions: Development of a design codebook and web-based study. *Journal of Medical Internet Research, 24*(4), e32630. https://doi.org/10.2196/32630

Niu, S., Ma, J., Yin, Q., Wang, Z., Bai, L., & Yang, X. (2025). Modelling patient longitudinal data for clinical decision support: A case study on emerging AI healthcare technologies. *Information Systems Frontiers, 27*(2), 409–427. https://doi.org/10.1007/s10796-024-10513-x

Noor, A. A., Manzoor, A., Mazhar Qureshi, M. D., Qureshi, M. A., & Rashwan, W. (2025). Unveiling explainable AI in healthcare: Current trends, challenges, and future directions. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 15*(2), e70018. https://doi.org/10.1002/widm.70018

OECD. (2019). *Recommendation of the council on artificial intelligence (OECD/LEGAL/0449)*. Organisation for Economic Co-operation and Development (OECD) Publishing. https://doi.org/10.1787/eb3a0a9c-en

Orlova, A., Warner, D., & Reyes, S. (2017). AHIMA leading and influencing international standards for HIM practices. *Journal of AHIMA, 88*(11), 22–29.

Thunström, A. O., Carlsen, H. K., Ali, L., Larson, T., Hellström, A., & Steingrimsson, S. (2024). Usability comparison among healthy participants of an anthropomorphic digital human and a text-based chatbot as a responder to questions on mental health: Randomized controlled trial. *JMIR Human Factors, 11*, e54581. https://doi.org/10.2196/54581

Pahud de Mortanges, A., Luo, H., Shu, S. Z., Kamath, A., Suter, Y., Shelan, M., Pöllinger, A., & Reyes, M. (2024). Orchestrating explainable artificial intelligence for multimodal and longitudinal data in medical imaging. *npj Digital Medicine, 7*(1), 195. https://doi.org/10.1038/s41746-024-01190-w

Papachristou, K., Katsakiori, P. F., Papadimitroulas, P., Strigari, L., & Kagadis, G. C. (2024). Digital twins' advancements and applications in healthcare, towards precision medicine. *Journal of Personalized Medicine, 14*(11), 1101. https://doi.org/10.3390/jpm14111101

Partin, A., Brettin, T., Zhu, Y., Dolezal, J. M., Kochanny, S., Pearson, A. T., Shukla, M., Evrard, Y. A., Doroshow, J. H., & Stevens, R. L. (2023). Data augmentation and multimodal learning for predicting drug response in patient-derived xenografts from gene expressions and histology images. *Frontiers in Medicine, 10,* 1058919. https://doi.org/10.3389/fmed.2023.1058919

Parvin, N., Joo, S. W., Jung, J. H., & Mandal, T. K. (2025). Multimodal AI in biomedicine: Pioneering the future of biomaterials, diagnostics, and personalized healthcare. *Nanomaterials, 15*(12), 895. https://doi.org/10.3390/nano15120895

Pati, S., Kumar, S., Varma, A., Edwards, B., Lu, C., Qu, L., Wang, J. J., Lakshminarayanan, A., Wang, S., Sheller, M. J., Chang, K., Singh, P., Rubin, D. L., Kalpathy-Cramer, J., & Bakas, S. (2024). Privacy preservation for federated learning in health care. *Patterns, 5*(7), 100974. https://doi.org/10.1016/j.patter.2024.100974

Peng, J., Zou, K., Zhou, M., Teng, Y., Zhu, X., Zhang, F., & Xu, J. (2021). An explainable artificial intelligence framework for the deterioration risk prediction of hepatitis patients. *Journal of Medical Systems, 45*(5), 61. https://doi.org/10.1007/s10916-021-01736-5

Pfeifer, B., Baniecki, H., Saranti, A., Biecek, P., & Holzinger, A. (2022). Multi-omics disease module detection with an explainable Greedy Decision Forest. *Scientific Reports, 12*(1), 16857. https://doi.org/10.1038/s41598-022-21417-8

Phang, K. C., Ng, T. C., Singh, S. K. G., Voo, T. C., & Alvis, W. A. (2025). Navigating artificial intelligence in Malaysian healthcare: Research developments, ethical dilemmas, and governance strategies. *Asian Bioethics Review, 17,* 631–665. https://doi.org/10.1007/s41649-024-00314-4

Pinto-Coelho, L. (2023). How artificial intelligence is shaping medical imaging technology: A survey of innovations and applications. *Bioengineering, 10*(12), 1435. https://doi.org/10.3390/bioengineering10121435

Rahaman, M. A., Garg, Y., Iraji, A., Fu, Z., Kochunov, P., Hong, L. E., Van Erp, T. G. M., Preda, A., Chen, J., & Calhoun, V. (2024). Imaging-genomic spatial-modality attentive fusion for studying neuropsychiatric disorders. *Human Brain Mapping, 45*(17), e26799. https://doi.org/10.1002/hbm.26799

Räz, T., Pahud de Mortanges, A., & Reyes, M. (2025). Explainable AI in medicine: Challenges of integrating XAI into the future clinical routine. *Frontiers in Radiology, 5,* 1627169. https://doi.org/10.3389/fradi.2025.1627169

Revathi, G., & Mathew, O. C. (2026). Region guided Mask R-CNN with Haralick ResNet fusion for accurate coronary artery disease detection in computed tomography angiography images. *Scientific Reports, 16*(1), 12231. https://doi.org/10.1038/s41598-026-43951-5

Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H., Albarqouni, S., Bakas, S., Galtier, M. N., Landman, B., Maier-Hein, K., Ourselin, S., Sheller, M., Summers, R. M., Trask, A., Xu, D., Baust, M., & Cardoso, M. J. (2020). The future of digital health with federated learning. *npj Digital Medicine, 3,* 119. https://doi.org/10.1038/s41746-020-00323-1

Russ, P., Mross, P. M., Kräling, G., Lechner, F., Eldakar, M., Schlicker, N., Bedenbender, S., Brouwer, S., Zantvoort, K., Jerrentrup, A., Grgic, I., & Hirsch, M. C. (2025). Feasibility of a multimodal AI-based clinical assessment platform in emergency care: An exploratory pilot study. *Frontiers in Digital Health, 7,* 1657583. https://doi.org/10.3389/fdgth.2025.1657583

Sadeh-Sharvit, S., Camp, T. D., Horton, S. E., Hefner, J. D., Berry, J. M., Grossman, E., & Hollon, S. D. (2023). Effects of an artificial intelligence platform for behavioral interventions on depression and anxiety symptoms: Randomized clinical trial. *Journal of Medical Internet Research, 25*, e46781. https://doi.org/10.2196/46781

Sadeghi, Z., Alizadehsani, R., Cifci, M. A., Kausar, S., Rehman, R., Mahanta, P., Bora, P. K., Almasri, A., Alkhawaldeh, R. S., Hussain, S., Alatas, B., Shoeibi, A., Moosaei, H., Hladik, M., Nahavandi, S., & Pardalos, P. M. (2024). A review of explainable artificial intelligence in healthcare. *Computers & Electrical Engineering, 118,* 109370. https://doi.org/10.1016/j.compeleceng.2024.109370

Safak, O., Hekim, M. T., Çakmak, T., Demir, F., Akbulut, Y., Kadiroğlu, Z., Sengür, A., & Mehmet, A. K. (2026). Detection of occluded coronary arteries in non-ST-elevation myocardial infarction (NSTEMI) patients with deep learning models and ReliefF-based weighted subspace SVM ensembles (RBWSSE) algorithm. *PeerJ Computer Science,* e3576. https://doi.org/10.7717/peerj-cs.3576

Sakaguchi, M., Yoshizawa, A., Masui, K., Sakai, T., & Komori, T. (2025). AI-powered histology for molecular profiling in brain tumors: Toward smart diagnostics from tissue. *Cancers, 18*(1), 9. https://doi.org/10.3390/cancers18010009

Sartini, M., Carbone, A., Demartini, A., Giribone, L., Oliva, M., Spagnolo, A. M., Cremonesi, P., Canale, F., & Cristina, M. L. (2022). Overcrowding in the emergency department: Causes, consequences, and solutions—a narrative review. *Healthcare, 10*(9), 1625. https://doi.org/10.3390/healthcare10091625

Schlicher, M., Li, Y., Murthy, S. M. K., Sun, Q., & Schuller, B. W. (2025). Emotionally adaptive support: A narrative review of affective computing for mental health. *Frontiers in Digital Health, 7,* 1657031. https://doi.org/10.3389/fdgth.2025.1657031

Seza, K., Tawada, K., Kobayashi, A., & Nakamura, K. (2025). Multimodal artificial intelligence using endoscopic USG, CT, and MRI to differentiate between serous and mucinous cystic neoplasms. *Cureus, 17*(6), e85547. https://doi.org/10.7759/cureus.85547

Shamszare, H., & Choudhury, A. (2023). Clinicians' perceptions of artificial intelligence: Focus on workload, risk, trust, clinical decision making, and clinical integration. *Healthcare, 11*(16), 2308. https://doi.org/10.3390/healthcare11162308

Sheller, M. J., Edwards, B., Reina, G. A., Martin, J., Pati, S., Kotrotsou, A., & Bakas, S. (2020). Federated learning in medicine: Facilitating multi-institutional collaborations without sharing patient data. *Scientific Reports, 10*, 12598. https://doi.org/10.1038/s41598-020-69250-1

Shin, H., Park, J. E., Jun, Y., Eo, T., Lee, J., Kim, J. E., Lee, D. H., Moon, H. H., Park, S. I., Kim, S., Hwang, D., & Kim, H. S. (2023). Deep learning referral suggestion and tumour discrimination using explainable artificial intelligence applied to multiparametric MRI. *European Radiology, 33*(8), 5859–5870. https://doi.org/10.1007/s00330-023-09710-0

Shoshani, A., Gurfinkel, B., Kor, A., Ben-Haim, Y., Kanarek, O., Segev, R., Shafir, O., & Arbel, R. (2026). Efficacy of a conversational AI agent for psychiatric symptoms and digital therapeutic alliance: A randomized clinical trial. *JAMA Network Open, 9*(4), e266713. https://doi.org/10.1001/jamanetworkopen.2026.6713

Simon, B. D., Ozyoruk, K. B., Gelikman, D. G., Harmon, S. A., & Türkbey, B. (2025). The future of multimodal artificial intelligence models for integrating imaging and clinical metadata: A narrative review. *Diagnostic and Interventional Radiology, 31*(4), 303–312. https://doi.org/10.4274/dir.2024.242631

Sipos, D., Goyal, R., & Zapata, T. (2024). Addressing burnout in the healthcare workforce: Current realities and mitigation strategies. *The Lancet Regional Health Europe, 42,* 100961. https://doi.org/10.1016/j.lanepe.2024.100961

Sirapangi, M., & Gopikrishnan, S. (2024). MAIPFE: An efficient multimodal approach integrating pre-emptive analysis, personalized feature selection, and explainable AI. *Computers, Materials & Continua, 79*(2), 2229–2251. https://doi.org/10.32604/cmc.2024.047438

Tabl, A. A., Alkhateeb, A., ElMaraghy, W., Rueda, L., & Ngom, A. (2019). A machine learning approach for identifying gene biomarkers guiding the treatment of breast cancer. *Frontiers in Genetics, 10,* 256. https://doi.org/10.3389/fgene.2019.00256

Thacharodi, A., Singh, P., Meenatchi, R., Tawfeeq Ahmed, Z. H., Kumar, R. R. S., V, N., Kavish, S., Maqbool, M., & Hassan, S. (2024). Revolutionizing healthcare and medicine: The impact of modern technologies for a healthier future—a comprehensive review. *Health Care Science, 3,* 329–349. https://doi.org/10.1002/hcs2.115

Tong, B. G., Liang, Z., He, X., Yang, F., Yang, L., & Gao, L. (2025). AI-driven dynamic psychological measurement: Correcting university student mental health scales using daily behavioral and cognitive data. *Frontiers in Digital Health, 7,* 1615250. https://doi.org/10.3389/fdgth.2025.1615250

Tunyasuvunakool, K., Adler, J., Wu, Z., Green, T., Zielinski, M., Žídek, A., Bridgland, A., Cowie, A., Meyer, C., Laydon, A., Velankar, S., Kleywegt, G. J., Bateman, A., Evans, R., Pritzel, A., Figurnov, M., Ronneberger, O., Bates, R., Kohl, S. A. A., . . . Hassabis, D. (2021). Highly accurate protein structure prediction for the human proteome. *Nature, 596*(7873), 590–596. https://doi.org/10.1038/s41586-021-03828-1

Upreti, D., Yang, E., Kim, H., & Seo, C. (2024). A comprehensive survey on federated learning in the healthcare area: Concept and applications. *Computer Modeling in Engineering & Sciences, 140*(3), 2239–2274. https://doi.org/10.32604/cmes.2024.048932

Vajrobol, V., Saxena, G., Pundir, A., Singh, S., Gaurav, A., Bansal, S., Attar, R., Rahman, M., & Gupta, B. (2025). A comprehensive survey on federated learning applications in computational mental healthcare. *Computer Modeling in Engineering & Sciences, 142*(1), 49–90. https://doi.org/10.32604/cmes.2024.056500

Veeramani, N., S. A., R., S., S., S. P., S., S., & Jayaraman, P. (2025). NextGen lung disease diagnosis with explainable artificial intelligence. *Scientific Reports, 15,* 33052. https://doi.org/10.1038/s41598-025-07603-4

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly, 27*(3), 425–478. https://doi.org/10.2307/30036540

Venkatesh, V., Thong, J. Y. L., & Xu, X. (2012). Consumer acceptance and use of information technology: Extending the unified theory of acceptance and use of technology. *MIS Quarterly, 36(*1), 157–178. https://doi.org/10.2307/41410412

Wang, L., Zhang, Z., Wang, D., Cao, W., Zhou, X., Zhang, P., Liu, J., Fan, X., & Tian, F. (2023). Human-centered design and evaluation of AI-empowered clinical decision support systems: A systematic review. *Frontiers in Computer Science, 5,* 1187299. https://doi.org/10.3389/fcomp.2023.1187299

Wang, X., Xie, Y., Chen, X., Yang, J., Li, R., Gao, W., Yan, Z., Zhou, H., & Ye, Z. (2026). Securing federated learning with blockchain in the medical field: Systematic literature review. *Journal of Medical Internet Research, 28*, e79052. https://doi.org/10.2196/79052

World Health Organization. (2021). *Ethics and governance of artificial intelligence for health.* WHO guidance. https://www.who.int/publications/i/item/9789240029200

Wu, Y., Song, H., Ye, C., Lin, R., Huang, W., Wang, Y., & Yang, X. (2026). A pilot randomized controlled trial of AI-delivered vs. human-delivered iCBT for depression in young adults. *BMC Psychiatry, 26*(1), 239. https://doi.org/10.1186/s12888-026-07925-1

Xu, M., Zhai, F., & Zhang, T. (2025). Privacy protection in the application of artificial-intelligence technology in corporate governance. *International Journal of Information Security and Privacy, 19*(1), 1–17. https://doi.org/10.4018/IJISP.389733

Yan, R., Luo, H., Lu, J., Liu, D., Posluszny, H., Dhaliwal, M. P., MacLeod, J., Qin, Y., Yang, C., Hartman, T. J., & Hu, X. (2025). DietAI24 as a framework for comprehensive nutrition estimation using multimodal large language models. *Communications Medicine, 5*(1), 458. https://doi.org/10.1038/s43856-025-01159-0

Yeom, J. C., Kim, J. H., Kim, Y. J., Kim, J., & Kim, K. G. (2024). A comparative study of performance between federated learning and centralized learning using pathological images of endometrial cancer. *Journal of Digital Imaging, 37*(4), 1683–1690. https://doi.org/10.1007/s10278-024-01020-1

Yoojin, S., Lee, M., Lee, Y., Kim, K., & Kim, T. (2025). Artificial intelligence-powered quality assurance: Transforming diagnostics, surgery, and patient care—innovations, limitations, and future directions. *Life, 15*(4), 654. https://doi.org/10.3390/life15040654

Yu, Q., Ma, Q., Da, L., Li, J., Wang, M., Xu, A., Li, Z., Li, W., & Alzheimer's Disease Neuroimaging Initiative. (2024). A transformer-based unified multimodal framework for Alzheimer's disease assessment. *Computers in Biology and Medicine, 180,* 108979. https://doi.org/10.1016/j.compbiomed.2024.108979

Zhai, K., Masoodi, N. A., Zhang, L., Yousef, M. S., & Qoronfleh, M. W. (2022). Healthcare fusion: An innovative framework for health information management. *Electronic Journal of Knowledge Management, 20*(3), 179–192. https://doi.org/10.34190/ejkm.20.3.2968

Zhang, F., Kreuter, D., Chen, Y., Dittmer, S., Tull, S., Shadbahr, T., & Roberts, M. (2024). Recent methodological advances in federated learning for healthcare. *Patterns, 5*(6), 101006. https://doi.org/10.1016/j.patter.2024.101006

Zhang, K., Wang, D., Lin, F., Xie, J., & Zhou, W. (2026). A comprehensive review of explainable artificial intelligence in healthcare: Methods, evaluation, and clinical integration. *iScience, 29*(3), 115026. https://doi.org/10.1016/j.isci.2026.115026

Zhang, L., Zhong, Y., Yang, G., Huang, L., Deng, A., Ao, M., & Li, J. (2026). Artificial intelligence-assisted diagnosis and histopathological grading of bladder cancer: Current status, challenges, and future directions. *Frontiers in Digital Health, 8,* 1708289. https://doi.org/10.3389/fdgth.2026.1708289

Zhang, R., Chen, Y., Yue, W., Zhang, Y., Li, X., Feng, S., Yuan, F., & Luo, M. (2026). Multimodal artificial intelligence in medicine: A task-oriented framework for clinical translation. *Frontiers in Medicine, 12,* 1736272. https://doi.org/10.3389/fmed.2025.1736272

Supplementary Materials

Extended Table 1

Comparative Evaluation of Multimodal Artificial Intelligence Systems for AI-Driven Precision Healthcare Across Clinical Domains (2018–2026)

Clinical Domain	Clinical Application (Model)	Sample Characteristics & Pre-processing	Data Modalities Used	Relative Modality Burden	Fusion Strategy	AI Algorithm Used	XAI Integration	Validation (Internal/ External)	Performance Metrics (Metric vs. Control/ Benchmark)	Clinical Utility	Implementation Barriers & Mitigation Strategies	Generalizability Risk (High/Med/Low)	Clinical Readiness Level
Cardiology (Demir et al., 2025)	CHD Early Detection (SMOTE Integrated ML)	4,240 (Framingham) & 303 (UCI); Imbalance: 3,596 healthy vs 644 patients; Missingness: Handled via Hotdecking (nearest neighbor); Dimensionality: 16 attributes; Normalization: StandardScaler; Augmentation: SMOTE synthetic samples.	Clinical attributes + Lab tests (cholesterol) + Demographic attributes (age, sex, smoking)	Low: Easily accessible basic clinical patient records and laboratory cardiovascular risk indicators	Type I (Feature-level): Clinical and lab features integrated into a single vector prior to classification. Handles imbalanced data through synthetic sampling	SVM (with Hotdecking & SMOTE); XGBoost, LSTM, ANN	Feature Importance (Linear SVM coefficients and Input layer weights).	Internal: Random 80/20 splitting	AUC 0.925 / F1 92.75% (SVM on UCI).	Long-term cardiovascular risk prediction and individualized clinical management.	B: Imbalanced datasets,Bias & lack of diversity. M: Data preprocessing via Hotdecking and SMOTE for consistency.	Medium: Uses well-known public datasets but requires broader testing for generalization.	Validation Stage: Demonstrated improved stability via confidence interval analysis.
Cardiology (Safak et al., 2026)	NSTEMI Occluded Artery Prediction (RAL-CNN)	Expert-labeled digital 12-lead 907 ECG records; NSTEMI dataset from ER. Imbalance: LAD majority. Augmentation: Hybrid balancing (random oversampling, pitch shifting, noise injection, and class-weighting applied only to training dataset).	12-lead ECG spectrograms.	Medium: Real-time digital acquisition and multi-lead ECG processing.	Type II (Classifier-level): Parallel Attention and Residual branches process signals before merging in LSTM layers. Relief-Based Weighted Subspace Ensembles (RBWSSE)	RAL-CNN (Residual+ Attention+ LSTM)	Spatial Uncertainty (SUE) and Grad-CAM features.	Internal: 10-fold cross-validation.	Accuracy 81.3% (vs. base CNN 65.3%).	Non-invasive triage to determine need for urgent invasive CAG.	B: Computational complexity, interpretability, bias. M: Optimization for embedded/real-time use, hybrid balancing, lightweight design.	High: Limited size, Single-center training, LAD occlusion bias.	Validation Stage: Baseline established for multi-center studies. Needs real-time monitoring validation.

Clinical Domain	Clinical Application (Model)	Sample Characteristics & Pre-processing	Data Modalities Used	Relative Modality Burden	Fusion Strategy	AI Algorithm Used	XAI Integration	Validation (Internal/ External)	Performance Metrics (Metric vs. Control/ Benchmark)	Clinical Utility	Implementation Barriers & Mitigation Strategies	Generalizability Risk (High/Med/Low)	Clinical Readiness Level
Cardiology Kim et al. (2025)	CVD Screening (SleepCVD-Net)	194 subjects (80 controls, 114 CVD), 232,817 1-s ECG segments from Sleep Heart Health Study. Demographic bias: 89.1% White; Imbalanced: Control signals (96k) vs. Stroke (42k). Two-layer batch normalization for raw signal normalization, detrending, correcting baseline wandering	Nocturnal single-lead ECG, Airflow, and SpO2	Medium: Requires standard nocturnal physiological recording equipment and sensors	Type II (Intermediate/Feature-level): Concatenation of modality-specific feature maps into a single vector before dense layers.	1D CNN (3 modes)	None mentioned	Internal: Training/ Validation /Test split	Mean Accuracy: 97.55% vs. SVM (87.40%) and biLSTM (84.70%); F1-score: 0.96 (Stroke) to 0.97 (CHF/Angina).	Automatic screening and monitoring of CVD during sleep to reduce economic burden of diagnosis	(B) Bias & lack of diversity: homogeneous White cohort. High economic burden of CVD treatments. (M) SleepCVD-Net reduces costs and time via automated nocturnal screening.	High (Single database source [SHHS] from multicenter cohort, demographic homogeneity, no external validation).	Validation Stage (Comparison with common ML benchmarks (biLSTM, SVM) but remains an extension tool and not primary diagnostic standard).
Cardiology Revathi et al., (2026)	CAD Detection (DeepConvNet)	206 patients' CCTA images (5.130 billion pixels). Data augmentation to adjust for imbalance, Demographics: potential imbalance in age/sex. Pre-processing: Min-max normalization, histogram equalization, Gaussian smoothing; flipping augmentation. Acquisition bias (variability in scanners/BMI); Annotation bias from human segmentations.	CCTA images unimodal source but 2 features extracted: Haralick texture + Deep hierarchical ResNet	High: Requires CCTA imaging and expert annotation	Type II (Intermediate/Feature-level): Hand-crafted Haralick texture and deep hierarchical ResNet feature maps fusion using fully connected layers (comparable to multimodal fusion strategies)	Region-Guided Mask R-CNN; DeepConvNet	XAI method used. Feature ranking (ASM/IDM identified as top Haralick features).	Internal: Cross-validation	Accuracy: 98.3%, F1-score: 98.2%, AUC: 0.98; Dice: 0.91 and IoU 0.87 vs U-Net (0.89 Dice, 0.85 IoU) and V-Net (0.90 Dice, 0.86 IoU),	High-precision segmentation and classification of coronary structures	(B) Model interpretability gaps: high complexity. Dependency on high-quality labeled data. (M) Optimization reduces computation burden time to 5.5 min.	High (Inherent complexity + variability of CCTA images, may not generalize well across all patient demographics or scan qualities).	Validation Stage (Extensive algorithmic benchmarking and internal cross-validation but lacks multi-center prospective implementation).

Clinical Domain	Clinical Application (Model)	Sample Characteristics & Pre-processing	Data Modalities Used	Relative Modality Burden	Fusion Strategy	AI Algorithm Used	XAI Integration	Validation (Internal/ External)	Performance Metrics (Metric vs. Control/ Benchmark)	Clinical Utility	Implementation Barriers & Mitigation Strategies	Generalizability Risk (High/Med/Low)	Clinical Readiness Level
Chronic Disease Management Sirapangi & Gopikrishnan, 2024	Pre-emptive Analysis (MAIPFE), IoT Monitoring	35k samples. Multiple IoMT datasets (Breast Cancer, Diabetes, Heart Disease), sparse text/sensor data. Preprocessing: FFO feature optimization; Imputation of missing values; feature scaling.	Physiological signals (IoT wearable device) + Environmental sensors + patient demographics	Medium: Requires continuous wearing of sensors and user adherence	Type I (Feature-Level): Firefly Optimizer (FFO) selects optimal subset of features from disparate IoT and environmental sources, then integrated into a unified feature vector and processed by RNN	Firefly Optimizer (FFO) feature selection + RNN + Fuzzy C Means (FCM)	DeepSHAP model for feature attribution and rankings	Internal: Stratified training and validation splits.	Acc: 94.15%, Precision: 96.72%, AUC: 92.31% (NTS 35k samples) vs. FETLM models	Real-time health monitoring, preemptive analysis of health risks and actionable recommendations	B: Computational intensity, data heterogeneity (noise); privacy. M: Firefly Optimizer for low-delay personalized feature selection, feature engineering, RFE	Medium: Tested on multiple standard datasets; real-world noise/adherence is a risk.	Prototype Stage: Hypothetical results presented vs. existing methods, rigorous evaluation
Critical Care Nemati et al., 2018	Real-time Sepsis Prediction (AISE)	~42,000 patients (MIMIC-III freely accessible); 27,527 patients in Emory University Hospital cohort. Imbalance: Sepsis-3 definitions (rare events); Dimensionality: Coupling varying resolutions (EMR + Vitals).	EMR data (low-res) + physiological vital signs (high-res BP/HR)	Medium: Requires continuous bedside monitoring systems.	Type II (Classifier-level): Coupled resolution learning integrates time-series vitals (varying resolution inputs) with static EMR data	Coupled Resolution Learning	Notification basis visualization to show factor contribution	Internal & External: Multi-center (MIMIC-II/Emory)	AUROC 0.87 (4-hour prediction window)	Real-time sepsis prediction in intensive care units before organ failure	B: EMR recall/information bias,lack of diversity, Interpretability gaps. M: Tele-ICU monitoring layer to mitigate recall bias	Low: Developed on Emory data and externally validated on MIMIC-III database.	Prospective validation stage: Large-Scale External Validated, integrated remote monitoring team with notification basis visualization, performance reliability prior to clinical recognition

Clinical Domain	Clinical Application (Model)	Sample Characteristics & Pre-processing	Data Modalities Used	Relative Modality Burden	Fusion Strategy	AI Algorithm Used	XAI Integration	Validation (Internal/ External)	Performance Metrics (Metric vs. Control/ Benchmark)	Clinical Utility	Implementation Barriers & Mitigation Strategies	Generalizability Risk (High/Med/Low)	Clinical Readiness Level
Critical Care Lundberg et al., 2018	Hypoxaemia Prediction (Prescience)	50,000+ surgeries university-affiliated hospitals; some exclusions for non-physiological bypass data. Preprocessing: Exponentially decaying averages to capture time-registered data.	Static EHR data (drugs, history) + Intraoperative dynamic features (real-time vitals)	Low: Diagnostic data routinely collected in standard surgical EHR	Type I: Summarized unevenly sampled data into fixed-length feature vectors (grouped of time-series and static data) before classification.	XGBoost (Gradient Boosting)	SHAP values (Succinct visual summaries)	Internal: Validated on held-out operating room test sets.	AUC: 0.81 (Prescience) vs. Anaesthesiologist average (AUC: 0.66)	Real-time prediction of desaturation to enable preemptive treatment.	B: High feature dimensionality (3,905 features). M: Exponentially decaying averages to capture time-registered data, SHAP-based feature importance summaries for experts.	Medium: Large sample but limited to specific academic medical centers.	Validation Stage: Validated on held-out surgical test sets. Prospective trials during live procedures are required before clinical deployment
Critical Care / Emergency Care Lauritsen et al., 2020	Early Warning System (XAI-EWS)	163,050 admissions, Danish cohort (CROSS-TRACKS). Imbalance: Sepsis 2.44%, AKI 0.75%. Preprocessing: 1-hour interval aggregation; no resampling used.	EHR (27 labs + 6 vitals)	Low: Routine non-invasive EHR data (history/basic symptoms).	Type I: Disparate vitals and labs integrated into a single unified sequential matrix (time-steps x features) before Temporal Convolutional Network (TCN) prediction + DTD explanation.	Temporal Convolutional Network (TCN)	Deep Taylor Decomposition (DTD) explanation module for temporal relevance	Internal: Temporal 5-fold cross-validation on a large Danish representative cohort.	AUROC: 0.79-0.92 for Sepsis, AKI, and ALI vs. MEWS/SOFA	Real-time prediction of acute critical illness to enable timely intervention	B: Low disease prevalence (imbalance), lack of transparency in deep learning. M: Weighted relevance attribution, Visual explanations (DTD) for clinical trust.	Medium: Highly representative of the Danish population; needs other ethnic validation.	Validation Stage: Temporal cross-validation performed on a large Danish cohort. Needs validation across ethnic and geographic diversity

Clinical Domain	Clinical Application (Model)	Sample Characteristics & Pre-processing	Data Modalities Used	Relative Modality Burden	Fusion Strategy	AI Algorithm Used	XAI Integration	Validation (Internal/ External)	Performance Metrics (Metric vs. Control/ Benchmark)	Clinical Utility	Implementation Barriers & Mitigation Strategies	Generalizability Risk (High/Med/Low)	Clinical Readiness Level
Critical Care / Emergency Care Russ et al., 2025	Modular ED Assessment Platform	20 ambulatory non-urgent patients. Accessibility: Excludes cognitive impairment + language barriers. Preprocessing: Brief data collection; survey feedback.	Vitals (Sensors) + Speech (audio/text via LLM) + Structured Clinical History	Low: Short, automated, non-invasive clinical history and vital sign capture	Type II: Voice and sensor subnetworks processed independently then merged by LLM.	Locally running LLM (Mistral Small 3)	Staff-reviewed explanations and understandable diagnoses	Internal: Exploratory pilot feasibility and usability study.	SUS Score: 90.6 (Excellent usability benchmark is >80.3)	Streamlining ED triage and documentation to reduce staff workload.	B: Bias & lack of diversity (small sample). EHR integration and potential clinician rejection. M: User-centric design and human-in-the-loop validation.	High: Small-scale pilot; single-center; spectrum bias (low-acuity only).	Prototype Stage: Exploratory pilot study focused on feasibility, usability, and patient trust rather than diagnostic accuracy or clinical impact
Emergency Medicine Kadri et al., (2022)	LOS Prediction (GAN-P)	Pediatric Emergency Department historical data (2011–2012), 1400-2200 arrivals/month. Missingness: Missing EHR data corrected via KNN/Random Forest imputation. Pre-processing: Min-max normalization; Mapping descriptors to LOS.	Historical EHR data, medical, and laboratory info (12 heterogeneous descriptors)	Low: Retrospective EHR and lab data	Type I (Early/Feature-level): Mapping integrated observation descriptors to LOS values (same fusion logic as multimodal system)	GAN-driven predictor; Supervised fine-tuning of learned distribution s	None mentioned	Internal: Regional pediatric ED cohort. Actual data from Lille Regional Hospital center, France. Comparison with CNN, DBN, RF	R2: 0.871 (GAN-P) vs. 0.836 (CNN); RMSE: 100.309 (Best/Lowest); and Deep Belief Network (0.858).	Forecasting future trends of patient flow/admissions for hospital resource planning and overcrowding prevention	Missing data in EHR datasets. Mitigation: k-nearest neighbors and Random Forest regressor used for data imputation to improve data quality.	High (Single-center, single regional hospital source).	Validation Stage (Extensive comparative testing against DBN/SAE + pediatric medical experts, but single-site retrospective evaluation).

Clinical Domain	Clinical Application (Model)	Sample Characteristics & Pre-processing	Data Modalities Used	Relative Modality Burden	Fusion Strategy	AI Algorithm Used	XAI Integration	Validation (Internal/ External)	Performance Metrics (Metric vs. Control/ Benchmark)	Clinical Utility	Implementation Barriers & Mitigation Strategies	Generalizability Risk (High/Med/Low)	Clinical Readiness Level
Gastro-enterology/ Oncology Seza, K., et al. (2025)	Cyst Differ-entiation (MI-Quad)	25 patients (SCN) + 24 patients (MCN). Small cohort from multicenter study. Data augmentation: 39,200 images per modality. Imbalance: limited subjects. Pre-processing: Cropped to squares, resized to 256x256 pixels, converted to 8-bit grayscale.	EUS, ECECT, T2, MRP, and Sex	High: Requires multiple advanced imaging modalities (MRI, CT, specialized EUS)	Type II (Intermediate/Feature-level): Classifier-level fusion of feature maps extracted after modality-specific deep learning	ResNet (Residual Network) as the deep learning backbone	None mentioned	Internal: Comparison with Expert accuracy	Accuracy: 99.0% (MI-Quad) vs. 81.0% (Human Experts); Sensitivity: 98.0%; Specificity: 100%	Differentiating serous and mucinous cystic neoplasms to prevent unnecessary surgery and guide treatment	(B) Data heterogeneity & domain shift: unknown optimal fusion. Small participant cohort. (M) AI minimizes info loss by learning directly from raw images compared to discrete clinical values.	High (Small sample)	Prototype Stage (Proof-of-concept AI combination outperforming human experts in differentiating neoplasms but optimal fusion strategy remains unknown and must be adapted to different clinical factors).

Clinical Domain	Clinical Application (Model)	Sample Characteristics & Pre-processing	Data Modalities Used	Relative Modality Burden	Fusion Strategy	AI Algorithm Used	XAI Integration	Validation (Internal/ External)	Performance Metrics (Metric vs. Control/ Benchmark)	Clinical Utility	Implementation Barriers & Mitigation Strategies	Generalizability Risk (High/Med/Low)	Clinical Readiness Level
Hepatology Peng et al., 2021	Hepatitis Deterioration Risk (XAI Framework)	155 patients; 19 features (UCI). Imbalance: 20.6% death rate; Missingness: Imputation. Nominal features set to majority, continuous to average; Augmentation: SMOTE applied.	19 different Clinical + Biological inspections	Low: Standard clinical and laboratory tests.	Type I (Feature-level): Combines 19 clinical + biological features integrated into a single structured input vector. Type III (Decision-level) : Complex ensemble integration of individual ML outputs (RF, XGBoost, SVM).	Random Forest (Ensemble)	SHAP, LIME, and PDP for global/local interpretation .	Internal: K-fold cross-vali dation (K=20).	Accuracy 91.9% (RF vs. LR Accuracy 87.5% at K=20).	Identifying disease deterioration risk and guiding personalized treatment.	B: Data heterogeneity, Interpretability gaps, distrust. M: Global/local XAI to improve transparency.	High: Small single-center public dataset (UCI benchmark). More real-world data is needed for validation	Validation Stage: A feasible, interpretable and tested framework reliably forecasts hepatitis exacerbation risk using benchmark data
Infectious Disease Malik & Rathee, 2024	COVID-19 Screening (XAI Screening Tool)	50,000+ individuals (Israeli Ministry of Health dataset). Imbalance: RT-PCR labels; Bias: Simulation of self-reporting bias by removing negative symptoms.	Demographics (age/sex) + contact history + binary clinical symptoms	Low: Based on basic patient records and subjective reported symptoms.	Type I (Feature-level): Integration of binary clinical features and history into a tree-based gradient boosting. Sequentially trains weaker learners (DTs) leveraging knowledge.	LightGBM (Gradient Boosting)	SHAP algorithm and dependency graphs for feature influence.	Internal: RT-PCR assay as ground truth; validation on independe nt held-out test set (30%)	Accuracy 92.82% (but low AUC 0.58).	Efficient screening + triaging, optimizing resource management during pandemic waves in resource-limit ed settings.	B: Dependency on subjective symptom reporting (reduced trust), Interpretability gaps. M: XAI (SHAP) for transparent factor attribution, trust	High: Relies on single-source Israeli dataset with self-reported symptoms	Validation Stage: Feasibility study for assisting clinical triage and resource allocation.

Clinical Domain	Clinical Application (Model)	Sample Characteristics & Pre-processing	Data Modalities Used	Relative Modality Burden	Fusion Strategy	AI Algorithm Used	XAI Integration	Validation (Internal/ External)	Performance Metrics (Metric vs. Control/ Benchmark)	Clinical Utility	Implementation Barriers & Mitigation Strategies	Generalizability Risk (High/Med/Low)	Clinical Readiness Level
Laboratory Medicine Hu et al., 2023	PCD Diagnosis (IFE Recognition)	12,703 IFE images (Cohort-based). Imbalance: 84% normal vs 16% abnormal. Preprocessing: Two-stage classification (normal/abnormal then patterns); Image-specific Score-CAM band identification.	Immunofixation Electrophoresis (IFE) images (unimodal input)	High: Requires complex lab preparation and invasive blood/urine collection	Type III (Decision-Level): Integrates outputs from ensemble of 3 deep CNN architectures (ResNet18, VGG16, MobileNetV2) to produce final image recognition labels	Ensemble of 3 Deep CNNs (ResNet18, VGG16, MobileNet V2)	Score-CAM heatmaps for band identification	Internal & External: Tested across different Helena IFE systems.	Acc: 99.8%, Sens: 92.56% vs. Junior Experts (Acc: 99.40%, lower precision)	Automation to reduce interpretative variability and expert workload. Advanced ensemble complexity to evaluate AI robustness	B: Shortage of experts and extremely PCD rare patterns, Bias & lack of diversity. M: AI-driven automation for consistency and rule-based systems for rare patterns.	Low: Validated on external datasets from different imaging systems.	Deployment-ready: Human expert-level performance; multi-system validated proving high generalizability.
Longitudinal Care / CDSS Niu et al., 2025	Disease Risk Prediction (DSLAM), Longitudinal EHR	9,759 unstructured medical notes (MIMIC-III subset). Missingness: High sparsity. Preprocessing: Stop-word removal; Clinical-BERT embedding.	Unstructured medical notes + Risk label descriptions	Low: Retrospective, non-invasive data from EHR medical notes	Type II (Intermediate): Adaptive cross-attention mechanism integrating label-dependent attention network encoded separately, then latent state-space update.	Deep State-space (DSSM) with Label-dependent Attention	Label-dependent attention scores to highlight clinical terms	Internal: Comparison on MIMIC-III and N2C2 datasets.	F1: 0.899, ROCAUC: 0.876 (MIMIC-III) vs. RETAIN (F1: 0.882)	Tracking latent health trajectories across multiple hospital visits	B: High dimensionality, unstructured text noise and data sparsity. M: Clinical-BERT for embedding joint word-label and deep state-space modeling to capture time-varying information for transitions.	Medium: Validated on two major public longitudinal datasets	Validation Stage: Performance evaluated against two real-world EHR datasets and outperformed RNN-based baselines but captured latent states are still challenging to interpret at a broader medical domain level

Clinical Domain	Clinical Application (Model)	Sample Characteristics & Pre-processing	Data Modalities Used	Relative Modality Burden	Fusion Strategy	AI Algorithm Used	XAI Integration	Validation (Internal/ External)	Performance Metrics (Metric vs. Control/ Benchmark)	Clinical Utility	Implementation Barriers & Mitigation Strategies	Generalizability Risk (High/Med/Low)	Clinical Readiness Level
Mental Health Khan et al., 2026	Depression Detection (Comprehensive Multimodal Framework, DGN+MHAN)	Three public datasets: DAIC-WOZ (N=189), StudentSADD (N=300), and Moodable (N=200+). Imbalance: Inherent in mental health sets; Augmentation: Pitch shift, color jittering, and temporal shifting.	Audio (acoustic) + Video (visual/facial) + Text (linguistic)	High: High-dimensional data requiring synchronized real-time pipelines for recording, automated text transcription and linguistic analysis.	Type II (Adaptive): Dynamic Gating Network (DGN) weighs modalities; Multi-Head Attention (MHAN) captures interactions.	Transformer encoders with RoBERTa/ Wav2Vec2.0/ResNet	Attention maps mapped back to video frames, facial/acoustic/text tokens and SHAP.	Internal: Subject-independent splits (70/15/15).	93.0% Accuracy (DAIC-WOZ classification); F1-Score 91.4% (DAIC-WOZ dataset vs. static fusion ~82.4%).	Automated continuous screening, timely interventions, continuous monitoring of symptoms, and decision support	B: Data imbalance, algorithmic bias, and real-world noise. M: Fairness regularization, context-aware-gating, XAI integration, and data augmentation.	Medium: Tested on 3 datasets with limited diversity in cultural/ethnical backgrounds.	Validation Stage: High performance achieved, Experts rated clarity/plausibility, future work for long-term temporal tracking.
Molecular Biology Tunyasuvunakool et al., 2021	Human Proteome structure (AlphaFold)	20,296 proteins, 2.7M residues (human proteome); binned by resolution/confidence. Preprocessing: MSA search on UniRef90; template search on PDB seqres.	1D Protein sequence data + 3D PDB structural templates	High: Requires extremely expensive advanced multi-modal real-time synchronized experimental structures (PDB) data.	Type II: Sequence and structural templates processed via independent subnetworks, then attention-based integrated in Evoformer. Full chain prediction for inter-domain packing	Evoformer (Transformer-based)	Per-residue pLDDT confidence maps (color-coded)	Internal & External: Validated on held-out PDB set and CAMEO benchmark.	AUC: 0.897 for pLDDT confidence vs experimental resolved head	Structure-based function and druggable pocket detection	B: Computational cost of training, Data heterogeneity. M: Cached intermediates and multi-GPU workers for open-source inference.	Low: Covers 98.5% of human proteins; externally validated.	Deployment-ready: Validated against PDB/CAMEO benchmarks covering 98.5% of human proteins. Broadly available tool for biological research.

Clinical Domain	Clinical Application (Model)	Sample Characteristics & Pre-processing	Data Modalities Used	Relative Modality Burden	Fusion Strategy	AI Algorithm Used	XAI Integration	Validation (Internal/ External)	Performance Metrics (Metric vs. Control/ Benchmark)	Clinical Utility	Implementation Barriers & Mitigation Strategies	Generalizability Risk (High/Med/Low)	Clinical Readiness Level
Neurology Yu et al., 2024	AD Diagnosis & MCI conversion (AD-Transformer)	1,651 subjects (ADNI). Imbalance: AD, MCI, CN cohorts; Missingness: Variables >30% missing excluded, RF imputation. Bias: ~90% white, mostly North American (lacks racial/ethnic diversity). Augmentation: Patch-based CNN.	3D sMRI + clinical data (neuropsych/functional) + genetic data	High: Requires advanced 3D neuroimaging and genomic sequencing, expensive acquisition/complex pipelines.	Type II (Adaptive): Unified Transformer processes linear projections of non-image data and image tokens jointly across modalities.	Hybrid Patch-CNN + Transformer Encoder	Grad-CAM and SHAP for voxel/token importance, Saliency Maps	Internal: 5-fold cross-validation.	AUC 0.99 (vs. image-only ResNet 0.87).	Early diagnostic aid for personalized management and proactive disease management.	B: Missing image modalities, high modality burden, bias, low diversity. M: Imputation techniques, gather diverse global data, future use of masked modeling	High: Lacks racial diversity; paucity of external genetic data.	Validation Stage: Feasibility proven on ADNI dataset.
Neurology Gu, S., et al. (2025)	NMO Diagnosis (MAiDS-NMO)	1,618 images (330 NMO+, 1,288 NMO-); 330 associated with clinical records. Demographic bias: Single-center Chinese population; Imbalanced: 219 NMO vs. 798 non-NMO in training. Pre-processing: QC filtering (focus clarity ≥70%); segmentation model used to remove motion artifacts.	UWF fundus images + Clinical reports (MRI optic nerve/spine, AQP4 antibody tests)	High: Specialized UWF imaging and specialist clinical MRI/lab tests.	Type III (Late/Decision-level): Weighted integration of image scores and clinical criteria scores, Conditional decision approach	Deep Neural Network; Inception-V3	Heatmaps to visualize optic disc hyperemia. Specialized technicians ensured scanning quality for visual interpretability.	Internal: 5-fold cross-validation	AUC: 0.9923 (Multimodal) vs. 0.9751 (Image-only) from single 80:10:10 train/validation/test split; Sensitivity: 97.0%; Specificity: 96.9%	Assisting diagnosis in settings with limited specialized expertise	(B) High modality burden: Access to MRI/specialists. (M) Optimized for diagnostic accuracy for suspected NMO in resource-limited clinics.	High (Single-center, imbalanced, no multi-ethnic cohort).	Validation Stage (Systematic internal validation + Comparative test with ophthalmologists, Lack of an external validation dataset).

Clinical Domain	Clinical Application (Model)	Sample Characteristics & Pre-processing	Data Modalities Used	Relative Modality Burden	Fusion Strategy	AI Algorithm Used	XAI Integration	Validation (Internal/ External)	Performance Metrics (Metric vs. Control/ Benchmark)	Clinical Utility	Implementation Barriers & Mitigation Strategies	Generalizability Risk (High/Med/Low)	Clinical Readiness Level
Neuropsychiatry Rahaman et al., 2024	Schizophrenia (SZ) Classification (Spatio-modality Fusion)	N=437 (162 SZ vs 275 HC). Datasets: COBRE, fBIRN, MPRC. Pre-processing: Segmentation/normalization, GICA for sMRI; GWAS for SNPs; standard imaging protocols. Joint training.	sMRI (Structural) + fMRI (Functional) + Genomics (SNPs)	High: Extremely expensive and invasive MRI scans and genomic sequencing	Type II: Modality-specific subnetworks (AE, MLP, bi-LSTM) processed independently with mid-fusion of subnetworks regulated by spatio-modality attention module (inspired by Bottleneck Attention, BAM), then merged before final layer.	Spatio-modality mid-fusion with bottleneck attention (BAM)	Inherent interpretability via attention scores for modality/context	Internal: Reproducibility checked via multiple random data splits.	Acc: 94.1%, F1: 0.697 vs. Unimodal Genomics (Acc: 70.8%), Unimodal fMRI (Acc: 81.1%)	Discovery of reliable biomarkers (e.g., CSMD1, ATK3 genes) for mental disorders	B: High-dimensional, Data heterogeneity, noisy multimodal data, High modality burden. M: Spatio-modality attention regulation to identify relevant subspaces.	Medium: Merged cohorts increase diversity but needs ethnic validation. High-dimensional data is noisy.	Validation Stage: Reproducibility checked via multiple random splits and merged diverse cohorts. Remains a research framework for biomarker discovery rather than a deployed clinical tool

Clinical Domain	Clinical Application (Model)	Sample Characteristics & Pre-processing	Data Modalities Used	Relative Modality Burden	Fusion Strategy	AI Algorithm Used	XAI Integration	Validation (Internal/ External)	Performance Metrics (Metric vs. Control/ Benchmark)	Clinical Utility	Implementation Barriers & Mitigation Strategies	Generalizability Risk (High/Med/Low)	Clinical Readiness Level
Nutrition Yan et al., 2025	Dietary Estimation (DietAI24)	3,000 ASA24 images + 1,000 Nutrition5k images. Bias: Behavioral Hawthorne effect; Diversity: Nutrition5k limited to U.S. cafeteria foods; Accessibility: controlled photo settings.	Smartphone food Images + authoritative FNDDS nutrient database	Medium: Real-time smartphone photography + Nutrient databases	Type II (Adaptive): Specialized Retrieval-Augmented Generation (RAG) integrates GPT vision model outputs with nutrient databases.	RAG-based Multimodal LLM (GPT Vision)	RAG grounds the model's visual reasoning in authoritative nutrition databases (FNDDS) to prevent hallucinations. Expert review (two dietitians) qualitative evaluation (Likert scale, clinical plausibility)	Internal: Validated on standardized dietary image datasets.	98.9% Success Rate (Food Recognition)	Epidemiological studies and individualized precision nutrition	B: Behavioral bias, Interpretability gaps (hallucinations). M: RAG grounds prediction models in authoritative databases.	Medium/high: Controlled data settings; lacks international cuisines.	Prototype + Validation stage: Deployment requires end-to-end encryption for image transmission and integrating diverse databases to ensure equitable performance
Oncology Tabl et al., 2019	Breast Cancer Gene Biomarkers (Proposed Model, Hierarchical RF)	347 patients. Imbalance: High. Handled via one-vs-rest scheme and SMOTE; Dimensionality: Narrowed 24k genes via feature selection (mRMR).	High-throughput Genomic profiles (DNA Microarray) + Clinical data	High: High-throughput genomic sequencing required; high-dimensional data requiring complex pipelines.	Type I (Feature-level): Integrated genomic/clinical features used at five sequential tree-based classification nodes.	Hierarchical Random Forest	Circos plots for biomarker gene correlation	Internal: 10-fold cross-validation	Accuracy 80.9% - 100% across five hierarchical nodes	Personalizing treatment (Surgery/Radiation/Hormone) based on survivability biomarkers	B: High dimensionality, Bias, Imbalanced datasets, lack of heterogeneity. M: Cost-sensitive classifiers, Feature selection (mRMR) narrows 24k genes to a handful.	High: Small public benchmark dataset; highly imbalanced nodes.	Validation Stage: Feasibility study to identify gene biomarkers

Clinical Domain	Clinical Application (Model)	Sample Characteristics & Pre-processing	Data Modalities Used	Relative Modality Burden	Fusion Strategy	AI Algorithm Used	XAI Integration	Validation (Internal/ External)	Performance Metrics (Metric vs. Control/ Benchmark)	Clinical Utility	Implementation Barriers & Mitigation Strategies	Generalizability Risk (High/Med/Low)	Clinical Readiness Level
Oncology Partin et al., 2023	Pharmacology Drug Response Prediction in PDX (MM-Net/UMH-Net)	6,962 drug response values from 89 patients. PDX models. Imbalance: 917 non-response vs 42 response; Augmentation: Pseudo drug-pair	Gene Expression (GE) + Histology WSIs + Drug chemical descriptors	High: Requires multi-omics profiling, WSI scanning and digital pathology.	Type II (Classifier-level): Modality-specific subnetworks fused via a feature concatenation layer.	MM-Net, CNN (Subnetworks + Concatenation)	Qualitative visual evaluation of enriched feature space.	Internal: 100 repeated data splits (10-fold repeated).	MCC 0.310 (vs. LGBM baseline 0.259), 0.798 AUROC	Guiding precision oncology drug screening and pharmaceutical development.	B: High dimensionality (overfitting), Limited sample sizes, lack of diversity. M: Data augmentation (pseudo drug-pairs), homogenization of representations.	High: Highly imbalanced classes (PDX models); metrics vary greatly across splits.	Prototype: Tool for identifying candidates for preclinical screening
Oncology Guo et al., 2019	Sarcoma Lesion Segmentation (Multimodal 3D CNN)	50 patients. Balance: Equal number of positive/negative patches extracted. Pre-processing: Rigid registration via commercial software. Addressed voxel misalignment.	PET + CT + T1/T2-weighted MRI	High: Requires multiple concurrent radiologic scanning modalities.	Type I (Feature-level): Fusion within convolutional layers via 3D tensors (best performance).	Supervised Cross-Modality CNN (and U-Net)	3D Surface Visualization and label maps allowing comparison to ground truth manual annotations	Internal: 10-fold cross-validation.	DICE 0.85 (PET+CT+T2 vs. PET-only DICE 0.76).	Automated contouring/segmentation of soft tissue sarcoma lesions for treatment planning (radiotherapy).	B: Voxel-level registration errors across modalities (misalignment), data heterogeneity, domain shift, high modality burden. M: In-network fusion, Shared feature iterative segmentation/registration.	High: Small sample (N=50); assumes perfect voxel correspondence.	Prototype / Validation Stage: Proven robust to low-quality/noised images. Test more complex structures in future work.

Clinical Domain	Clinical Application (Model)	Sample Characteristics & Pre-processing	Data Modalities Used	Relative Modality Burden	Fusion Strategy	AI Algorithm Used	XAI Integration	Validation (Internal/ External)	Performance Metrics (Metric vs. Control/ Benchmark)	Clinical Utility	Implementation Barriers & Mitigation Strategies	Generalizability Risk (High/Med/Low)	Clinical Readiness Level
Oncology Emegano et al., 2025	Prostate Cancer Classification (ResNet50)	1,276 histological prostate biopsy images from mono-tertiary institutions. Imbalance: 761 Benign vs. 515 Malignant. Preprocessing: Data augmentation (rotation/flip) to 10k samples; feature scaling.	Histopathology-based prostate biopsy images (unimodal input)	High: Highly invasive surgical biopsy procedure, 2D tissue staining and processing	N/A: Unimodal (Exclusively uses a single model architecture, ResNet50)	ResNet50 (Residual Learning)	Grad-CAM for visualization of stromal and atypical arrangement, clinical feature matching	Internal: Held-out validation data reserved.	ResNet50 Acc: 98.0%, AUC: 0.98, F1: 0.98 benign vs F1: 0.96 malignant vs. Swin Transformer benchmark Acc: 95%	Standardizing diagnostic interpretation to reduce over/under-di agnosis, and improve pathological workflows. Domain benchmarking + high clinical utility in critical precision healthcare tasks with expert-level performance.	B: High computational requirements, Data heterogeneity (staining variability), inability to include radiological images and lack of multi-site external validation. M: Transfer learning, data augmentation, ResNet to mitigate redundancy and training time.	High: Single-center data; lacks diverse scanner/staining validation.	Validation Stage: Benchmarked against SOTA models; needs multi-center trial validation.

Clinical Domain	Clinical Application (Model)	Sample Characteristics & Pre-processing	Data Modalities Used	Relative Modality Burden	Fusion Strategy	AI Algorithm Used	XAI Integration	Validation (Internal/ External)	Performance Metrics (Metric vs. Control/ Benchmark)	Clinical Utility	Implementation Barriers & Mitigation Strategies	Generalizability Risk (High/Med/Low)	Clinical Readiness Level
Oncology Abbas et al., (2024)	Cancer Classification (Adaptive FL)	35,000 synthetic images (histology, CT, MRI) across 7 subclasses. Imbalanced: Subclasses (Glioma, normal kidney, etc.) split across 3 hospitals, Pre-processing: RGB to grayscale; resizing to 28x28x1; noise removal.	Histology slides, CT scans, and MRIs (Fused across brain, kidney, breast datasets).	Medium to High: Advanced clinical imaging required	Type II (Intermediate/Adaptive): Self-adaptive weights synchronization with a global server to handle multidisciplinary data.	CNN + Self-adaptive Federated Learning	None mentioned	Internal: Cloud-based simulation validation	Accuracy: 90.0% (Adaptive FL) vs. 87.3% (Conventional FL); F1-score: 0.86 (Kidney tumor) to 1.0 (Breast).	Cooperative multidisciplinary model training without sharing private patient data	(B) Data heterogeneity & domain shift: lack of Non-IID data. No information-privileged access and cyber-attack risks. (M) Federated Learning edge security.	Medium (Multi-institutional simulation but imbalanced).	Prototype (Simulated using MATLAB environment, requires the integration of the design into a real-world healthcare setting).
Oncology Deacon et al., (2025)	ROBIN Assay (Methylome), Neurosurgery	Intraoperative nanopore-based DNA sequencing 50 cases (prospective cohort). Missingness: Data sparsity during early intraoperative sequencing. Pre-processing: Adapted transposase protocol for rapid turnaround	unimodal source but multiple genomic features derived (methylation, CNV, SNV)	High: Requires intraoperative tissue and nanopore sequencing equipment.	Type II (Intermediate/Feature-level): Unified profiling integrating genomic variants and methylome features	Neural network (Sturgeon/ CrossNN)	None mentioned	External/ Prospective (Prospective validation in surgical settings).	Classification: 76% (38/50 cases) classified within 1 hr of sequencing.	Real-time intraoperative tumor unified molecular profiling with multidisciplinary precision approach	(B) High modality burden: Slow turnaround for molecular features. (M) Adapted transposase protocol enables sample prep in 90 min.	Medium (Prospective validation but limited single-workflow cohort).	Validation (Research use only; comprehensive assessment of clinical impact is needed).

Clinical Domain	Clinical Application (Model)	Sample Characteristics & Pre-processing	Data Modalities Used	Relative Modality Burden	Fusion Strategy	AI Algorithm Used	XAI Integration	Validation (Internal/ External)	Performance Metrics (Metric vs. Control/ Benchmark)	Clinical Utility	Implementation Barriers & Mitigation Strategies	Generalizability Risk (High/Med/Low)	Clinical Readiness Level
Oncology Luo et al., 2019	GI Cancer Detection (GRAIDS)	Over 1 million images. Bias: Demographic variability (age, sex, prevalence) across hospitals. Pre-processing: Strict exclusion of non-endoscopic or inconsistent images	High-scale (over 1 Million) endoscopic images (White light) unimodal source	Medium to High: Requires expert endoscopic capture.	Type III (Late/Decision-level): Cloud-based real-time analysis	Deep learning (GRAIDS system)	None mentioned	External (Multi-center: 5 provincial hospitals across China) & Prospective.	Accuracy: 0.915–0.977 (External) vs. Trainee experts (0.886).	Real-time clinical detection of upper GI cancer. Benchmark value to determine if multimodal systems (like MI-Quad) can provide a statistically significant synergistic effect over unimodal standard	(B) Data heterogeneity & domain shift. Stable cloud connection dependency for real-time analysis. (M) Implementation of dedicated monitors adjacent to endoscopy hardware.	Low (Multi-center external validation + large diverse cohort).	Deployment-ready (Implemented in SYSUCC's endoscopic clinical practice).
Oncology / Multi-Domain Makarov et al., 2025	Trajectory Digital Twin (DT-GPT)	N=52,767 (merged 16,496 patients Flatiron Health EHR database (NSCLC), 35,131 patients MIMIC-IV dataset (ICU), 1,140 patients Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset). Missingness: 94.4%. Preprocessing: Text encoding medical history; two-step outlier filtering.	Longitudinal Electronic Health Records (EHR) + Labs + Vitals (Sensors) + Drug administrations + Speech (LLM) + Cognitive scores (MMSE, CDR-SB, and ADAS11)	Low: Standard clinical records from demographics and history.	Type I (Feature-level): Heterogeneous longitudinal EHR data encoded into a single chronological text prompt before adaptive LLM processing.	Digital Twin—GPT (Fine-tuned LLM BioMistral-7B)	Conversational chatbot for prediction rationale, reasoning and important variables. Understandability/Predictability rated at 4.12/5.0	Internal: Exploratory pilot study; Benchmarked against 14 SOTA models across 3 datasets.	SUS Score: 90.6 (Excellent usability), Scaled MAE: 0.55 (NSCLC) vs. LightGBM (0.57); AUC: 0.70 (classification)	Streamlining ED triage and documentation to reduce staff workload, proactive health management, zero-shot prediction of non-target labs and adverse event mitigation	B: Data sparsity, Data heterogeneity & domain shift (noise/misspellings), bias. M: Human-in-the-loop, Fine-tuned Foundation models capable of handling missing data and noise, zero-shot capability.	Medium: Large datasets used but primarily retrospective/curated EHR data	Validation Stage: Benchmarked against 14 models across 3 datastes. Robust to real-world EHR noise, but needs to enhance sensitivity to high-risk outcomes

Clinical Domain	Clinical Application (Model)	Sample Characteristics & Pre-processing	Data Modalities Used	Relative Modality Burden	Fusion Strategy	AI Algorithm Used	XAI Integration	Validation (Internal/ External)	Performance Metrics (Metric vs. Control/ Benchmark)	Clinical Utility	Implementation Barriers & Mitigation Strategies	Generalizability Risk (High/Med/Low)	Clinical Readiness Level
Oncology / Neurology Shin et al., 202	Brain Tumor Triage & Referral Suggestion (Interpretable DL System)	877 patients; independent test set (n=130). Imbalance: 30 entities (rare and common); Bias: Single-institution retrospective; Normalization: Skull stripping and de-identification pre-processing.	Multi-parametric MRI (T1, T1CE, FLAIR, DWI, ADC) sequences	High: Multiparametric MRI (T1, T1CE, FLAIR, DWI, ADC) sequences required.	Type III (Decision-level) : Hierarchical CNN. Sequential classifiers discriminate tumor vs. non-tumor then recommend triage.	Hierarchical CNN	Layer-wise Relevance Propagation (LRP) for multi-contrast heatmaps.	Internal: Tested on retrospective held-out independent test set (n=130)	AUC 0.90 (vs. neuroradiologists 0.81-0.88; vs. residents 0.51-0.92).	Specialist-level triage and referral tool for neuroradiologists in ER settings to prevent erroneous resection.	B: Subjectivity in management and referral pathways, interpretability and diversity gaps. M: Multi-contrast heatmaps for quantified decision basis (transparency).	High: Retrospective single-center design; limited to lesions >2cm.	Validation Stage: Comparable to neuroradiologists in triage tasks.
Oncology / Neurology Hewitt et al., 2023	Diffuse Brain Tumor (Glioma) Subtype Prediction (Direct Image-to-Subtype, WHO CNS)	N=2,845 2016/2021 WHO guideline cohorts: UCL (UK, n=1,882), TCGA (n=864), & CPTAC (n=99). Imbalance: UCL cohort chosen for more balanced classes; Domain Shift: Addressed across different scanner platforms. Missingness: Addressed in alterations. Preprocessing: Normalization of tiles to mitigate domain shift.	Digitized Whole-slide histopathology images (unimodal input)	High: Requires highly invasive neurosurgical procedures for 2D pathological tissue slides + Genetic/molecular profiling.	Type III (Sequential): Stepwise stacking reconstruction of WHO multimodal diagnostic algorithms using separate unimodal predictive modules to determine final output.	Stepwise Sequential Deep Learning (Convolutional, attMIL: attention multiple-instance learning)	Morphological heatmaps generated for prediction drivers + feature transparency.	Internal & External: 5-fold cross-val and independent test cohorts, TCGA/CPTAC	AUROC: 0.95 (Glioblastoma) and 0.95 (IDH alteration) vs. Internal sets.	Accurate + faster diagnostic classification prediction of core molecular alterations directly from histology for individualized treatment. Unimodal input predicting multimodal outputs, bridging the gap between morphology + genetics.	B: Data heterogeneity, Domain shift between scanners, high modality burden. M: Normalization of tiles, training with multi-centric cohorts.	Low: Validated on independent external cohorts (TCGA/CPTAC) with different scanners.	Validation Stage: Robust performance across independent external cohorts and 2016/2021 WHO diagnostic criteria. Domain shift needs addressing in diverse datasets.

Clinical Domain	Clinical Application (Model)	Sample Characteristics & Pre-processing	Data Modalities Used	Relative Modality Burden	Fusion Strategy	AI Algorithm Used	XAI Integration	Validation (Internal/ External)	Performance Metrics (Metric vs. Control/ Benchmark)	Clinical Utility	Implementation Barriers & Mitigation Strategies	Generalizability Risk (High/Med/Low)	Clinical Readiness Level
Ophthalmology Ma et al. (2025)	Self-Triage Chatbot (IOMIDS)	15,640 entries from 9,825 subjects across 10 subspecialties. Missingness: Diagnostic accuracy drops (63.5% to 20.2%) if patient data is incomplete. Pre-processing: YOLOv7 for eye detection in smartphone images; score rescaling to [-1, 1].	Patient text (History) + Slit-lamp or Smartphone images	Low to Medium: Smartphone (low) vs. Slit-lamp (medium)	Type III (Late/Decision-level): Integration of image labels and preliminary diagnoses via interactive prompts	ChatGPT-3.5 + ResNet50 + YOLOv7	None mentioned	External: Multi-center: Shanghai, Nanjing, Suqian	Accuracy (External): 81.1% (Text+Smartphone) vs. 72.5% (Text only)	Patient self-diagnosis and self-triage Decision Support System	(B) Model interpretability gaps: General LLMs lack precision for complex cases and localized medical exams. (M) Proprietary LLM-based prompt engineering via clinical dialogs.	Low (External validation + diverse cohort across 10 subspecialties).	Validation (Two-stage clinical evaluation with patients but prospective evaluations are still necessary).
Pediatrics / Psychiatry Albahri et al., 2024	Autism Triage Model (Fuzzy MCDM)	538 autistic patients. Imbalanced classes: Minor: 241, Moderate: 259, Urgent: 38. Preprocessing: Cost-sensitive learning and oversampling; imputation of missing values.	42 Medical and Sociodemographic criteria	Medium: Requires extensive patient history and expert psychologist participation	Type I: Unified vector created via Fuzzy MCDM (FWZIC) weighting of 19 criteria.	Logistic Regression (optimized)	LIME for local approximations of feature contribution	Internal: 10-fold cross-validation on two labeled sets.	F1: 0.977, AUC: 0.999 (Logistic Regression) vs. SVM (F1: 0.165)	Early identification and triage to reduce family stress and costs	B: Highly imbalanced classes, dataset noise, Bias & lack of diversity. M: Cost-sensitive learning, Fuzzy set theory to handle uncertainty, oversampling	High: Small, imbalanced single-center dataset; lacks external validation.	Prototype: Initial novel framework combining FDM, FWZIC, and PTAP tested on a relatively small, single-center dataset

Clinical Domain	Clinical Application (Model)	Sample Characteristics & Pre-processing	Data Modalities Used	Relative Modality Burden	Fusion Strategy	AI Algorithm Used	XAI Integration	Validation (Internal/ External)	Performance Metrics (Metric vs. Control/ Benchmark)	Clinical Utility	Implementation Barriers & Mitigation Strategies	Generalizability Risk (High/Med/Low)	Clinical Readiness Level
Psychiatry Guleria, 2025	Text Classification (Sentiment Model), NLP	4,998 samples. Imbalanced: Consult/History entries (high) vs. Dermatology (low). Pre-processing: Tokenization, stop word removal, and lemmatization; TF-IDF.	Medical transcripts (22 subspecialties) unimodal source	Low: Textual reports/EHR notes.	Type I: Unified vectorization via BERT embeddings before primary classification	LSTM + BERT; Multi-head attention.	Word ranking visualization (visualizing most significant phrases).	Internal: Comparison with SVM, RF, NB (MATLAB/ Jupyter)	Accuracy: 94% (LSTM) vs. SVM (65%); F1-score: 90%; Precision: 87%; Recall: 93%.	Automated classification of complex medical transcripts.	(B) Data heterogeneity & domain shift: High-dim medical vocabulary. (M) BERT pre-trained embeddings improve semantic understanding.	High (Single dataset source)	Validation (testing and training validation, LSTM and BERT against ML classifiers)
Psychiatry Chen et al., 2024a	Depression Assessment (ANN-Depress), Digital Health	183 participants (101 patients with Major Depressive Disorder and 82 controls), 156 subjects (80 MDD, 76 controls) actigraphy, 108 subjects (61 MDD, 47 controls) app-based measurement. Including expression, voice, and actigraphy features. Pre-processing: PRAAT voice feature extraction; nparACT actigraphy analysis. Bias: Small sample	Passive actigraphy, App usage, Active facial expression + voice, and NLP.	Low: Passive data collected via smartphone sensors and actigraphy.	Type II (Intermediate/Feature-level): Fusion of diverse digital modality features for common analysis	Artificial Neural Network (ANN);	None mentioned.	Internal (Leave-one-out cross-validation).	F1-score: 0.81 (Multimodal) vs. Subjective happiness only (0.49) or Actigraphy only (0.55).	Detecting depression state and trait via passive monitoring.	(B) Bias & lack of diversity: Data privacy and ethical considerations in passive monitoring. (M) Secure home-use digital assessment systems.	High (Single-center, small sample)	Prototype (small sample size, needs larger dataset validation to support its clinical feasibility)

Clinical Domain	Clinical Application (Model)	Sample Characteristics & Pre-processing	Data Modalities Used	Relative Modality Burden	Fusion Strategy	AI Algorithm Used	XAI Integration	Validation (Internal/ External)	Performance Metrics (Metric vs. Control/ Benchmark)	Clinical Utility	Implementation Barriers & Mitigation Strategies	Generalizability Risk (High/Med/Low)	Clinical Readiness Level
Pulmonology Biswas et al., 2025	Pneumonia Detection (FLPneXAINet)	8,402 CXR images (Benchmark Kaggle set). Imbalance: 3,904 normal vs 4,498 pneumonia. Corrected via synthetic data generation. Preprocessing: CycleGAN for data augmentation; RFE, ANOVA, and RF for feature optimization.	Chest X-ray (CXR) + CycleGAN synthetic/augmented data	Medium: Requires specialized imaging equipment (X-rays) but is non-invasive	Type II: Modality-specific subnetworks fused before the final layer via concatenation of extracted features (VGG16/MobileNet) from parallel CNN extractors.	Ensemble Deep Learning (VGG16 + MobileNet) with SVM classifier	Grad-CAM heatmaps and LIME segmentation	Internal: Local/Federated environment with four diverse local entities.	Acc: 97.61%, F1: 98.36%, Recall: 98.13% vs. EfficientNetB0 (Acc: 95.19%)	Secure, automated screening in resource-limited or data-sensitive settings.	B: Data heterogeneity & domain shift, Small local datasets. M: Federated Learning for privacy and Cycle-GAN for augmentation.	Medium: Uses benchmark sets but lacks broad multi-site external testing.	Validation Stage: Demonstrated high performance in federated environments. Lacks real-world multi-site clinical testing
Pulmonology Veerammi et al., (2025)	Lung Disease (XAI- TRANS)	7,560 Chest X-ray images. Balanced dataset: 5 classes, ~1500 images/class (COVID, TB, Bacterial or Viral Pneumonia, Normal). Pre-processing: Improved U-Net segmentation; pixel normalization and data augmentation.	Chest X-ray images unimodal source	Medium: Standard CXR required.	Type II: Modality-specific (segmentation + classification) feature maps fused before final diagnosis	Inception-V3 + U-Net.	LIME and Grad-CAM integration for localized lung opacity.	Internal (Kaggle dataset)	Accuracy: 97.53% (Multiclass) vs. VGG16 (84%); Precision: 98.95%; F1-score: 0.9797.	Rapid screening of infectious lung diseases (COVID-19, TB, pneumonia). Foundational insights into XAI-TRANS, core requirement for adaptive AI governance.	(B) Public dataset dependency (M) external validation	Low-Medium (Balanced set; public source)	Validation (systematic technical validation with public Kaggle datasets)

Note. Fusion approaches were classified as described in Guo et al., 2019. Abbreviations: Acc, accuracy; AD, Alzheimer's disease; AUC/AUROC, area under the receiver operating characteristic curve; BAM, bottleneck attention module; CAD, coronary artery disease; CCTA, coronary computed tomography angiography; CDSS, clinical decision support system; CV, cross-validation; CVD, cardiovascular disease; DSSM, deep state-space model; DTD, Deep Taylor Decomposition; ECG, electrocardiography; EMR, electronic medical record; F1, F1-score; FI, feature importance; FFO, firefly optimizer; GI, gastrointestinal; NMO, neuromyelitis optica; SMOTE, synthetic minority oversampling technique; SNP, single nucleotide polymorphism; SUS, system usability scale; TCN, temporal convolutional network; WSI, whole-slide imaging.

WESTCLIFF
UNIVERSITY
Educate. Inspire. Empower.

Published by: Westcliff University | 17877 Von Karman Ave, 4th Floor, Irvine CA 92614
949-825-5999 | wijar.westcliff.edu | wjournal@westcliff.edu