Proceedings of ICBioMed 2025 Symposium: AI for Healthcare: Advanced Medical Data Analytics and Smart Rehabilitation
DOI: 10.54254/2753-8818/2025.AU27956

Research on the Application of Data Science in the Medical Field

Boshen Zhu

Beijing Institute of Petrochemical Technology, Beijing, China
3218057401@qq.com

Abstract. The application of data science in the medical field is profoundly transforming the development model of modern healthcare systems. With the digital transformation in healthcare, data science has introduced new opportunities and methodologies for disease prediction, diagnosis, and treatment. This study examines data science's key applications in medicine, emphasizing infectious disease early warning, medical imaging, and clinical decision support. In infectious disease surveillance, the integration and intelligent analysis of diverse data sources have greatly improved early warning system efficacy, bolstering public health management. Applying deep learning technologies in medical imaging has substantially improved diagnostic accuracy and efficiency, optimizing clinical workflows. For clinical decision support, techniques such as natural language processing help extract valuable insights from medical records, assisting physicians in developing more precise treatment plans. Innovative applications in pharmacogenomics and chronic disease management are propelling personalized medicine. Data integration, algorithmic robustness, and ethical regulations remain challenging. Future research necessitates interdisciplinary collaboration, improved technical standards, and regulatory frameworks for sustainable data science in healthcare.

Keywords: Data Science, Healthcare, Artificial Intelligence, Disease Prediction, Clinical Decision-Making

1. Introduction

In recent years, with the rapid advancement of information technology, the application of data science in healthcare has become a significant force driving transformation in the medical industry.Globally, medical data has grown exponentially due to electronic health records, digitized imaging, wearables, and genomic sequencing, enriching medical data science. At the same time, governments worldwide have introduced policies to support the digital transformation of healthcare, such as the United States' "Precision Medicine Initiative" and the European Union's "Digital Europe Programme," further highlighting the strategic importance of medical data at the national level. Despite advances in medical data science, data integration, algorithm generalizability, clinical translation, and ethical regulations still face challenges. Data silos impede data use, and algorithmic performance varies across populations.

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-8818/2025.AU27956

This paper aims to systematically explore the current applications, key technologies, and future directions of data science in healthcare, specifically focusing on its practical implementations in infectious disease early warning, medical image analysis, and clinical decision support. By reviewing the latest domestic and international research, this study adopts a methodology combining literature review and case analysis to examine data science's effectiveness and potential challenges in medical settings. Beyond technological innovations, the research also investigates the broader value of data science in healthcare management, policy-making, and ethical governance.

This study offers theoretical insights and practical direction for healthcare's digital shift. It highlights current tech strengths/weaknesses to guide research, suggests solutions/policies for data science challenges in medicine, and forecasts medical data science's future. This aids healthcare organizations, businesses, and regulators in decision-making, fostering data science-healthcare integration to boost service quality/efficiency, resource use, and public health emergency response.

Three distinct features characterize international research in medical data science. First, it facilitates multidisciplinary collaboration. A notable example is the United States, where an innovation triangle model has emerged involving "academia (e.g., the MIT Jameel Clinic), healthcare institutions (e.g., Mayo Clinic), and technology companies (e.g., Google Health)." Second, there is a noticeable shift toward empirical methodologies. According to Nature Medicine statistics, most medical AI papers published between 2020 and 2023 adopted prospective clinical trial data for validation, reflecting an increase of several percentage points compared to the previous three-year period. Finally, applications are becoming more specialized and refined. The field now encompasses a wide range of internationally recognized sub-disciplines, covering complete life-cycle health services—from gene editing to end-of-life care.

2. Disease surveillance and public health applications

2.1. Multi-source data fusion

Modern infectious disease early warning systems have moved beyond the limitations of traditional single-source data, evolving into multi-dimensional data collection networks. Search engine query data, exemplified by Google Trends, offers a means to detect changes in population-level symptom-seeking behavior; research indicates that increases in search volumes for influenza-related terms serve as leading indicators of increases in outpatient visits.. By employing natural language processing (NLP), social media sentiment monitoring analyzes the volume of disease-related discussions on platforms such as Twitter and Weibo. A notable example is Boston Children's Hospital’s HealthMap system, which issued an alert several days before the official recognition of the 2014 Ebola outbreak by monitoring various global data sources.

Mobile device location data offers novel approaches for transmission chain tracing. Leveraging anonymized mobile signaling data, an MIT research team has effectively modeled the propagation of the dengue fever epidemic in Singapore, achieving a spatial resolution of up to 100 meters. It is important to note that the integration of multi-source data presents difficulties pertaining to data heterogeneity [1]. The newly developed Spatio-Temporal Graph Convolutional Network (ST-GCN) can effectively integrate data with varying spatiotemporal granularities, enhancing early warning systems' accuracy.

DOI: 10.54254/2753-8818/2025.AU27956

2.2. Advances in predictive modeling

Deep learning models are revolutionizing the paradigm of infectious disease prediction. Long Short-Term Memory (LSTM) networks excel at capturing temporal patterns in disease transmission. A predictive model for COVID-19 trends, developed by researchers at Beijing Normal University, leveraged a twelve-dimensional feature set encompassing meteorological factors and population mobility data, demonstrating a notably low prediction error [2]. On the other hand, Graph Neural Networks (GNNs) are particularly adept at modeling spatial spread dynamics. The EpiGNN model, crafted by the NIH, adeptly forecasted influenza virus transmission patterns across regions by scrutinizing intercity traffic flow data [3].

Meanwhile, ensemble learning methods have demonstrated powerful robustness. A hybrid model integrating Random Forest, XGBoost, and neural networks significantly outperformed all single models' overall performance during an African malaria prediction challenge. A recent advancement involves the introduction of physics-informed AI models. For instance, the EpiLearn framework, developed by Tsinghua University, incorporates epidemiological SEIR equations into neural networks. This approach preserves the significant advantage of being data-driven while also adhering to disease transmission dynamics, resulting in reduced prediction errors in testing.

3. Clinical Decision Support (CDS)

3.1. Intelligent diagnosis of medical images

The field of medical imaging AI has undergone a paradigm shift from traditional image processing to deep learning. Early computer-aided detection (CAD) systems relied on handcrafted features—such as texture and shape—which resulted in relatively low sensitivity in detecting microcalcifications in mammography After 2012, convolutional neural networks (CNNs) instigated a paradigm shift. The triumph of AlexNet in the ImageNet challenge catalyzed the medical research community to investigate transfer learning methodologies. Recent studies demonstrate that 3D ResNet-152 achieves exceptionally high sensitivity in detecting pulmonary nodules in CT scans and significantly reduces false positive rates.

The integration of multi-modal imaging has emerged as a significant new trend. For example, the CheXNeXt system developed at Stanford University is capable of concurrently analyzing chest X-rays, CT scans, and ultrasound data, thereby facilitating comparative lesion analysis via a cross-modal attention mechanism [4]. The RadFusion framework, reported in Nature Medicine in 2023, integrates features derived from multiple imaging devices and has substantially improved diagnostic accuracy for rare diseases. It is also noteworthy that radiomics—by extracting high-throughput quantitative features—has identified novel imaging biomarkers, such as those indicating breast cancer heterogeneity.

Significant progress has been made in the intelligent transformation of radiology workflows. Numerous AI-based imaging products have been approved by the U.S. FDA, covering eleven specialities including thoracic, neurological, and musculoskeletal imaging. For instance, Tencent's MI Imaging platform has deployed a pneumonia CT-assisted diagnosis system in multiple hospitals across China, which has reduced average reporting time by nearly half and significantly decreased the rate of missed diagnoses. In the field of pathological slide analysis, Google's LYNA system has achieved exceptionally high accuracy in detecting lymph node metastasis of breast cancer, surpassing the performance of human pathologists [5].

DOI: 10.54254/2753-8818/2025.AU27956

AI-guided interventional imaging is demonstrating growing clinical value. For instance, Siemens' AI-Rad Companion is capable of annotating vascular stenosis in real time during interventional procedures, which leads to a marked reduction in procedural duration [6]. More cutting-edge applications include dynamic tumor margin identification using intraoperative MRI and automated puncture path planning under ultrasound guidance. These technologies are reshaping clinical protocols—though careful attention must be paid to responsibility allocation in human-AI collaboration.

Medical imaging AI faces a "lab-to-clinic" translation gap. An MIT investigation revealed that the performance of the lung nodule algorithm varied considerably across hospitals, which can be attributed to differences in equipment configurations and reconstruction algorithms [7]. To address this, NEMA introduced the AI-DICOM standard to regulate data input and output formats. The China Association of Medical Equipment has issued the "Medical Imaging AI Product Quality White Paper," which defines an evaluation framework consisting of more than one hundred parameters [8].

Robustness enhancement has become a research hotspot. Adversarial Training significantly improves the model's adaptability to variations in CT scanning parameters, while Test-time Augmentation techniques enhance performance on external validation sets without requiring retraining. The most recent FDA regulations stipulate that AI-driven medical devices must incorporate ongoing learning protocols to guarantee sustained model integrity within clinical settings [9].

3.2. Auxiliary diagnosis and clinical decision-making

Natural language processing (NLP) technologies have unlocked the value of medical record texts. BERT achieves exceptionally high F1 scores in automated encoding of Chinese discharge diagnoses, significantly outperforming traditional ICD coders. The clinical event extraction system developed by Mayo Clinic can identify multiple types of key information from unstructured texts, supporting the automatic generation of progress notes that are compliant with the SOAP format. The most recent innovation employs temporal relationship modeling—for instance, Harvard researchers' Temporal-BERT—to precisely reconstruct the progression trajectory in patients with diabetes mellitus [10].

Integrating multi-source clinical data enhances diagnostic accuracy. Johns Hopkins University's DeepDiagnosis system, through joint analysis of laboratory tests, vital signs, and medication records, enables earlier identification of sepsis. The "Smart Medical Assistant" developed by the First Affiliated Hospital of USTC (University of Science and Technology of China) automatically retrieves historical medical records upon input of patient complaints to generate a differential diagnosis list. Clinical tests show an exceptionally high consistency rate with diagnoses made by attending physicians.

Automated extraction of literature knowledge accelerates evidence-based practice. IBM Watson for Oncology analyzes vast amounts of newly published literature monthly to maintain the timeliness of treatment recommendations. The EBM-AI system developed by the Peking University team constructs dynamic evidence maps using relation extraction technology, significantly reducing the time spent on clinical trial screening. The latest research explores integrating large language models (e.g., GPT-4) with evidence-based databases, though caution is warranted against potential "scientific hallucinations" such models may generate.

Pharmacogenomic AI models reduce the risk of adverse drug reactions. The University of Chicago's PREDICT system integrates multiple gene loci related to drug metabolism to predict

DOI: 10.54254/2753-8818/2025.AU27956

personalized dosages for medications such as warfarin, significantly reducing severe adverse reactions. Chinese researchers have developed an HLA-AI tool that leverages deep sequencing data to forecast the potential toxicity risks associated with immune checkpoint inhibitors. A current challenge lies in addressing racial disparities, as pharmacogenomic data from African populations remains severely underrepresented in global databases.

Reinforcement learning is a promising approach to chronic disease management. For instance, Google DeepMind's algorithm for glycemic control leverages continuous glucose monitoring data to dynamically modulate insulin delivery, thereby increasing the duration of time-in-range for individuals with type 1 diabetes. MD Anderson Cancer Center's chemotherapy optimization system dynamically adjusts dosing plans based on patient tolerance, improving treatment completion rates. These systems must resolve the "exploration-exploitation" dilemma to balance efficacy validation and personalized experimentation.

Preoperative virtual simulation enhances surgical safety. France's PhysIKA system generates digital twins of livers from patient CT data to predict functional compensation under different resection plans. Beijing Tiantan Hospital's AI-powered neuronavigation automatically corrects MRI registration errors during surgery, achieving high success rates in protecting the optic nerve during pituitary tumor operations. Future directions include integrating augmented reality technology to enable holographic visualization of anatomical structures.

4. Conclusion

This study systematically examines the current applications, key technologies, and future development trends of data science in the medical field. Infectious disease early warning systems, medical imaging analysis, and clinical decision support are just a few of the key areas where data science is revolutionizing contemporary healthcare paradigms. These advancements offer cutting-edge technological tools for disease prediction, diagnosis, and treatment. The integration of multi-source data and sophisticated algorithms has greatly improved epidemic monitoring and early warning capabilities in the fight against infectious diseases. In medical imaging, deep learning technologies have substantially improved diagnostic accuracy and efficiency. In clinical decision-making, natural language processing and knowledge graphs offer more precise diagnostic and therapeutic support for physicians.

However, this study also reveals several critical challenges in the current development of medical data science. First, data silos remain prevalent, with insufficient data sharing and standardization among different medical institutions. Second, the ability of algorithms to generalise across diverse populations and scenarios needs improvement. Third, data privacy protection and ethical frameworks require further refinement. Finally, responsibility attribution and regulatory frameworks for AI-assisted diagnosis and treatment are not yet mature. These issues, to some extent, limit the broad application of data science in healthcare.

This study has several areas for improvement: First, the scope of literature coverage could be expanded, particularly by including more case studies from developing countries. Second, research methods could be more diversified, such as incorporating qualitative approaches like field surveys or expert interviews. Third, the exploration of emerging technologies (e.g., federated learning, generative AI) in healthcare could be more thorough.

Future research should prioritize: 1) secure, cross-institutional medical data sharing; 2) interpretable, robust AI models; 3) tailored clinical solutions; 4) ethical guidelines and regulations; 5) optimized human-AI collaboration. Post-pandemic, data science must enhance public health emergency response. Integrating 5G, IoT, and blockchain will broaden applications. Medical

DOI: 10.54254/2753-8818/2025.AU27956

institutions should digitize, enterprises must develop/validate core tech, and regulators need robust standards. Collaboration is crucial to leverage data science for improved healthcare, resource optimization, and health equity, achieving "Healthy China."

References

[1] Candeago, L., Fadda, E., Var-Sarkis, T., et al. (2022). Fine-grained dengue forecasting using telephone mobility data. Scientific Reports, 12(1), 938. https: //doi.org/10.1038/s41598-022-05237-2

[2] Zhang, H., Zhang, Y., Zhou, Y., et al. (2021). Predicting the trends of COVID-19 amidst the impacts of meteorological and traffic factors based on the LSTM model. International Journal of Environmental Research and Public Health, 18(1), 278. https: //doi.org/10.3390/ijerph180100278

[3] Gao, Z., Zhou, F., Zhang, J., et al. (2022). EpiGNN: Exploring spatial-temporal graph neural networks for epidemiological forecasting. IEEE Transactions on Big Data, 9(3), 839-851. https: //doi.org/10.1109/TBDATA.2021.3072088

[4] Rajpurohit, P., Irvin, J., Zhu, K., et al. (2018). CheXNeXt: Better chest X-ray interpretation with deep learning on chest CT. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (pp. 5123-5133). Association for Computational Linguistics.

[5] Liu, Y., Kohli, M., Kornblith, S., et al. (2018). LYNA (lymph node assistant): A deep learning system for detecting breast cancer metastases in lymph nodes. Retrieved May 16, 2024, from https: //ai.googleblog.com/2018/10/applying-deep-learning-to-metastatic.htm

[6] Meek, J., Heberle, C., Vogt, W., et al. (2021). AI-Rad companion: A novel platform providing AI-powered post-processing and reporting for medical imaging. Journal of Digital Imaging, 34(3), 621-631. https: //doi.org/10.1007/s10278-021-00434-7

[7] Zebelman, E. M., Ker, J., Perumal, S., et al. (2022). Domain generalization of deep learning-based lung nodule detection in chest computed tomography. Radiology: Artificial Intelligence, 4(3), e210214. https: //doi.org/10.1148/ryai.2022210214

[8] China Medical Equipment Association. (2021). White paper on the quality assessment of medical imaging AI products. Beijing: China Medical Equipment Association.

[9] U.S. Food and Drug Administration. (2023, October 27). Predetermined change control plans for artificial intelligence/machine learning (AI/ML)-enabled device software functions: Guidance. Retrieved September 4, 2025, from https: //www.fda.gov/regulatory-information/search-fda-guidance-documents/predetermined-change-control-plans-artificial-intelligence-machine-learning-aiml-enabled-device

[10] He, Y., Zhang, Y., An, J., et al. (2022). Health-BERT: A vision-and-language-fused model for multimodal health representation learning. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (pp. 4436-4449). Association for Computational Linguistics.