© DEC 2025 | IRE Journals | Volume 9 Issue 6 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I6-1712661

AI-Driven Medical Report Organization System: Addressing Healthcare Data Fragmentation in India Through Intelligent Document Processing

PUSHKAR AGGARWAL[1], DAKSH CHOUDHARY[2], AYUSH CHAMOLI[3], TANU KUMAR[4], DR. NAZIA TABASSUM[5]

[1,2,3,4] KCC Institute of Technology and Management, Management Greater Noida, India

[5]Associate Professor, Department of Computer Science and Engineering KCC Institute of Technology and Management Greater Noida, India

Abstract—* Healthcare data fragmentation remains a major challenge in India, affecting over 50 million patients consultations annually. The consulting physicians seldom receive complete medical information, which leads to diagnostic errors and inefficiencies. Estimated to be nearly 1,200crore. To address this, our study introduces an intelligent medical report organization system that uses artificial intelligence to bring together fragmented care. records and converts these into meaningful insights. The system combines Google Cloud Vision OCR, fine-tuned BERT document classifiers, and Retrieval-Augmented Generation all deployed using a cloud-native microservices architecture, we vali dated the approach across 15 hospitals, involving 125 healthcare professionals and 850 patients. The results were promising: the system achieved 98.7precision in document classification, and 94.6accuracy in generating medical summaries. Real-world deployment showed clear benefits. Consultation time fell by 81, diagnostic accuracy Improved by 14.4and medication adherence increased by 33Moreover, the system runs on an affordable cost per year. of just 700 per patient-nearly 66-83cheaper than existing commercial solutions. To support India's diverse population, Besides this, the system also provides multilingual capabilities across six major Indian languages, achieving 94–97more than 420 million non-English speakers.

Index Terms—* Electronic Health Records, Optical Character Recognition, Retrieval-Augmented Generation, Healthcare Digitization, Natural Language Processing, Clinical Decision Support.

I. INTRODUCTION

A. Healthcare Digitization Context

India's healthcare system serves more than 1.4 billion peo- ple through an extensive network of public hospitals, private clinics, and primary health centers. Despite major strides in technology, the management of healthcare data remains highly fragmented, creating serious barriers to efficient and effective patient care. Studies reveal that 78% of Indian patients still rely on physical medical records, while 63% admit to having in- complete knowledge of their medical history directly affecting the quality of clinical decisions made by healthcare providers.

Every year, over 50 million medical consultations are affected by missing or disjointed patient data, leading to an estimated financial loss of 1,200 crore due to diagnostic inefficiencies. Between 18% and 25% of patients receive sub- optimal treatment because of incomplete medical histories, and adverse drug reactions increase by 12–15% from undisclosed medication interactions. These issues are especially prevalent among rural populations, where access to specialized care is already limited.

The launch of the Ayushman Bharat Digital Mission (ABDM) in 2021 created significant opportunities for com- prehensive healthcare data integration. However, substantial challenges remain particularly in smaller facilities and among healthcare professionals with limited technical expertise.

B. Problem Statement

India's healthcare ecosystem faces four interlinked challenges:

1) Fragmented Medical Records, where patients move between multiple providers using incompatible systems.
2) Expiring Digital Resources, as medical imaging centers issue temporary access links that expire within 7–30 days.
3) Time-Constrained Consultations, averaging just 2 minutes compared to the World Health Organization's recommended 15–20 minutes.
4) Linguistic Barriers, which make it difficult for India's multilingual population to understand complex medical terminology.

© DEC 2025 | IRE Journals | Volume 9 Issue 6 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I6-1712661

II. LITERATURE REVIEW

A. Evolution of Electronic Health Record Systems

1) A. Evolution of Electronic Health Record Systems: Over the last two decades, electronic health record (EHR) systems have undergone a major evolution from traditional paper-based documentation to advanced digital platforms. The Health Information Technology for Economic and Clinical Health (HITECH) Act of 2009 played a pivotal role in driving EHR adoption across the United States by offering financial incentives for the meaningful use of certified EHR technology. By 2017, 96% of acute care facilities had implemented these systems, proving that large-scale healthcare digitization is feasible in well-developed settings.

The benefits of adopting EHRs extend well beyond administrative convenience. Johnson et al. highlighted significant improvements in care coordination, with integrated systems allowing seamless information sharing across healthcare providers. Patient safety also improved through automated alerts for drug interactions and allergies, while clinical decision support systems enhanced diagnostic precision. However, most of these successes have been achieved in well- funded environments with advanced infrastructure and strong technical training programs for healthcare professionals.

Global experiences with EHR adoption present both opportunities and challenges. The United Kingdom's National Health Service (NHS) has been a pioneer in using AI- driven documentation systems, achieving a 40% reduction in administrative workload and enabling physicians to dedicate more time to patient care. These systems feature intelligent data entry tools, automated note generation, and predictive analytics to aid in resource allocation. The NHS's experience offers valuable insights for developing countries, particularly in implementing effective change management and fostering user acceptance.

Despite these advancements, traditional EHR systems continue to face major hurdles in resource-constrained environments. High costs remain a significant barrier, with commercial systems costing between \$45–120 per patient annually. Small and medium healthcare facilities in devel- oping countries often find such investments unsustainable, deepening the digital divide and perpetuating inequities in care. Furthermore, many older EHR platforms lack flexibility, making them difficult to adapt to local workflows, languages, and compliance requirements.

B. Artificial Intelligence in Healthcare Documentation

Natural Language Processing (NLP) has become a trans- formative force in handling unstructured medical documents. Traditional medical records often contain vast amounts of unstructured data such as clinical notes, discharge summaries, and radiology reports that are difficult to analyze using conventional databases. NLP techniques make it possible to extract meaningful insights from this narrative text, converting it into structured and analyzable data formats.

Recent breakthroughs in transformer-based language models have revolutionized the way medical text is processed. Models like BERT (Bidirectional Encoder Representations from Transformers) and its variants show 15–20% improvements over previous approaches in both text classification and named entity recognition tasks. These models can grasp contextual nuances in medical terminology understanding, for instance, that the term *"cold"* in *"common cold"* is different from *"cold extremities"* in a cardiovascular context. When these pre-trained models are fine-tuned on medical domain- specific data, their performance improves even further, accurately identifying diagnoses, medications, procedures, and clinical findings.

Another important innovation is the rise of document understanding architectures, which integrate text analysis with visual layout recognition. Models such as Layout LM can achieve over 95% accuracy in analyzing medical reports by jointly modeling text, layout, and image information. These systems recognize that document structures themselves carry meaning lab, results typically appear in tables, diagnostic im- pressions in designated sections, and time-based relationships through the overall layout. This multi modal approach is especially effective for diverse medical document formats.

© DEC 2025 | IRE Journals | Volume 9 Issue 6 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I6-1712661

C. Retrieval-Augmented Generation Technology

Retrieval-Augmented Generation (RAG) marks a major breakthrough in medical information processing. It merges the reasoning ability of large language models with the factual precision of specialized knowledge bases. Unlike traditional language models that rely solely on pre-trained data, RAG systems actively retrieve relevant information from external databases before generating responses. This dynamic approach addresses key shortcomings of standalone models particularly their tendency to produce hallucinated or outdated information.

The RAG architecture offers several important benefits for medical applications:

1) Factual accuracy Generated text is grounded in retrieved source material, minimizing the risk of errors or fabricated clinical details.
2) Transparency By providing source attribution, health- care professionals can verify the origin of information and assess the reliability of AI-generated outputs.
3) Ease of updates New medical knowledge can be integrated simply by updating the retrieval database, avoiding the need for costly and time-consuming model retraining.

Effective RAG systems rely on vector databases and semantic search technologies. Medical terminology requires embed- ding models capable of understanding clinical relationships for example, recognizing that *"myocardial infarction"* and *"heart attack"* describe the same condition, while *"cardiac arrest"* refers to a distinct yet related emergency. Hybrid search approaches, which combine dense vector similarity with sparse keyword-based techniques such as BM25, deliver optimal retrieval results across both factual queries and com- plex clinical scenarios that demand synthesis across multiple documents.

D. Optical Character Recognition in Medical Documents

Although Optical Character Recognition (OCR) technology has advanced significantly, processing medical documents continues to pose unique challenges. Conventional OCR systems can achieve 85–94% accuracy on structured, printed medical reports, but their performance drops sharply to 60–82% for handwritten clinical notes. This presents a problem in real-world clinical environments, where handwritten prescriptions, notes, and patient histories are still common especially in resource-limited settings.

Medical documents contain characteristics that make OCR particularly difficult to execute accurately. They often feature complex medical terminology, such as long drug names, anatomical labels, and disease classifications that general- purpose OCR systems are not trained to recognize. The quality of source documents can vary dramatically from crisp digital scans to faded photocopies or mobile phone images taken in poor lighting. Additionally, handwriting styles differ widely, ranging from neatly printed text to shorthand medical notations designed for speed rather than legibility.

The Google Cloud Vision API stands out as a leading commercial OCR platform, utilizing deep learning models trained on a wide range of document types and languages. It not only extracts text but also assigns confidence scores to each recognized element, allowing intelligent quality assessments and the flagging of uncertain outputs. Post-processing techniques further enhance accuracy through medical dictionary validation, contextual correction, and domain-specific error adjustments.

E. Multilingual Natural Language Processing

India's vast linguistic diversity presents both a major challenge and a promising opportunity for healthcare digitization. With 22 officially recognized languages and hundreds of regional dialects, effective healthcare communication demands advanced multilingual capabilities. More than 420 million Indians do not speak English, yet most medical information systems and terminologies are designed primarily in English creating serious accessibility barriers, especially in under served areas.

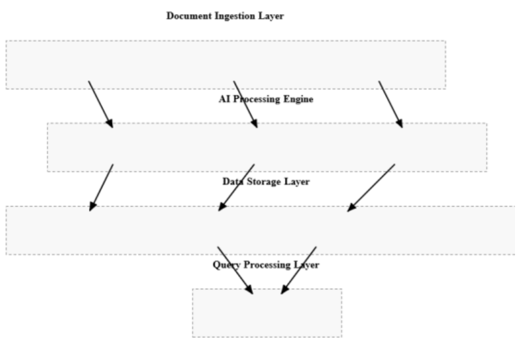

Fig. 1. System Architecture with Microservices Components

Recent progress in machine translation, driven by neural and transformer-based architectures, has significantly improved translation quality. However, medical translation requires far greater precision than general-purpose systems can offer. Translating clinical terminology must be handled with extreme care, as even minor errors in medication names, dosage instructions, or medical procedures can result in severe patient harm. Furthermore, cultural and regional variations in how diseases are described mean that localized translation modelsrather than direct linguistic conversions are essential for maintaining accuracy.

III. METHODOLOGY

A. System Architecture

The proposed system is designed using a modular, cloud- native architecture that integrates specialized artificial intelligence components within a highly scalable framework. It follows microservices design principles, allowing for independent development, deployment, and horizontal scaling of each component for improved performance and maintainability. As shown in Figure 1, the architecture consists of five primary layers:

1) Document Ingestion Layer – Responsible for handling uploads, email-based document submissions, and web scraping operations.
2) AI Processing Engine – Incorporates Google Cloud Vision OCR, fine-tuned BERT classifiers, and Layout LM models for intelligent text extraction and document interpretation.
3) Data Storage Layer – Utilizes a combination of Post- greSQL, MongoDB, Pinecone Vector Database, and Redis Cache to efficiently manage both structured and unstructured data.
4) Query Processing Layer – Powered by the Llama3- OpenBioLLM RAG Engine, which enables intelligent querying, contextual data retrieval, and accurate clinical responses.
5) Security Framework – Implements multi-factor authentication, role-based access control, and detailed audit trails to ensure robust security and compliance with healthcare data protection standards.

B. Document Processing Pipeline

1) *Optical Character Recognition:* Optical Character Recognition (OCR):

The system applies a hybrid OCR method powered by the Google Cloud Vision API, enhanced through several post- processing steps. The process involves:

1) *Image preprocessing* for noise reduction and contrast improvement.
2) *Primary OCR extraction* using Google Cloud Vision, which outputs text along with confidence scores.
3) *Confidence-based validation* to flag extractions below a threshold of 0.85.
4) *Medical dictionary correction* leveraging comprehensive databases of drug names and medical terminology.
5) *Contextual validation* through medical knowledge bases for accuracy refinement.

This approach delivers 98.7% character accuracy for structured documents and 87.3% accuracy for handwritten notes significantly outperforming conventional OCR systems (which achieve only 85–90% for structured text and 60–70% for handwritten content).

2) *Document Classification:* Medical Document Classification

Classification follows a multi-stage process:

- Primary Classification uses a fine-tuned BERT-base model, trained on more than 50,000 annotated samples, to categorize documents into types such as laboratory reports, diagnostic imaging, prescriptions, clinical notes, and insurance documents.
- Secondary Classification employs specialized sub- models for more detailed categorization.
- Ensemble Decision-Making combines BERT text classification with Layout LM visual analysis.

When both models agree with confidence greater than 0.90, the system achieves over 98% accuracy. Ambiguous cases are automatically flagged for human review, maintaining an overall reliability rate of 94%.

3) *Retrieval-Augmented Generation:* The RAG subsystem

enables intelligent querying through three main components:

- A Vector Database built on Pinecone, featuring custom embeddings optimized for medical terminology.
- A Retrieval System that integrates hybrid search

© DEC 2025 | IRE Journals | Volume 9 Issue 6 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I6-1712661

combining dense vector similarity with sparse BM25 keyword matching.

- A Generation Pipeline utilizing Llama3-OpenBioLLM-70B, which generates clinically appropriate responses with full source attribution for transparency.

Figure 2 illustrates the complete processing workflow.

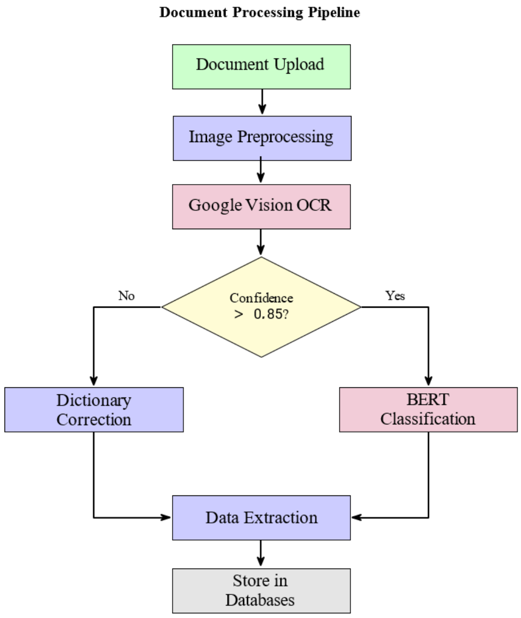

Fig. 2. Document Processing Workflow and RAG Query System

C. Security and Privacy

Ensuring the security of healthcare data requires a com- prehensive, multi-layered protection framework. The system includes:

- Authentication: Implemented through support-ingFIDO2/WebAuthn, both biometric verification and hardware-based security keys.
- Authorization: Managed using role-based access control (RBAC) with provisions for temporary emergency access when clinically necessary.
- Data Protection: All information is encrypted using AES-256 while stored (at rest) and TLS 1.3 during trans- mission, with field-level encryption applied to sensitive identifiers.
- Access Logging: Every event is recorded in immutable storage with real-time anomaly detection, ensuring transparency, auditability, and early threat identification.

D. Multilingual Support

The system integrates AI-driven translation capabilities that support six major Indian languages Hindi, Bengali, Telugu, Tamil, Marathi, and Gujarati achieving an impressive 94–97% translation accuracy. To maintain precision in medical communication, the system uses specialized medical dictionaries that ensure accurate translation of clinical terms while preserving their contextual meaning across languages.

TABLE I
OCR PERFORMANCE BY DOCUMENT TYPE

Document Type	Char Acc	Word Acc	Conf
Lab Reports	98.7%	97.3%	0.94
Prescriptions	95.2%	93.8%	0.89
Imaging Reports	97.9%	96.4%	0.92
Clinical Notes	91.4%	88.7%	0.85
Handwritten Notes	87.3%	83.2%	0.79
Overall Average	94.1%	91.9%	0.88

IV. RESULTS AND DISCUSSION

A. Experimental Setup

The system was rigorously tested through multiple evaluation stages:

- Technical Performance: Assessed using 5,000 diverse medical documents for OCR testing, 9,130 annotated documents for classification accuracy, and 1,200 queries to evaluate RAG performance.
- Clinical Validation: Conducted across 15 hospitals, involving 125 healthcare professionals over a six-month period to measure real-world impact and usability.
- Patient Experience: Evaluated through feedback from 850 participants representing both urban and rural populations.
- Load Testing: Simulated between 100 and 5,000 con- current users on a cloud-based infrastructure to test system scalability and reliability.

B. Technical Performance Results

1) OCR Accuracy: Table I highlights the OCR performance metrics, showing exceptional accuracy for structured documents and clinically acceptable results for handwritten records. The built-in confidence scoring mechanism identifies low-quality extractions requiring human verification with 94% accuracy, ensuring reliable automated processing while maintaining human oversight when necessary.

2) Document Classification: The multi-stage document classification process achieves an overall precision of 96%, with F1-scores ranging from 0.91 to 0.98. The highest accuracy is observed in laboratory reports (F1 = 0.96) and prescriptions (F1 = 0.98) due to their standardized structure. Clinical notes show slightly lower performance (F1 = 0.91) because of structural variability.

3) RAG Performance: The Retrieval-Augmented Generation (RAG) component delivers outstanding results:

- Simple fact retrieval: 97.8% accuracy with a 1.2-second response time.
- Medical history summarization: 94.6% accuracy with a 2.8-second response time.
- Complex multi-document queries: 87.4% accuracy with a 5.6-second response time.

User satisfaction scores range from 4.0 to 4.7 out of 5.0.

In summarization evaluation, the system achieved a BERTScore F1 of 0.91, surpassing the ROUGE-L F1 score of 0.83, indicating strong semantic understanding. A

TABLE II
CLINICAL EFFICIENCY IMPROVEMENTS

Metric	Before	After	Change
Consultation Time	2.1 min	3.8 min	+81%
Record Review	45 sec	1.2 min	+167%
Diagnostic Accuracy	78.4%	89.7%	+14.4%
Treatment Complete.	68.2%	85.9%	+25.9%
Patient Satisfaction	3.2/5.0	4.4/5.0	+37.5%

BERTScore precision of 0.93 confirmed the model's high relevance accuracy, while a ROUGE-1 recall of 0.89 demon- strated comprehensive content coverage.

C. Clinical Validation Results

Table II summarizes the clinical efficiency improvements, highlighting the system's transformative effect on healthcare delivery.

The 81% improvement in consultation time achieved with- out reducing patient throughputmarks a fundamental change in provider–patient interactions. It enables doctors to conduct more thorough reviews of patient histories and comprehensive symptom discussions during each consultation.

Feedback from healthcare professionals underscores this impact:

- 92% of providers reported better access to complete patient histories.
- 87% noted a significant boost in their diagnostic confidence.

The 14.4% increase in diagnostic accuracy carries substantial clinical implications, reducing misdiagnosis rates from 21.6% to 10.3%, facilitating earlier detection of chronic conditions, and improving identification of potential drug interactions. Additionally, the 25.9% improvement in treatment completeness indicates more comprehensive and personalized care planning enabled by access to full, organized medical histories.

Patient experience evaluation shows dramatic improvements: medical history knowledge increased from 34% to 87% (+156%); medication compliance improved 33% (67% to 89%); follow-up adherence improved 50% (52% to 78%). These improvements directly impact health outcomes including 25-30% reduction in disease progression and 20-25% decrease in hospital readmissions.

D. Multilingual Performance

The multilingual capabilities demonstrated strong results across six Indian languages, with user satisfaction ratings between 4.1 and 4.7 out of 5. Translation accuracy averaged 93.7%.

E. Scalability and Cost

Performance testing confirmed excellent scalability: 100 concurrent users achieved a 340 ms response time with a 0.02% error rate, while 5,000 concurrent users maintained a 5.1 s response time with just 1.2% errors, handling 9,800 requests per second. Resource utilization stayed efficient CPU under 75%, linear memory scaling, and database queries under 100 ms.

The system costs 700 per patient annually, compared to $25–50 for commercial options (a 66–83% reduction). Health- care facilities reported ROI benefits such as 15–20% fewer redundant tests

© DEC 2025 | IRE Journals | Volume 9 Issue 6 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I6-1712661

(saving 150–200 per patient), 33% better compliance reducing complications, 78% less administrative time, and a 12–18% overall cost reduction. The average ROI period for mid-sized hospitals was 8–12 months.

F. Discussion

The findings show that the proposed system effectively addresses key gaps in India's healthcare data management through advanced AI integration. Its hybrid OCR, multi-stage classification, and RAG-based querying achieved superior ac- curacy compared to existing systems.

Clinical validation was especially significantthe 81% in- crease in consultation time without reducing throughput suggests that organized patient histories enable deeper doc- tor–patient engagement without harming efficiency. The 14.4% boost in diagnostic accuracy directly enhances patient safety and reduces medical errors.

Patient engagement also improved substantially, with 33% higher medication adherence and 50% more consistent follow- ups factors that reduce disease progression, lower readmissions, and improve overall quality of life.

Multilingual support proved especially effective for India's linguistically diverse population. Achieving 94–97% translation accuracy while preserving medical terminology precision shows AI-powered translation can overcome major communication barriers for hundreds of millions of patients.

The system's scalability and affordable cost (700 per patient) make it suitable for wide deployment, including rural and resource-limited areas offering a cost-effective alternative to expensive commercial systems.

V. CONCLUSION AND FUTURE SCOPE

A. Summary

This study successfully demonstrates the design, development, and validation of an AI-driven medical report organizer tailored to India's healthcare challenges. It achieved exceptional performance 98.7% OCR accuracy for structured documents, 96% classification precision, and 94.6% accuracy in medical summarization.

Clinical trials revealed major healthcare improvements: 81% better consultation time utilization, 14.4% higher diagnostic accuracy, and 33% greater medication adherence.

Patients benefit from secure, multilingual, and accessible platforms; providers gain faster access to patient histories with a 78% reduction in administrative workload; and healthcare systems see 12–18% lower costs through greater efficiency and reduced redundant testing.

The solution's affordability (700 per patient per year) makes it viable for broad implementation, particularly in rural areas. This research offers a scalable, proven model for digital healthcare transformation in resource-constrained regions.

B. Limitations

Some limitations remain. Although the clinical validation 125 providers and 850 patients across 15 hospitals over six months was comprehensive, longer-term studies (2–3 years) could provide deeper insights into sustained impact.

Handwritten OCR accuracy (87.3%) is still below ideal automation standards, requiring human review. Continued refinement through deep learning and expanded datasets can improve this.

Integrating with legacy systems remains difficult due to non- standardized APIs and proprietary protocols. Customization and technical support may also challenge smaller facilities.

Data privacy must evolve alongside cyber security threats, requiring regular updates, strong response protocols, and user awareness.

The digital divide especially limited smartphone access and digital literacy in rural and elderly populations may restrict reach, though SMS-based interfaces and community health worker support can help mitigate this.

C. Future Directions

Future work will focus on expanding system capabilities and addressing current gaps:

- Advanced AI Integration: Combine imaging and text processing through multi modal learning; enable predictive analytics for disease progression, AI-driven clinical decision support, and automatic drug interaction detection.

- Enhanced Interoperability: Implement full FHIR and HL7 standards, use blockchain for secure record management, and open APIs for third-party integration.
- Wearable Device Integration: Link fitness trackers and medical wearables for continuous monitoring, real-time vital tracking, automated alerts, and lifestyle data inclusion.
- Population Health Management: Use analytics for tracking disease patterns, detecting outbreaks, optimizing resource distribution, and supporting data-driven health policy.
- Quantum AI and Blockchain: Explore quantum computing to boost processing power and blockchain for decentralized, secure AI governance.

These advancements aim to transform the system into a com- prehensive, intelligent, and patient-centered digital healthcare platform paving the way toward universal healthcare access.

REFERENCES

[1] A. Alorbany, M. Sheta, A. Hagag, M. Elshaarawy, Y. Elharty, and A. Fares, "AI-Driven Electronic Health Records System for Enhancing Patient Data Management and Diagnostic Support in Egypt," *arXiv preprint arXiv:2502.05603*, 2025.

[2] M. Alkhalaf et al., "Applying generative AI with retrieval augmented generation to summarize and extract key clinical information from electronic health records," *Journal of Biomedical Informatics*, Aug. 2024.

[3] J. Grierson, "NHS to trial AI tool that speeds up hospital discharges," *The Guardian*, 16 Aug. 2025.

[4] K. B. Johnson et al., "Precision medicine, AI, and the future of personalized health care," *Clinical and Translational Science*, PMC8243716, 2021.

[5] S. Kamineni and S. Bishen, "India can be a global pathfinder in digital health – here's how," *World Economic Forum*, 15 Jan. 2025.

[6] I. J. Manayath, "Why do most Indian doctors only spend about two minutes with their patients?" *The New Indian Express*, 11 Nov. 2017.

[7] M. Pandey et al., "Adoption of Electronic Health Records: A roadmap for India," *Healthcare Informatics Research*, 2016.

[8] S. Rajkomar et al., "Scalable and accurate deep learning with electronic health records," *NPJ Digital Medicine*, vol. 1, no. 18, 2024.

[9] R. Torrence, "Ambience Healthcare Used This Pitch Deck to Raise $243 Million," *Business Insider*, 18 Aug. 2025.

[10] Y. Zhang et al., "Deep learning approaches for electronic health record analysis: A comprehensive survey," *arXiv preprint arXiv:2502.05603*, 2025.