Journal of Science Research International (JSRI)
ISSN: 2456-6365 (Online)

OPEN ACCESS

A SURVEY ON AI-DRIVEN TECHNIQUES FOR AUTOMATED MEDICAL DIAGNOSIS AND CONSULTATION SYSTEM

Sabitha Praveen Madamby*, Nidhi Pravin Patil,

Ashlesha Sunil Parte, Saradnya Harshad Majarekar and Awaiz Asif Shaikh

Department of Computer Science,

Pillai College of Arts, Commerce and Science (Empowered Autonomous), New Panvel

**Corresponding author E-mail: sabitha.praveen@mes.ac.in*

Received: 13 January 2026	Revised: 13 February 2026	Accepted: 29 March 2026	Published: 17 April 2026

DOI: **https://doi.org/10.5281/zenodo.19622904**

Abstract:

The integration of artificial intelligence (AI) into healthcare is transforming medical diagnosis and patient consultation. This survey examines advances in automated diagnostic systems that combine natural language processing, machine learning, and data extraction to interpret clinical records and support evidence-based decision-making. Key technologies include text recognition for digitizing handwritten and scanned records, large-scale language models (LLMs) for generating clinical summaries, and intelligent retrieval methods that improve chatbot-driven consultations. We review approaches ranging from domain-adapted LLMs to OCR-NLP pipelines and retrieval-augmented generation techniques, highlighting shared trends, core strengths, and limitations. Special attention is given to multi-agent system simulations for testing AI reasoning in clinical workflows and privacy-preserving machine learning protocols that protect patient data while meeting regulatory standards. Persistent challenges include model generalizability, maintaining high diagnostic accuracy, scaling to real-world clinical volumes, and ethical, transparent deployment. Future directions include multimodal AI architectures, federated and continual learning, and interpretable models that build trust among clinicians and patients. This work provides a reference for researchers and practitioners developing AI-powered medical consultation systems aimed at safer, more accessible, patient-centered healthcare.

Keywords: *Artificial Intelligence, Healthcare, Diagnostic Systems, Natural Language Processing, Machine Learning, Clinical Decision Support,*

Introduction

Artificial intelligence (AI) is transforming healthcare by advancing automated diagnosis, clinical decision support, patient consultation, and administrative processes, largely driven by large language models (LLMs) and generative AI. These technologies provide scalable, accessible, and cost-efficient solutions, addressing clinician

shortages and expanding care in underserved areas through virtual consultations and symptom-assessment tools. Modern AI systems, including transformer-based models like GPT-4, Med-PaLM 2, and LLaMA-2, can interpret complex patient data—ranging from unstructured clinical notes to imaging studies—and generate multi-task predictions from electronic health records (EHRs). Techniques such as retrieval-augmented generation (RAG), OCR-based digitization, and privacy-preserving protocols enhance reliability, accuracy, and regulatory compliance. This survey examines AI methodologies for automated medical diagnosis and consultation, including LLMs, predictive modeling, explainable AI, data intake automation, and hybrid multi-agent systems, highlighting trends, clinical relevance, and deployment challenges. By consolidating practical insights, comparative evaluations, and open research directions, this work provides a foundational reference for developing transparent, trustworthy, and effective AI-driven healthcare solutions that work synergistically alongside human clinicians.

Table 1: Key Requirements and Challenges in Automated AI Healthcare

Requirement/Challenge	**Description**
Data Availability and Annotation	Sufficient, labeled, high-quality multi-institutional datasets
Standardization and Interoperability	Use of standard clinical terminologies, seamless EHR integration
Privacy and Security	Patient data anonymization (k- anonymity, federated learning), compliance with GDPR/HIPAA
Transparency and Explainability	Model interpretability, rationale generation for decisions, trust
Clinical Validation	Prospective, multi-site, real-world evaluation studies
Ethical Trust and Fairness	Bias mitigation, equitable model performance across groups
Human-AI Collaboration	Co-pilot models, hybrid doctor +AI workflows, multi-agent simulation
Robustness and Generalizability	Performance across diverse sites, tasks, and populations

Literature analysis

Large language models (LLMs) such as GPT-4, Med-PaLM 2, LLaMA-2, BioBERT, and Health-LLM are adapted for healthcare, excelling in question answering, clinical reasoning, triage, report generation, and doctor recommendation, with Med-PaLM 2 reaching 86.5% on USMLE-style tasks. Multi-agent simulations like AI Hospital highlight challenges in context retention and dynamic reasoning. Explainability methods (LIME, SHAP, attention maps, saliency, counterfactuals) support trust and regulatory compliance. Multi-task and multi-source EHR models (e.g., GenHPF) improve cross-institutional diagnosis and outcome prediction, aided by transfer and federated learning. OCR/NLP pipelines structure clinical documents using UMLS/SNOMED, scaling via semi-supervised learning. Privacy-preserving approaches (k-anonymity, differential privacy, federated learning) align with regulations like HIPAA and GDPR. Benchmarking shows LLMs often match or surpass clinicians, and hybrid human-AI workflows enhance accuracy and trust. Overall, medical LLMs demonstrate expert-level capability, though challenges persist in interpretability, generalization, compliance, and seamless human-AI collaboration.

AI-driven medical diagnosis and consultation employ a spectrum of methods, from classical machine learning (SVM, logistic regression, random forests) and deep learning (CNNs, RNNs) to large language models (LLMs) and hybrid frameworks. Classical ML excels on structured data for risk prediction and disease classification, while DL handles imaging, sequential EHR data, and biomedical signals. Graph neural networks (GNNs) model complex

patient interactions, and LLMs (GPT-3/4, Med-PaLM 2, Health-LLM, BioBERT, LLaMA-2) combined with retrieval-augmented generation (RAG) provide advanced reasoning, clinical QA, report generation, triage, and personalized recommendations.

Other approaches include multi-agent simulations (AI Hospital, Dr Rank) for end-to-end clinical evaluation, OCR/NLP pipelines for digitizing and structuring unstructured records, and privacy-preserving frameworks (k-anonymity, federated learning) to ensure compliance. Hybrid human-AI co-pilot workflows reduce diagnostic errors, improve trust, and mitigate LLM hallucinations, though challenges remain in multi-turn reasoning, context retention, explainability, and large-scale deployment. Overall, the literature shows a progression from early expert systems to sophisticated, multi-modal, interpretable AI systems with hybrid human oversight.

Performance metrics and clinical impact

The reviewed literature predominantly employs standard classification and regression metrics to evaluate diagnostic and analytic performance:

- Accuracy and Macro-F1 Score: These metrics assess disease prediction, report generation quality, diagnostic triage accuracy, and entity recognition. For example, Health-LLM achieved 83% accuracy and an F1 score of 0.76 in the IMCS-21 multi-turn dialogue diagnosis bench- mark, outperforming traditional and baseline models [1].
- Area Under the Curve (AUC): Widely used in cross- institution EHR benchmarks and risk stratification tasks, multi-task models like Gen HPF exhibit AUROC values between 0.78 and 0.82 across several hospital sites [15].

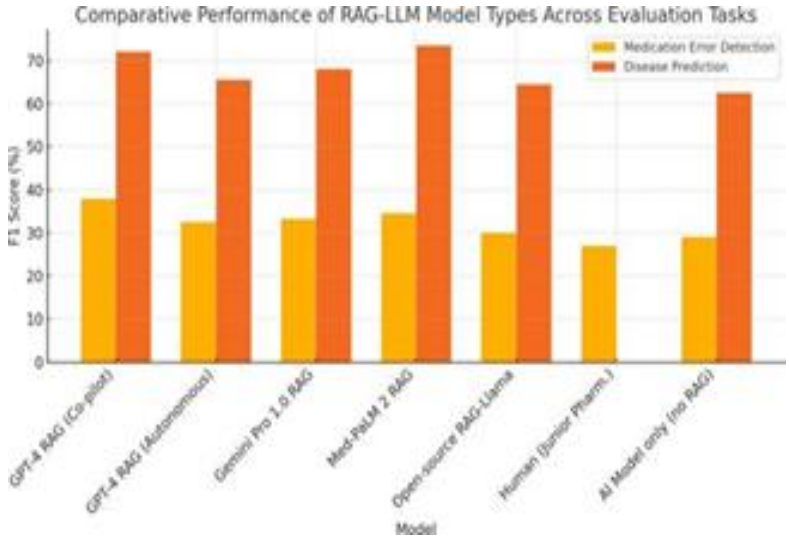

Figure 1: F1 score comparison among various RAG-LLM models on medication error detection and disease prediction tasks

Table 2: Comparative Summary: AI Methodologies, Benchmarks, and Performance

Model Type	Primary Task	Benchmarks	Top Metric	Noted Limitation
GPT-4/Med-PaLM 2	QA/Diagnosis	MedQA, PubMed QA	86.5% Accuracy (USMLE)	Hallucination, bias, computational cost
Health-LLM/RAG	Disease Prediction	IMCS-21, Custom	83.3% Accuracy, F1: 0.76	Standardization, hyperparameter tuning
Gen HPF (MTL)	Multi-task EHR	Multi-hospital datasets	AUROC: 0.82	Data drift, necessity for text encoding
CNN/RNN	Imaging/Sequential	Radiology, Labs	Accuracy: 75–77%	Large labeled data requirement
OCR-Tesseract (NLP)	Prescription Extraction	Prescription datasets	Accuracy: 98%	Handling handwriting variability
Dr Rank (LLM RecSys)	Doctor Ranking	Dr Rank dataset	NDCG@10: 77.41	Coverage bias, subjectivity
K-anonymity, Federated Learning	Privacy	Custom/Open datasets	Privacy-Utility Trade-off	Data loss, computational cost
AI Hospital/Lab	Multi-agent Evaluation	MVME, clinical simulation	Multi-dimensional scores	Realism, language coverage limitations

Evaluation metrics

- NLP metrics (Exact Match, BLEU, ROUGE, Perplexity) assess chatbot QA and generative quality [4].
- NDCG@10 measures ranking relevance and fairness in explainable recommenders like Dr Rank [9].
- Interpretability scores evaluate LLM explanations via clinical audits or human review [19].
- **Clinical Impact:**
- Med-PaLM 2 achieved 86.5% on MedQA (USMLE), with hybrid human-AI teams outperforming individual models by up to 10% [4], [6].
- Gen HPF and multi-source RAG models show better robustness and cross-institutional calibration [15].
- OCR/NLP pipelines achieve ~98% accuracy, reducing manual effort and errors [12].
- Privacy-preserving methods (k-anonymity, federated learning) balance data utility with confidentiality [3], [18].

Model limitations and research gaps

- **Model limitations:** LLMs and multi-agent simulators struggle with ambiguous or atypical clinical scenarios, multi-turn reasoning, and hallucinations. CNN/RNN models need large labeled datasets and often generalize poorly without adaptation. Explainability tools (LIME, SHAP, attention) provide only partial transparency, while OCR/NLP pipelines falter on poor handwriting, rare terms, or degraded images.
- **Clinical and regulatory gaps:** Benchmarks rely on curated datasets or simulations, lacking large-scale real-world validation. Fairness and bias mitigation are limited, and stronger privacy often reduces utility. Human-AI collaboration lacks standardized protocols, complicating accountability and trust.

- **Open research gaps:** Integrating multimodal EHR, imaging, genomics, and text remains challenging. Language and cultural diversity in models is limited, and continuous updating, error correction, and feedback incorporation are underdeveloped, posing safety and obsolescence risks.

Table 3: Summary Comparison: Model, Dataset, Metric, and Best Scores

Model	Architecture	Dataset	Metric	Score
GPT-4 (RAG)	LLM + RAG	MedQA (USMLE)	Accuracy	0.86
Med-PaLM 2	LLM	MedQA, Med M- CQA	Accuracy	0.865
Health- LLM	LLM + Auto ML	IMCS-21	Accuracy / F1	0.83 / 0.76
Gen HPF	Multi-task EHR	Multi-site EHR	AUROC	0.78–0.82
CNN / RNN	Deep Learning	Radiology, EHR	AUC / Accuracy	0.73–0.77
Dr Rank	LLM + XAI	Dr Rank dataset	NDCG@10	0.77
OCR-Tesseract	OCR / NLP	Prescription images	Extraction Accuracy	0.98
AI Hospital	Multi- agent LLM simulation	MVME	Composite	—

Figure 2: Heatmap illustrating AI methodologies versus clinical tasks

Figure 2 illustrating AI methodologies versus clinical tasks. Large language models enhanced with retrieval-augmented generation (RAG) excel at question answering and triage tasks; Gen HPF leads in multi-task EHR prediction; CNNs remain the strongest in medical imaging. Remaining gaps include fairness, longitudinal reasoning, and human–AI collaboration integration.

Summary

While LLMs and hybrid AI frameworks are rapidly advancing toward expert-level performance for key clinical question answering and disease prediction tasks, persistent challenges in explainability, generalizability, fairness, and privacy hinder widespread real-world clinical deployment. Addressing these limitations will require richer

benchmarking standards, multisite prospective trials, dynamic learning protocols, and stronger collaboration among domain experts, AI developers, and regulatory stakeholders.

Comparative tables of techniques, datasets, and results

Table 4 summarizes key AI models and techniques, their benchmark datasets, evaluation metrics, and highest reported performance, providing a convenient format for cross- comparison and highlighting state-of-the-art methods across categories.

Table 4: Summary Comparison: Model, Dataset, Metric, and Best Performance Scores

Model	Architecture	Dataset	Metric	Score
GPT-4 (RAG)	LLM + RAG	MedQA (USMLE)	Accuracy	0.86
Med-PaLM 2	LLM	MedQA, Med M- CQA	Accuracy	0.865
Health- LLM	LLM + Auto ML	IMCS-21	Accuracy / F1	0.83/0.76
Gen HPF	Multi-task EHR	Multi-site EHR	AUROC	0.78–0.82
CNN/RNN	Deep Learning	Radiology, EHR	AUC / Accuracy	0.73–0.77
Dr Rank	LLM + XAI	Dr Rank dataset	NDCG@10	0.77
OCR-Tesseract	OCR / NLP	Prescription images	Extraction Accuracy	0.98
AI Hospital	Multi- agent LLM simulation	MVME	Composite Score	—

Table notes

1. Med-PaLM 2 and Health-LLM top clinical QA, triage, and multi-task benchmarks.
2. Gen HPF excels in multi-source EHR predictions across hospitals.
3. OCR-Tesseract achieves near-perfect prescription data digitization.
4. Dr Rank demonstrates explainable, fair medical recommendations.

Benchmark visualization

1. Radar plot (Figure 3) compares AI methods across QA, diagnosis, EHR analytics, and explainability.
2. LLM and RAG-LLM models lead in question answering and diagnostic tasks.

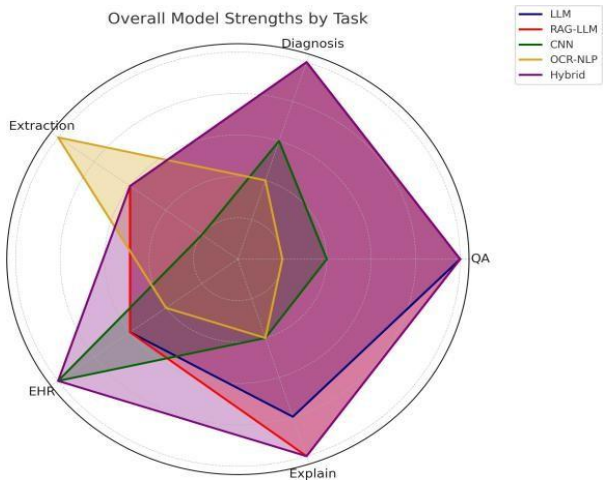

Figure 3: Radar plot of AI method performance across key clinical tasks

- CNN-based architectures dominate imaging-related clinical pipelines.
- OCR-NLP techniques excel in clinical document digitization and information extraction.
- Hybrid AI frameworks, including co-pilot models and multi-agent systems, demonstrate balanced effectiveness across all key clinical domains, reflecting the importance of human-AI collaboration for safe deployment.

Summary

The combined tables and visualizations succinctly portray the relative strengths, limitations, and evolving trends within AI-powered medical diagnosis and consultation, serving as a guide for practitioners and researchers in navigating this complex field.

Visuals and case summary table

Table 5: Deployment Scenarios: Key Features and Lessons

System	Deployment Setting	Integration Features	Lessons Learned
Babylon Health	Live: multi- country telehealth	EHR/telemedicine, rule-based+ LLM triage	Human QA, user training, iterative workflow
Google Med-PaLM	Hospital QA pilots	FHIR/HL7, RAG+LLM, physician-in- the-loop	AI can out per- form doctors on QA, but over- sight critical
AI Hospital	Benchmarking research	Multi-agent sim, workflow orchestration	Reveals LLM gaps, supports co-piloting

Conclusion of case studies

Widespread deployment demonstrates that LLM-powered and hybrid AI consultation systems are capable of matching or closely approaching physician-level performance for key tasks, thus improving access to care and reducing clinician burden. The most effective deployments blend advanced AI with human supervision, strict operational protocols, and sustained clinical and regulatory validation. Continuous attention to usability, transparency, and soliciting user feedback is essential for earning trust and achieving safe, equitable, and impactful healthcare transformation [6], [4].

Future research directions

AI is rapidly reshaping medical diagnosis and consultation, but technical, ethical, and operational challenges remain. Emerging trends include:

1. **Multimodal AI:** Integrating clinical text, imaging, genomics, and wearable data allows richer context, early detection, and personalized diagnostics. Agent-based architectures distribute processing across modalities, enabling cohesive clinical reasoning and precision medicine.
2. **Federated learning and privacy:** Decentralized model training protects patient data while improving generalizability across hospitals and rare or underrepresented populations. Key challenges include data heterogeneity, system interoperability, and regulatory-standardized protocols.
3. **Agentic and continual learning:** AI systems are evolving to actively seek information, learn from new cases, and adapt continuously, maintaining relevance amid changing medical knowledge and emerging diseases. Ensuring safe, stable, and auditable continual learning remains a major frontier.

4. **Visual/spatial AI and explainability:** Combining computer vision, spatial analytics, and LLMs supports real-time, complex clinical workflows (e.g., surgery, patient monitoring). Explainable methods are needed to interpret multi-dimensional data streams and foster clinician trust.

Open challenges

1. **Interpretability and explainability:** LLMs and deep learning models lack inherent transparency, limiting clinical trust and regulatory approval. Future work should embed explainability within AI architectures, communicate reasoning, uncertainty, and actionable rationales, and integrate clinical narratives for better alignment with human decision-making.
2. **Generalization and robustness:** AI models often struggle across diverse institutions, populations, and emerging diseases due to data heterogeneity and distribution shifts. Prospective multi-site validation, transfer learning, domain adaptation, and advanced data augmentation are key to building reliable, generalizable systems.
3. **User Trust and socio-technical alignment:** Trust requires accuracy, transparency, accountability, and alignment with user values. Systems should provide error recourse, customizable controls, and clear risk-benefit communication, while sociotechnical research ensures smooth integration into clinical workflows.
4. **Fairness, ethics, and regulatory compliance:** Bias and inequitable performance across socio-demographic groups remain under-addressed. Standardized fairness metrics, representative datasets, ongoing audits, and proactive regulatory alignment are essential for safe, ethical, and equitable AI deployment.

Conclusion

Artificial intelligence, particularly large language models and multimodal systems, is transforming automated medical diagnosis and consultation. This survey highlighted key architectures—including LLMs, retrieval-augmented methods, deep learning, hybrid multi-agent systems, and predictive EHR analytics—along with advances in explainability and privacy. While AI models often approach clinician-level performance on benchmarks, challenges in interpretability, robustness, trust, and large-scale validation remain. Effective deployment envisions AI as an intelligent co-pilot, supporting clinicians with actionable insights. Future progress depends on multimodal integration, federated and continual learning, privacy-focused design, and ethical, regulatory alignment, enabling safe, equitable, and patient-centered healthcare.

References

1. Yu, Q., Jin, M., Shu, D., Zhang, C., Fan, L., Hua, W., Zhu, S., Meng, Y., Wang, Z., Du, M., & Zhang, Y. (2025). *Health-LLM: Personalized retrieval-augmented disease prediction system*. arXiv. https://arxiv.org/abs/2402.00746
2. Singhal, K., *et al.* (2023). *Towards expert-level medical question answering with large language models*. arXiv. https://arxiv.org/abs/2305.09617
3. Fan, Z., Tang, J., Chen, W., Wang, S., Wei, Z., Xie, J., Huang, F., & Zhou, J. (2024). *AI hospital: Benchmarking large language models in a multi-agent medical interaction simulator*. arXiv. https://arxiv.org/abs/2402.09742
4. Pahune, S., & Rewatkar, N. (2023). Large language models and generative AI's expanding role in healthcare. *International Journal of Research in Applied Science and Engineering Technology, 11*(8), 2288–2302.

5. AMIE Consortium. (2025). Towards conversational diagnostic artificial intelligence. *Nature*.
6. Maity, S., *et al.* (2025). Large language models in healthcare and medical applications. https://pmc.ncbi.nlm.nih.gov/articles/PMC12189880/
7. Alia, H., Qadir, J., Alam, T., House, M., & Shah, Z. (2023). ChatGPT and large language models in healthcare: Opportunities and risks. *International Journal of Medical Informatics* (preprint).
8. Karagiannis, S., Ntantogian, C., Magkos, E., Tsohou, A., & Ribeiro, L. L. (2024). Mastering data privacy: Leveraging k-anonymity for robust health data sharing. *International Journal of Information Security, 23*(8), 2189–2201.
9. Touvron, H., Lavril, G., *et al.* (2023). *LLaMA 2: Open foundation and fine-tuned chat models*. arXiv. https://arxiv.org/abs/2307.09288
10. Zhou, S., *et al.* (2025). Large language models for disease diagnosis: A scoping review. *Nature Medicine*.
11. Alkhanbouli, R., *et al.* (2025). The role of explainable artificial intelligence in disease prediction: Systematic review.
12. Al Siam, A., & Shohan, S. (2025). *Privacy-preserving AI for encrypted medical imaging: A framework for secure diagnosis and learning*. arXiv. https://arxiv.org/abs/2507.2106