Acta Scientific MEDICAL SCIENCES (ISSN: 2582-0931)

Volume 10 Issue 7 July 2026

Comprehensive Review

Artificial Intelligence in Human Health: A Comprehensive Review

Manabendra Debnath[1]* and Biplab De[2]

[1]Department of Human Physiology, Kabi Nazrul Mahavidyalaya, Sonamura, Sepahijala, Tripura 799131, India

[2]Regional Institute of Pharmaceutical Science and Technology, Abhoynagar, Agartala, Tripura 799005, India

***Corresponding Author:** Manabendra Debnath, Department of Human Physiology, Kabi Nazrul Mahavidyalaya, Sonamura, Sepahijala, Tripura 799131, India.

DOI: 10.31080/ASMS.2026.10.2260

Received: April 27, 2026

Published: July 06, 2026

Abstract

Background and Purpose: Artificial intelligence (AI) has emerged as a transformative force in modern healthcare, fundamentally reshaping diagnostics, therapeutics, drug discovery, personalised medicine and health systems management. This comprehensive review critically analyses current evidence on the application of AI technologies — including machine learning (ML), deep learning (DL), natural language processing (NLP), and computer vision — across multiple domains of human health.

Methods: A systematic narrative review was conducted across PubMed/MEDLINE, Scopus, Web of Science, IEEE Xplore, and arXiv. Peer-reviewed literature published between 2015 and 2025 was considered. Studies were selected based on methodological rigour, clinical relevance, dataset size and validation quality.

Results: AI demonstrates remarkable and increasingly validated performance in medical imaging (lung cancer CT: AUC 0.944; diabetic retinopathy: sensitivity 97.5%), clinical decision support (EHR mortality prediction: AUC 0.83–0.95), drug discovery (AlphaFold2; halicin) and mental health monitoring. Wearable AI algorithms detect atrial fibrillation at population scale. However, significant challenges persist: algorithmic bias systematically disadvantages minority populations; the majority of published AI studies lack external validation; regulatory frameworks lag technological development and clinical workflow integration remains poorly realized.

Conclusions: While AI demonstrates transformative potential in human health, realizing this potential safely and equitably requires rigorous prospective validation, bias mitigation frameworks, fit-for-purpose regulatory pathways and genuinely integrated clinical deployment. The future of AI in medicine lies in multimodal foundation models augmenting — not replacing — human clinical judgement.

Keywords: Artificial Intelligence; Machine Learning; Deep Learning; Medical Imaging; Algorithmic Bias; Electronic Health Records

Abbreviations

AI: Artificial Intelligence; ML: Machine Learning; DL: Deep Learning; NLP: Natural Language Processing; CNN: Convolutional Neural Network; EHR: Electronic Health Record; ICU: Intensive Care Unit; LLM: Large Language Model; GNN: Graph Neural Network; RL: Reinforcement Learning; GAN: Generative Adversarial Network; AUC: Area Under the Receiver Operating Characteristic Curve; SVM: Support Vector Machine; LSTM: Long Short-Term Memory; GRU: Gated Recurrent Unit; GDT: Global Distance Test; PPV: Positive Predictive Value; MIL: Multiple Instance Learning; WSI: Whole-Slide Image; LMIC: Low- and Middle-Income Country; HIC: High-Income Country; FDA: Food and Drug Administration; HIPAA: Health Insurance Portability and Accountability Act; GDPR: General Data Protection Regulation; SaMD: Software as a Medical Device; CDSS: Clinical Decision Support System; FEP: Free Energy Perturbation; ADMET: Absorption, Distribution, Metabolism, Excretion, Toxicity; TRIPOD-AI: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (Adapted for AI).

Introduction

The intersection of artificial intelligence (AI) and human health is rapidly becoming one of the most significant developments in modern medicine. Over the past decade, dramatic advances in computing power, the expansion of large biomedical datasets, and the evolution of sophisticated algorithms have enabled AI systems to carry out tasks once thought to require exclusively human expertise. Today, AI can help detect cancer in medical images, anticipate patient deterioration in intensive care units, accelerate the discovery of new drugs, and tailor cancer treatments to an individual's genetic profile [1]. These technological advances are unfolding at a time when healthcare systems worldwide are under growing strain — facing aging populations, rising rates of chronic disease, workforce shortages, and persistent disparities in access to care. Together, these pressures have intensified interest in AI as a potentially transformative force in medicine [2].

Yet the integration of AI into clinical practice has not been simple, nor has it been universally embraced. While AI offers powerful tools to address some of healthcare's most pressing challenges, its real-world implementation has raised important questions. Concerns about algorithmic transparency, data privacy, regulatory oversight, bias, and the appropriate balance between human clinical judgment and machine-generated recommendations remain at the forefront of debate [3]. In addition, there is often a substantial gap between what AI systems achieve in carefully controlled research settings and how they perform in everyday clinical environments — a gap that is frequently underestimated.

This review seeks to provide a clear, critical, and balanced assessment of the current state of AI in human health, organized around its major areas of clinical application. It is written for a diverse readership including experts in clinical informatics, biomedical engineering, computational medicine, and health policy. By combining technical evaluation with clinical and policy perspectives, the review aims to highlight both the meaningful progress that justifies cautious optimism and the complex challenges that require ongoing attention from researchers, clinicians, policymakers, and patient communities.

Review Methods

To capture the breadth and complexity of this field, we adopted a systematic narrative review methodology suited to synthesizing evidence across multiple clinical and technical domains. Literature searches were conducted in major biomedical and technical databases, including PubMed/MEDLINE, Scopus, Web of Science, IEEE Xplore, the Cochrane Library, and arXiv. These searches were complemented by manual review of reference lists from relevant articles, as well as focused searches within leading journals including Nature Medicine, The Lancet Digital Health, npj Digital Medicine, JAMA, New England Journal of Medicine, Nature, Cell, and Artificial Intelligence in Medicine.

Search strategies combined key terms related to artificial intelligence — such as "artificial intelligence", "machine learning", "deep learning," and "neural network" — with terms relevant to healthcare applications, including "health", "medicine", "clinical", "diagnostic,""therapeutic", "drug discovery," and "genomic." Boolean operators and Medical Subject Headings (MeSH) were used to tailor and refine searches within each thematic domain. The primary time frame covered publications from 2015 to 2025, reflecting the period of most rapid advancement in clinical AI; however, earlier foundational methodological studies were included where essential for conceptual clarity.

Studies were eligible for inclusion if they were peer-reviewed primary investigations, systematic reviews, or meta-analyses; employed validated AI methodologies; reported clear and

interpretable performance metrics; and demonstrated direct clinical relevance to the domains examined in this review. We excluded studies that lacked sufficient methodological transparency, reported only in vitro or animal findings without clinical data, or were not available in English.

Where applicable, study quality and reporting rigour were assessed using criteria aligned with the TRIPOD-AI reporting framework to ensure consistency, transparency, and methodological soundness across included studies.

Foundational AI technologies in healthcare

The AI technologies most relevant to healthcare can be organized into several broad categories, each with distinct strengths, limitations, and clinical application profiles. Table 1 provides a structured taxonomy of these methods.

AI Category	Key Architecture/ Method	Strengths	Limitations	Primary Healthcare Use Cases
Supervised ML	Random Forest, SVM, XGBoost	Interpretable; handles tabular data; robust on smaller datasets	Requires labelled data; limited on raw images/text	Risk stratification; EHR phenotyping; triage scoring
Deep Learning (CNN)	ResNet, VGG, EfficientNet	State-of-the-art image analysis; scalable with large datasets	Black box; needs large annotated datasets; computationally intensive	Radiology, pathology, dermatology, ophthalmology
Recurrent NN/ LSTM	LSTM, GRU, Bi-LSTM	Sequential temporal data modelling; captures time-dependencies	Vanishing gradients; slower than transformers for long sequences	Vital sign monitoring; EHR time-series; ICU prediction
Transformer/LLM	BERT, GPT-4, Med-PaLM, BioGPT	Exceptional NLP; clinical note understanding; multimodal capacity	Hallucination risk; high compute cost; regulatory uncertainty	Clinical decision support; medical Q&A; EHR summarisation
Graph Neural Network (GNN)	GCN, GAT, GraphSAGE	Relational data modelling; drug-target interaction; knowledge graphs	Scalability challenges; heterogeneous graph construction	Drug discovery; sepsis prediction; genomic pathway analysis
Reinforcement Learning	Deep Q-Network, PPO, A3C	Optimal sequential decision-making; adaptive treatment strategies	Sample inefficiency; reward function design; safety constraints	Dosing optimisation; robotic surgery; clinical trial design
Generative AI (GAN/VAE/Diffusion)	DALL-E, Stable Diffusion, cGAN	Synthetic data generation; data augmentation; molecular design	Mode collapse; quality/fidelity trade-offs; deepfake risk	Synthetic patient data; medical image augmentation; de novo drug design
NLP/Information Extraction	Named entity recognition, relation extraction	Unlocks unstructured clinical text at scale	Abbreviations; clinical jargon; language variability	Clinical note mining; adverse event detection; literature surveillance

Table 1: Taxonomy of AI Methods Applied in Healthcare: Architecture, Strengths, Limitations and Clinical Applications.

CNN = Convolutional neural network; SVM = Support vector machine; LSTM = Long short-term memory; GRU = Gated recurrent unit; LLM = Large language model; NLP = Natural language processing; GNN = Graph neural network; RL = Reinforcement learning; GAN = Generative adversarial network; VAE = Variational autoencoder; EHR = Electronic health record; ICU = Intensive care unit.

Machine learning and supervised learning

Machine learning refers to a broad class of algorithms that learn patterns from data in order to make predictions or decisions, rather than being explicitly programmed to perform a specific task [4]. Among its various approaches, supervised learning — where models are trained using labelled examples — has been especially influential in clinical settings. It has proven particularly effective in diagnostic applications, where large, well-annotated datasets of medical images or electronic health records are available for training [5].

Importantly, machine learning in healthcare is not limited to deep neural networks. Traditional supervised algorithms such as random forests, support vector machines (SVMs), and gradient-boosted decision trees (e.g., XGBoost) continue to perform strongly, especially when working with structured, tabular clinical data. In many cases, these methods offer a valuable advantage: they can be easier to interpret and explain, an attribute that remains critically important in high-stakes clinical decision-making.

Deep learning architectures

Deep learning, a branch of machine learning built on multilayered neural networks, has become especially influential in medical image analysis because of its ability to learn complex patterns directly from raw data. Rather than relying on manually engineered features, these models automatically extract layered representations — moving from simple visual elements to subtle disease-specific signatures. Architectures such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and, more recently, transformer-based models have demonstrated strong performance across a range of healthcare tasks. In carefully controlled research settings, some systems have matched or even surpassed human experts in specific diagnostic challenges [6]. More recently, the field has begun shifting from narrowly designed, task-specific models toward large-scale foundation models that are pre-trained on extensive, often multimodal datasets and then fine-tuned for particular clinical applications — a transition that marks the current frontier of deep learning in healthcare.

Natural language processing and large language models

Natural language processing (NLP) makes it possible to extract meaningful clinical information from the vast amounts of unstructured text that permeate healthcare — including electronic health records, physician notes, medical literature, and patient-reported narratives. Although such text contains rich insights, it has historically been difficult to analyze systematically at scale [7]. Recent advances in large language models (LLMs), such as GPT-4, MedPaLM 2, and BioGPT, have significantly expanded these capabilities, demonstrating strong performance in tasks like medical question answering, clinical note summarization, and even aspects of diagnostic reasoning. However, despite their fluency and apparent coherence, these systems can sometimes produce statements that sound convincing but are factually incorrect — a phenomenon often described as "hallucination." In clinical contexts, this limitation is more than technical; it represents a meaningful patient safety concern that must be carefully addressed before widespread deployment.

Reinforcement learning

Reinforcement learning (RL) is an approach in which algorithms learn to make decisions by interacting with an environment and receiving feedback based on the outcomes of those decisions. Rather than being trained on fixed labeled examples, RL systems improve over time by identifying which actions lead to better results. In healthcare, this framework has shown promise in areas such as optimizing sepsis treatment protocols, personalizing insulin dosing in diabetes management, and supporting robotic surgical systems [8]. Yet its clinical application raises important challenges. Perhaps the most fundamental is how to define reward functions that truly reflect meaningful, long-term patient outcomes. If models are optimized around narrow or short-term surrogate metrics, they risk recommending strategies that improve numbers on paper without genuinely improving patient health or well-being.

AI in medical imaging and diagnostics

Medical imaging is the area of healthcare where AI has produced some of its most rigorously validated clinical successes. Several factors have contributed to this progress: the availability of large, well-annotated imaging datasets, clearly defined diagnostic questions, and the natural alignment between image data and deep learning architectures. Together, these conditions have made fields such as radiology especially fertile ground for the development, testing, and benchmarking of AI systems [9]. Table 2 summarizes key AI performance benchmarks across major diagnostic applications.

Clinical Domain	AI Method	Key Performance Metric	Dataset Size	Human Benchmark	Source
Chest X-Ray (Pneumonia)	CheXNet CNN	AUC 0.761	112,120 images	Radiologist AUC 0.658	[5]
Lung Cancer CT Screening	3D CNN	AUC 0.944	42,290 NLST screens	6 radiologists	[10]
Skin Cancer Detection	Inception v3 CNN	Dermatologist-level AUC	129,450 images	21 board-certified dermatologists	[6]
Diabetic Retinopathy	Deep CNN (Google)	Sensitivity 97.5%; Specificity 93.4%	128,175 retinal images	Ophthalmologist-equivalent	[10]
Multi-ethnic Retinopathy	Deep CNN (SingHealth)	AUC 0.936	494,661 images; 108,554 pts	Multi-centre validation	[11]
Arrhythmia Detection (ECG)	34-layer ResNet	AUC 0.97 (12 classes)	91,232 ECGs; 53,549 pts	Cardiologist benchmark	[12]
EHR Mortality/Readmission	LSTM deep learning	AUC 0.83–0.95	46,864 adult patients	APACHE II (AUC ~0.75)	[13]
Protein Structure Prediction	AlphaFold2 (transformer)	Median GDT 92.4 (CASP14)	170,000+ PDB structures	Next-best: GDT 75.0	[14]
Antibiotic Discovery	Message-passing DNN	Novel broad-spectrum activity	6,111 compound screen	No prior chemical analogue	[15]
Computational Pathology	MIL deep learning	AUC 0.97–0.98	44,732 WSIs; 15,187 pts	Reduces pathologist workload ≥70%	[16]
Sepsis Early Prediction	Graph neural network	6 h prior warning	PhysioNet 2019 Challenge	SOFA/qSOFA scores	[17]
AF Screening (wearable)	Smartwatch algorithm	PPV 0.84 for AF	419,297 participants	Standard Holter monitor	[18]

Table 2: AI Performance Benchmarks in Clinical Diagnostic Applications: Method, Dataset, Metric, and Human Comparator.

AUC = Area under the receiver operating characteristic curve; GDT = Global distance test; PPV = Positive predictive value; CNN = Convolutional neural network; MIL = Multiple instance learning; WSI = Whole-slide image; NLST = National Lung Screening Trial; ResNet = Residual neural network; LSTM = Long short-term memory; pts = Patients.

Radiology and chest imaging

In chest radiology, Rajpurkar., *et al*. [5] showed that a convolutional neural network (CNN) trained on more than 100,000 frontal chest radiographs could detect pneumonia with performance exceeding that of board-certified radiologists, achieving an AUC of 0.761 compared to a mean radiologist AUC of 0.658. This study marked an important proof of concept: it demonstrated that AI systems could approach — and in defined tasks, surpass — expert-level interpretation of chest X-rays.

In computed tomography (CT) imaging, Ardila., *et al*. [1] developed a deep learning model for lung cancer screening CT that achieved an AUC of 0.944 for identifying malignant nodules, outperforming six radiologists on data from the National Lung Screening Trial. Notably, the model also demonstrated superior ability to predict malignancy at two-year follow-up, suggesting that AI systems may contribute not only to immediate diagnostic decisions but also to more nuanced, forward-looking risk stratification in cancer screening programs.

Dermatology and skin cancer detection

Dermatology has become another area where AI is showing performance on par with — or even surpassing — human specialists. In a landmark study published in Nature, Esteva., *et al*. [6] trained a convolutional neural network (CNN) using over 129,000 clinical images representing more than 2,000 different skin diseases. When tested against 21 board-certified dermatologists, the AI achieved accuracy comparable to these experts in identifying keratinocyte carcinomas and melanomas.

Building on this, Han., *et al*. [19] showed that a deep learning algorithm trained on dermoscopy images could detect melanoma with sensitivity and specificity similar to senior dermatologists. The AI maintained strong performance across images from different dermatoscopes and clinical environments — a key factor for real-world use that many single-center studies often overlook.

Ophthalmology: Diabetic retinopathy screening

Diabetic retinopathy is one of the clearest examples of how AI can make a real difference in medical imaging. The disease affects millions worldwide, early treatment can prevent blindness, and yet there is a serious shortage of ophthalmologists — especially in low- and middle-income countries (LMICs) where diabetes is rising quickly [11].

In a landmark JAMA study, Gulshan., *et al*. [10] developed a deep learning algorithm capable of detecting diabetic retinopathy from retinal fundus photographs. The AI demonstrated high sensitivity (97.5%) and specificity (93.4%), matching or even surpassing the performance of ophthalmologists and retinal specialists. This system eventually received FDA clearance, marking one of the first AI diagnostic tools to be deployed in real-world clinical practice.

Ting., *et al*. [11] further validated an AI system on nearly half a million retinal images from over 100,000 patients across ten diverse, multi-ethnic cohorts. The algorithm achieved an AUC of 0.936 for detecting referable diabetic retinopathy, demonstrating strong performance across different ages, ethnicities, and imaging devices — robustness especially important for LMICs.

Computational pathology

Digital pathology — the process of digitizing histological slides and analyzing them with AI algorithms — has become a truly transformative application of artificial intelligence. In a landmark Nature Medicine study, Campanella., *et al*. [16] developed a deep learning system using multiple instance learning, trained on over 44,000 whole-slide images from more than 15,000 patients spanning 25 different cancer types. The system achieved AUCs of 0.98 for prostate cancer, 0.97 for basal cell carcinoma, and 0.98 for breast cancer metastases.

Rather than replacing pathologists, this AI approach is designed to support them by flagging the large proportion of slides that are confidently negative, effectively reducing workload without compromising diagnostic quality. This workflow-compatible model represents a practical path to real-world deployment [20].

AI in genomics, proteomics, and precision medicine

AI is also transforming genomics and proteomics, arguably one of the most scientifically groundbreaking frontiers in precision medicine. The human genome contains around 3 billion base pairs, and interpreting genomic variation — figuring out which variants are pathogenic, predicting their functional impact, and translating this information into clinical decisions — is a monumental computational challenge. AI is uniquely suited to tackle this complexity [21].

A landmark breakthrough came with AlphaFold2, developed by DeepMind and published in Nature [14]. For the first time, AI achieved near-experimental accuracy in predicting protein structures from amino acid sequences — a problem that had eluded scientists for decades. On the CASP14 benchmark, AlphaFold2 reached a median GDT score of 92.4, far outperforming previous methods. These predicted structures are freely available through the European Bioinformatics Institute, accelerating research in drug discovery, understanding genetic diseases, and developing new therapeutics.

Beyond protein folding, AI is being applied to a wide range of genomic challenges: classifying genetic variants, developing polygenic risk scores, analyzing single-cell transcriptomics, and integrating multi-omics data to better stratify patients. Techniques like graph neural networks are enabling predictions of protein-protein interactions and drug-target relationships, opening new avenues for therapeutic discovery [21].

AI in drug discovery and development

Drug discovery is one of the most expensive and time-consuming processes in medicine — the development of a new

drug from initial discovery to regulatory approval typically requires ten to fifteen years and costs over USD 2 billion, with a failure rate exceeding 90% in clinical trials [15]. AI offers tools to accelerate and improve multiple stages of this process, from target identification through lead optimization to clinical trial design. Table 3 presents an overview of key AI platforms and their drug discovery achievements.

Tool/Platform	AI Method	Application Stage	Key Achievement	Developer	Ref.
AlphaFold2	Transformer + evolutionary attention	Target identification; structure prediction	Solved >200M protein structures; accelerated drug target discovery globally	DeepMind/ EMBL-EBI	[14]
Halicin (DMPNN)	Message-passing neural network	Lead compound identification	Novel antibiotic active against pan-resistant M. tuberculosis and CR Acinetobacter	MIT/Broad Institute	[15]
Insilico Medicine (Chemistry42)	Generative adversarial network; RL	De novo molecular design	INS018_055 to Phase II in 30 months vs industry avg. 6+ years for fibrosis target	Insilico Medicine	[22]
Schrödinger (FEP+)	Physics-informed ML + FEP	Lead optimisation; ADMET prediction	30–50% hit rate improvement vs traditional screening; multiple clinical candidates	Schrödinger Inc.	[23]
Atomwise (AtomNet)	3D CNN on molecular structures	Virtual screening; target binding	Identified Ebola and multiple sclerosis drug candidates through structure-based screening	Atomwise Inc.	[23]
Clinical trial outcome ML	Multi-modal ML (XGBoost, RF)	Clinical trial design/failure prediction	Predicted trial failure with accuracy substantially exceeding conventional approaches	Academic/ industry	[24]

Table 3: AI Platforms in Drug Discovery: Methods, Applications, and Key Achievements.

DMPNN = Directed message-passing neural network; RL = Reinforcement learning; GAN = Generative adversarial network; FEP = Free energy perturbation; ADMET = Absorption, distribution, metabolism, excretion, toxicity; 3D CNN = Three-dimensional convolutional neural network; RF = Random forest; CR = Carbapenem-resistant.

Target identification and molecular design

In a landmark study published in Cell, Stokes., *et al*. [15] trained a deep learning model on 2,335 molecules with known growth inhibition against Escherichia coli and used it to screen a library of over 6,000 compounds. The model identified halicin — a molecule structurally unlike conventional antibiotics — as a powerful bactericidal agent effective against a broad spectrum of drug-resistant pathogens, including Mycobacterium tuberculosis and carbapenem-resistant Acinetobacter baumannii. This work demonstrated that AI can discover entirely novel drug candidates by learning molecular features linked to bioactivity.

Similarly, Insilico Medicine's Chemistry42 platform uses generative adversarial networks and reinforcement learning for de novo molecular design. The platform advanced a novel fibrosis candidate (INS018_055) from initial target identification to Phase II clinical trials in roughly 30 months — dramatically faster than the industry average of six or more years [22].

Clinical trial optimization

AI has quickly moved from theoretical concept to real-world clinical use. In medical imaging, AI can classify skin cancers with dermatologist-level accuracy, detect diabetic retinopathy at the level of ophthalmologists, interpret radiographs as well as radiologists,

and analyze whole-slide histology with pathologist-level precision. In genomics and proteomics, AI systems such as AlphaFold2 have solved long-standing challenges in protein structure prediction. In drug development, AI has accelerated the discovery of novel therapeutics such as halicin, shortened molecular design timelines, and improved clinical trial efficiency through optimized patient recruitment, adaptive trial design, and predictive modeling of trial outcomes [23,24].

AI in electronic health records and clinical decision support

Predictive analytics from EHR data

Electronic health records (EHRs) hold enormous amounts of clinical information — diagnoses, medications, lab results, vital signs, clinical notes, and procedural records. In a landmark study, Rajkomar., *et al*. [13] trained deep learning models on de-identified EHR data from 46,864 adult patients across two academic medical centers. The models were able to predict a variety of clinical outcomes — including in-hospital mortality, 30-day unplanned readmission, prolonged hospital stays, and discharge diagnoses — with performance far exceeding that of traditional clinical prediction tools, such as APACHE II (mortality AUC ~0.75). Across different outcomes, the models achieved AUCs between 0.83 and 0.95, highlighting the potential of AI to support clinical risk stratification.

Sepsis prediction and early warning systems

Sepsis causes an estimated 270,000 deaths annually in the United States and is responsible for a significant proportion of ICU mortality [25]. Early recognition and treatment are critical to outcomes, making sepsis prediction a high-priority application for AI-based clinical decision support. Reyna., *et al*. [26] evaluated 27 ML models for sepsis prediction in the PhysioNet Challenge 2019, finding that best-performing models could predict sepsis six or more hours before clinical recognition with useful sensitivity and specificity. Moor., *et al*. [17] demonstrated that graph neural network approaches incorporating temporal dependencies in vital sign and laboratory data achieved particularly strong performance.

AI in mental health: digital phenotyping and risk prediction

Digital phenotyping and passive sensing

Traditionally, mental health assessments rely on patients' self-reports and occasional clinical visits, offering only brief glimpses into their day-to-day functioning and symptom patterns [27]. Digital phenotyping — the continuous measurement of behavior and physiology using personal devices such as smartphones, wearables, and social media — offers a way to monitor mental health more objectively and in real time [28]. AI can use diverse signals including GPS data, accelerometer readings, typing behaviors, and social media language to predict and track mental health outcomes.

Orru., *et al*. [29] reviewed 28 studies using machine learning on digital biomarkers for depression and anxiety, finding that algorithms combining multiple data streams achieved 70–90% accuracy in detecting depression. Interestingly, changes in digital biomarkers often appeared days or weeks before patients reported symptom changes, opening the door for early, pre-symptomatic interventions.

Suicide risk prediction

Predicting suicide risk is among the most difficult and high-stakes applications of AI in mental health. A meta-analysis by Franklin., *et al*. [30] found that over fifty years of research, traditional methods had barely improved in accuracy, with algorithms only slightly better than chance at predicting individual suicide attempts. More recent work shows promise: Walsh., *et al*. [31] used EHR data from 5,167 patients to develop a machine learning model that predicted near-term suicide attempts far more accurately than standard clinical scales. Similarly, Kessler., *et al*. [32] achieved AUCs above 0.84 in predicting suicide attempts and deaths in large epidemiological datasets.

Wearable technology and remote patient monitoring

Cardiac monitoring and arrhythmia detection

Wearable devices that continuously monitor physiological signals are opening new opportunities for AI-driven health surveillance outside traditional clinical settings. In the Apple Heart Study, Perez., *et al*. [18] enrolled over 419,000 participants and used a smartwatch-based algorithm to detect irregular pulses indicative of atrial fibrillation — the most common sustained cardiac arrhythmia and a major risk factor for stroke. Among those notified of an irregular pulse, 34% were confirmed to have atrial fibrillation on follow-up monitoring.

Similarly, Hannun., *et al*. [12] trained a deep neural network on over 91,000 single-lead ECG recordings from more than 53,000 patients to classify 12 different heart rhythm types. The model achieved cardiologist-level performance, with an AUC of 0.97 across rhythm classes, and outperformed individual cardiologists in detecting arrhythmias such as atrial fibrillation and idioventricular rhythm.

Ethical considerations, algorithmic bias, and health equity

The development and deployment of AI in healthcare raises profound ethical challenges. Table 4 provides a structured analysis of the principal challenge domains, exemplar evidence, current mitigation approaches, and outstanding research priorities.

Challenge Domain	Specific Issue	Exemplar Evidence	Current Mitigation Approaches	Research Priorities
Algorithmic Bias	Training data demographic imbalance; proxy variable misuse	Commercial algorithm underestimated Black patient needs by 26% using cost proxy [33]	Fairness-aware ML; diverse data curation; outcome variable audit	Bias metrics standardization; equity-by-design frameworks
Transparency and Explainability	Black-box deep learning; clinician trust deficit	LIME, SHAP post-hoc explanations; attention maps [34]	Explainable AI (XAI) methods; inherently interpretable architectures	Clinically meaningful explanation standards; regulatory XAI requirements
Data Privacy and Governance	Reidentification risk; cross-border data flows; consent validity	HIPAA (US), GDPR (EU) misalignment with AI training needs [35]	Federated learning; differential privacy; synthetic data	International health data governance frameworks; privacy-preserving ML standards
Regulatory Approval	Adaptive AI (algorithm drift); SaMD classification	FDA cleared >500 AI/ML devices by 2023; post-market surveillance incomplete [36]	FDA AI/ML action plan; EU AI Act Medical Device Regulation alignment	Continuous learning regulatory models; real-world performance monitoring
Generalisability	Single-centre training; dataset shift; temporal drift	Systematic review found majority of AI studies lack external validation [37]	Multi-site validation consortia; federated evaluation protocols	Standardized reporting (TRIPOD-AI); mandatory prospective validation
Clinical Workflow Integration	Alert fatigue; EHR incompatibility; clinical inertia	CDSS adoption failures despite strong algorithm performance [38]	Human-centred design; implementation science frameworks; change management	Real-world effectiveness trials; clinician-AI collaboration models
Health Equity	AI widening global diagnostic divide; LMICs left behind	AI retinopathy screening validated in HICs; infrastructure barriers in LMICs [11]	Mobile-first AI design; open-source models; global health partnerships	Equity-first AI development frameworks; LMIC-inclusive datasets

Table 4: Ethical and Regulatory Challenges in Healthcare AI: Evidence, Mitigations, and Research Priorities.

AI = Artificial intelligence; ML = Machine learning; LIME = Locally interpretable model-agnostic explanations; SHAP = SHapley Additive exPlanations; HIPAA = Health Insurance Portability and Accountability Act; GDPR = General Data Protection Regulation; FDA = Food and Drug Administration; SaMD = Software as a medical device; TRIPOD-AI = Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (adapted for AI); CDSS = clinical decision support system; LMIC = Low- and middle-income country; HIC = high-income country.

Algorithmic bias and disparate impact

One of the most pressing ethical challenges in healthcare AI is algorithmic bias — the tendency of AI systems trained on historical data to reinforce or amplify existing disparities [33]. A landmark study published in Science by Obermeyer., *et al*. [33] revealed that a widely used commercial risk stratification algorithm systematically

underestimated the healthcare needs of Black patients compared with White patients who had the same illness severity. The algorithm relied on healthcare costs as a proxy for health needs, but because Black patients historically incur lower healthcare costs due to limited access, they were assigned lower risk scores. This study highlights a crucial lesson: selecting an outcome variable in AI is not a neutral technical decision — it is an ethical one, with direct implications for equity in healthcare delivery.

Transparency, explainability, and clinical trust

A major barrier to adopting AI in clinical practice is the "black box" nature of many deep learning models — the difficulty in understanding why a specific prediction was made or which features of the input data influenced the output [39]. To address this, researchers have developed tools like attention mechanisms, saliency maps, and post-hoc explanation methods such as LIME and SHAP [34]. Yet, as Rudin [40] emphasizes, for high-stakes clinical decisions, it is often preferable to use models that are inherently interpretable rather than trying to explain opaque systems after the fact.

Privacy, data governance, and regulatory frameworks

The development of effective AI systems for healthcare requires access to large, high-quality datasets of patient information that are inherently sensitive and raise fundamental questions of privacy, consent, and data governance [35]. Regulatory frameworks — HIPAA in the United States, GDPR in Europe — were developed before the era of large-scale AI and provide imperfect guidance for the novel data practices that AI development requires. The FDA has published guidance on the regulation of software as a medical device and has developed a framework for AI/ML-based software that adapts over time, but important questions about post-market surveillance, algorithm drift, and the regulation of continuously learning systems remain unresolved [36].

Challenges and future directions

The validation gap

One of the biggest challenges for AI in healthcare is the gap between how these systems perform in research studies and how they work in real-world clinical settings. Most medical AI studies are retrospective, rely on data from a single center, and test performance on datasets very similar to the ones they were trained on. Reviews of AI diagnostic studies have highlighted worrying trends, including over fitting, poor external validation, and biased reporting, suggesting that reported performance is often overly optimistic [37]. To address these issues, two key steps are needed: adopting strict reporting guidelines such as the TRIPOD-AI extension, and requiring prospective, multi-site validation before AI systems are approved for clinical use.

Clinical workflow integration

Even AI systems that demonstrate strong performance in research settings frequently encounter significant barriers to clinical integration. Poorly designed user interfaces, disruption of established clinical workflows, alert fatigue, and the challenge of training clinical staff to appropriately calibrate their trust in AI recommendations are practical obstacles that have limited the adoption of otherwise promising systems [38]. The implementation science literature emphasizes that successful technology adoption requires sustained change management, iterative co-design with clinical users, and governance frameworks that define accountability for AI-assisted clinical decisions.

Foundation models and the multimodal future

One of the most exciting directions for AI in healthcare is bringing together different types of data — like imaging, genomics, electronic health records, wearable sensors, and patient-reported outcomes — into large-scale foundation models that can support a wide range of clinical tasks [41]. Models such as GPT-4, Med-PaLM 2, and BioMedBERT are already showing impressive skills in medical question answering and clinical reasoning.

Moor., *et al*. [41] describe a vision of "generalist medical AI": a single, continuously updated model that could assist clinicians across nearly every aspect of healthcare. By combining these diverse data streams with the reasoning power of large models, such systems could help deliver care that is more informed, personalized, and timely — potentially transforming the way medicine is practiced.

Discussion in Brief

This review has brought together evidence across nine major areas of AI application in human health, highlighting a picture of rapidly advancing capabilities alongside persistent, serious challenges. Several key themes emerge.

First, medical imaging is the most mature domain for AI. Large-scale validation studies in radiology, dermatology, ophthalmology, and pathology consistently show AI performance that meets or exceeds human specialists under controlled conditions. Yet even here, translating research success into real-world clinical impact has proven more difficult than expected. Prospective, randomized evaluations comparing AI-assisted and conventional diagnostic workflows often show more modest or mixed improvements than retrospective benchmark studies suggested.

Second, drug discovery represents some of the most transformative scientific applications of AI. Breakthroughs such as AlphaFold2 and halicin demonstrate AI's ability to enable discoveries that would have been impossible or decades away using traditional methods. AlphaFold2, in particular, has revolutionized structural biology, and its impact on drug discovery is likely to be one of the most significant scientific developments of the twenty-first century.

Third, the ethical dimensions of AI in healthcare — algorithmic bias, privacy, and accountability — are not secondary concerns to be addressed after technical development. They must be considered from the very beginning. The Obermeyer., *et al*. [33] study vividly demonstrated that even highly sophisticated AI systems can inadvertently harm the very populations they are intended to help if based on flawed assumptions. Going forward, the field must adopt equity-by-design frameworks with the same rigour currently applied to technical performance optimization.

Highlights

- AI diagnostic systems match or exceed specialist clinician performance across radiology, dermatology, ophthalmology, and pathology in controlled evaluations.
- Deep learning applied to electronic health records predicts critical outcomes — mortality, sepsis, readmission — with AUCs of 0.83–0.97.
- AlphaFold2 and deep neural networks have transformed drug target identification and novel antibiotic discovery.
- Algorithmic bias, limited external validation, regulatory gaps and workflow integration barriers remain critical unresolved challenges.
- Foundation models and multimodal AI represent the most transformative near-term trajectory for AI in clinical medicine.

Conclusions

Artificial intelligence has moved from a theoretical possibility to a clinical reality across many areas of human health, showing performance in certain tasks that can match — or even surpass — that of human specialists. The studies reviewed here offer reason for cautious optimism: AI has the potential to improve diagnostic accuracy, speed up drug discovery, enable more personalized treatment, extend specialist expertise to underserved populations, and transform the efficiency and effectiveness of healthcare systems.

Yet significant challenges remain. Without careful attention to equity, algorithmic bias could worsen health disparities rather than reduce them. Many AI systems are inadequately validated outside their original development settings, meaning published performance often overstates what can be achieved in real-world practice. Regulatory frameworks lag behind rapid technological progress, and integrating AI into clinical workflows requires changes in training, culture, and professional practice that are complex and slow.

The future of AI in healthcare will depend not only on technical innovation but on the wisdom, values, and commitment to fairness with which these tools are designed, tested, and deployed. The goal should not be to replace human judgment, but to enhance it — positioning AI as a genuine partner in the timeless human mission of healing, accountable both to patients and the health systems that support them.

Ethical Approval and Consent to Participate

Not applicable. This study is a systematic narrative review of previously published literature and does not involve primary data collection from human participants or animals.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflicts of Interest

The authors declare no conflicts of interest relevant to this work.

Author Contributions

Manabendra Debnath: Conceptualization, Methodology, Literature search, Writing – original draft, Review and editing. Biplab De: Conceptualization, Supervision, Writing – review and editing. Both authors read and approved the final manuscript.

Data Availability

Not applicable. This is a review article; no primary datasets were generated or analyzed.

Bibliography

1. Ardila D., *et al*. "End-to-End Lung Cancer Screening with Deep Learning on Low-Dose CT". *Nature Medicine* 25.6 (2019): 954-961.

2. Topol E J. "High-Performance Medicine: The Convergence of Human and Artificial Intelligence". *Nature Medicine* 25.1 (2019): 44-56.

3. Obermeyer Z and E J Emanuel. "Predicting the Future — Big Data, Machine Learning, and Clinical Medicine". *New England Journal of Medicine* 375.13 (2016): 1216-1219.

4. Char D S., *et al*. "Implementing Machine Learning in Health Care — Addressing Ethical Challenges". *New England Journal of Medicine* 378.11 (2018): 981-983.

5. Rajpurkar P., *et al*. "CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning". arXiv, arXiv:1711.05225 (2017).

6. Esteva A., *et al*. "Dermatologist-Level Classification of Skin Cancer with Deep Neural Networks". *Nature* 542.7639 (2017): 115-118.

7. Wang Y., *et al*. "A Clinical Text Classification Paradigm Using Weak Supervision and Deep Representation". *BMC Medical Informatics and Decision Making* 19.1 (2018): 1.

8. Yu C., *et al*. "Reinforcement Learning in Healthcare: A Survey". *ACM Computing Surveys* 55.1 (2021): 1-36.

9. Shen D., *et al*. "Deep Learning in Medical Image Analysis". *Annual Review of Biomedical Engineering* 19 (2017): 221-248.

10. Gulshan V., *et al*. "Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs". *JAMA* 316.22 (2016): 2402-2410.

11. Ting D S W., *et al*. "Development and Validation of a Deep Learning System for Diabetic Retinopathy and Related Eye Diseases Using Retinal Images from Multiethnic Populations with Diabetes". *JAMA* 318.22 (2017): 2211-2223.

12. Hannun A Y., *et al*. "Cardiologist-Level Arrhythmia Detection and Classification in Ambulatory Electrocardiograms Using a Deep Neural Network". *Nature Medicine* 25.1 (2019): 65-69.

13. Rajkomar A., *et al*. "Scalable and Accurate Deep Learning with Electronic Health Records". *npj Digital Medicine* 1.1 (2018): 18.

14. Jumper J., *et al*. "Highly Accurate Protein Structure Prediction with AlphaFold". *Nature* 596.7873 (2021): 583-589.

15. Stokes J M., *et al*. "A Deep Learning Approach to Antibiotic Discovery". *Cell* 180.4 (2020): 688-702.

16. Campanella G., *et al*. "Clinical-Grade Computational Pathology Using Weakly Supervised Deep Learning on Whole Slide Images". *Nature Medicine* 25.8 (2019): 1301-1309.

17. Moor M., *et al*. "Early Warning in the ICU by Prospective, Individualized, and Dynamic Prediction of Infection by a Novel ML-Based System". *PLOS Computational Biology* 17.2 (2021): e1008684.

18. Perez MV., *et al*. "Large-Scale Assessment of a Smartwatch to Identify Atrial Fibrillation". *New England Journal of Medicine* 381.20 (2019): 1909-1917.

19. Han S S., *et al*. "Classification of the Clinical Images for Benign and Malignant Cutaneous Tumors Using a Deep Learning Algorithm". *Journal of Investigative Dermatology* 138.7 (2018): 1529-1538.

20. Litjens G., *et al*. "A Survey on Deep Learning in Medical Image Analysis". *Medical Image Analysis* 42 (2017): 60-88.

21. Zou J., *et al*. "A Primer on Deep Learning in Genomics". *Nature Genetics* 51.1 (2019): 12-18.

22. Ren F., *et al*. "AlphaFold Accelerates Artificial Intelligence Powered Drug Discovery". *Chemical Science* 14.6 (2023): 1443-1452.

23. Harrer S., *et al*. "Artificial Intelligence for Clinical Trial Design". *Trends in Pharmacological Sciences* 40.8 (2019): 577-591.

24. Woo M., *et al*. "Prediction of Clinical Trials Outcomes Based on Target Choice and Clinical Trial Design with Multi-Modal Artificial Intelligence". *Clinical and Translational Science* 14.3 (2019): 1116-1124.

25. Singer M., *et al*. "The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3)". *JAMA* 315.8 (2016): 801-810.

26. Reyna M A., *et al*. "Early Prediction of Sepsis from Clinical Data: The PhysioNet/Computing in Cardiology Challenge 2019". *Critical Care Medicine* 48.2 (2019): 210-217.

27. Torous J., *et al*. "New Tools for New Research in Psychiatry: A Scalable and Customizable Platform to Empower Data Driven Smartphone Research". *JMIR Mental Health* 3.2 (2018): e16.

28. Insel TR. "Digital Phenotyping: Technology for a New Science of Behavior". *JAMA* 318.13 (2017): 1215-1216.

29. Orru G., *et al*. "Human Machine Decision-Making in Clinical Psychology and Neuroscience: A Systematic Review". *Frontiers in Psychology* 11 (2020): 491.

30. Franklin J C., *et al*. "Risk Factors for Suicidal Thoughts and Behaviors: A Meta-Analysis of 50 Years of Research". *Psychological Bulletin* 143.2 (2017): 187-232.

31. Walsh C G., *et al*. "Predicting Risk of Suicide Attempts over Time through Machine Learning". *Clinical Psychological Science* 5.3 (2017): 457-469.

32. Kessler RC., *et al*. "Suicide Prediction Models: A Critical Review of Recent Research with Recommendations for the Way Forward". *Molecular Psychiatry* 25.1 (2019): 168-179.

33. Obermeyer Z., *et al*. "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations". *Science* 366.6464 (2019): 447-453.

34. Lundberg S M and S I Lee. "A Unified Approach to Interpreting Model Predictions". *Advances in Neural Information Processing Systems* 30 (2017): 4765-4774.

35. Price W N and I G Cohen. "Privacy in the Age of Medical Big Data". *Nature Medicine* 25.1 (2019): 37-43.

36. Benjamens., *et al*. "The State of Artificial Intelligence-Based FDA-Approved Medical Devices and Algorithms: An Online Database". *npj Digital Medicine* 3.1 (2020): 118.

37. Liu X., *et al*. "A Comparison of Deep Learning Performance against Health-Care Professionals in Detecting Diseases from Medical Imaging: A Systematic Review and Meta-Analysis". *The Lancet Digital Health* 1.6 (2019): e271-e297.

38. Cresswell K M., *et al*. "Ten Key Considerations for the Successful Implementation and Adoption of Large-Scale Health Information Technology". *Journal of the American Medical Informatics Association* 20.e1 (2013): e9-e13.

39. Doshi-Velez F and B Kim. "Towards a Rigorous Science of Interpretable Machine Learning". arXiv, arXiv:1702.08608 (2017).

40. Rudin C. "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead". *Nature Machine Intelligence* 1.5 (2019): 206-215.

41. Moor M., *et al*. "Foundation Models for Generalist Medical Artificial Intelligence". *Nature* 616.7956 (2023): 259-265.