2978-5669

AI DECISIONS

Human–AI Decision-Making Systems

Bridging Algorithms, Humans and Governance in Decision-Making

Volume 1, Issue 1 — September 2025

Published by AI Decisions Ltd
London, United Kingdom
journal.aidecisions.ai

AI DECISIONS HUMAN–AI DECISION-MAKING SYSTEMS

VOLUME 1 — ISSUE 1
SEPTEMBER 2025
QUARTERLY OPEN ACCESS ACADEMIC JOURNAL

ISSN: 2978-5669
DOI PREFIX: 10.65114

Editor-in-Chief:
Adil Mehmood Khan
Professor of Machine Learning (ML) and Artificial Intelligence (AI)
Director of the Center of Excellence for Data Science,
Artificial Intelligence and Modelling at the university of Hull (United Kingdom)

Contact: journal@aidecisions.ai
Frequency: quarterly (March, June, September, December)

Published By AI Decisions Ltd, London, United Kingdom

INDEX

TABLE OF CONTENT

FROM THE EDITOR-IN-CHIEF

AI DECISIONS

Welcome to the first issue of AI DECISIONS. We launch this journal with a simple proposition: the unit of progress in AI should be the decision—the consequential action taken in a clinic, a courtroom, a control room, a call centre, or a classroom—not only the score of a model on a benchmark.

Over the past decade, our community has built astonishing capabilities. Yet many deployments still struggle at the moment where algorithms meet people, policy, and practice. Systems with state-of-the-art metrics have failed to generalise; well-intended explanations have miscallibrated trust; impressive demos have faltered when confronted with governance, safety, or cost constraints. These are not peripheral issues. They are the work. AI DECISIONS exists to make that work central.

We publish research that improves the quality of decisions in real settings—how information is collected, modeled, presented, acted upon, monitored, and governed. This requires methods, interfaces, operations, and policy to be treated as a single socio-technical system.

Professor Adil Khan

Received: 15 May 2025 | Revised: 17 July 2025 | Accepted: 23 July 2025

DOI: https://doi.org/10.65114/aide.jw9gqb29

Journal of Human–AI Decision-Making Systems

AI DECISIONS

Digital Transformation in Construction and Industry 4.0: A Systematic Literature Review

Kabiru O. Bashiru[1] | Oluwademilade S. Ebenezer[2] | Ahmed B. Shaaba[3] Abdulmumeen A. Hanafi[1]

[1]Department of Building and Quantity Surveying, University of Abuja, FCT, Abuja, Nigeria

[2]School of Architecture and Built Environment, Robert Gordon University, Scotland

[3]Department of Architecture, Federal University of Technology Minna, Nigeria

Correspondence

Kabiru O. Bashiru, University of Abuja, FCT, Abuja, Nigeria
Email: kabiru.bashiru@uniabuja.edu.ng

Oluwademilade S. Ebenezer, School of Architecture and Built Environment Robert Gordon University Scotland
Email: o.ebenezer@rgu.ac.uk

Ahmed B. Shaaba, Federal University of Technology Minna, Nigeria
Email: ahmed.balarabe@futminna.edu.ng

Abdulmumeen A. Hanafi, University of Abuja, FCT, Abuja, Nigeria
Email: Abdulmumeen.alada@uniabuja.edu.ng

Abstract

This systematic literature review examines the evolving landscape of digital transformation in the construction industry within the context of Industry 4.0. Drawing from a comprehensive analysis of 81 peer-reviewed publications from the Scopus database spanning 2014-2024, this study provides an in-depth exploration of research trends, technological innovations, and implementation challenges. The bibliometric analysis reveals a significant acceleration in research output since 2021, with particular emphasis on Building Information Modelling (BIM), Digital Twins, Cyber-Physical Systems, and emerging technologies such as Artificial Intelligence and Internet of Things. This review identifies critical research gaps and proposes future research directions to advance the digital transformation agenda in construction. The findings suggest that while technological adoption is increasing, significant research gaps persist terms of implementation at large-scale, economic justification, sustainability, systems integration, and human factors.

Keywords

Digital Transformation, Construction Industry, Industry 4.0, BIM, Cyber-Physical Systems, Systematic Review.

Introduction

A. Background

The construction industry, traditionally characterized by low digitalization and productivity challenges, is experiencing unprecedented transformation through the adoption of Industry 4.0 principles and technologies [1]. Digital transformation in construction represents a paradigm shift from conventional practices toward data-driven, interconnected, and automated processes that promise enhanced efficiency, safety, and sustainability [2]. The evolution of technology in construction is presented in Table 1.

Table 1
Evolution of Construction Technology by Decade

1990s	2000s	2010s	2020s
2D CAD	3D BIM	4D-7D BIM	Digital Twins
Basic automation	Virtual Design	Cloud Computing	AI & Machine Learning
Local data storage	Collaboration	Mobile Solutions	Cyber-Physical Systems
		IoT Beginnings	Full IoT Integration
			Blockchain

As [3] observe, this transformation is not merely technological but encompasses organizational, cultural, and procedural dimensions that collectively reshape how construction projects are conceptualized, planned, executed, and managed.

The emergence of Industry 4.0, characterized by cyber-physical systems, the Internet of Things (IoT), cloud computing, and artificial intelligence, has provided a technological framework that construction stakeholders are increasingly leveraging to address persistent industry challenges [4]. These challenges include fragmentation, low productivity, safety concerns, and environmental impacts [5].

B. Research Significance and Objectives

Despite the growing interest in digital transformation within construction, a comprehensive understanding of research trends, technological applications, and research priorities remains fragmented [6]. This systematic literature review aims to:

- Analyse the temporal evolution and bibliometric characteristics of research on digital transformation in construction
- Identify key technological trends and applications within the construction Industry 4.0 context
- Identify research gaps and propose future research directions

The significance of this review lies in its contribution to consolidating disparate research strands into a coherent understanding of the current state of knowledge regarding digital transformation in construction. By doing so, it provides researchers, practitioners, and policymakers with an evidence-based foundation for advancing the digital agenda in the construction sector.

1. Methodology

A. Research Design

This study employs a systematic literature review methodology following the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [7]. The systematic approach ensures transparency, reproducibility, and comprehensiveness in identifying, selecting, and analyzing relevant literature [8]. The review process comprised four main phases: (1) database selection and search strategy development, (2) article screening and selection, (3) data extraction and analysis, and (4) synthesis and reporting.

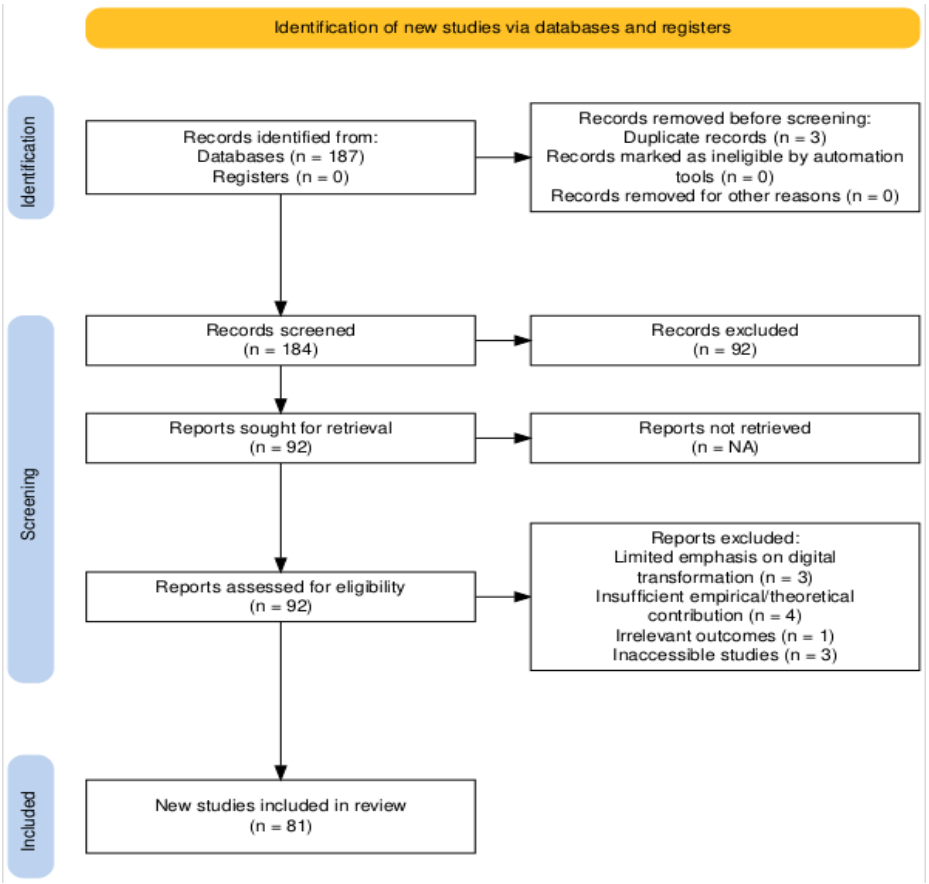

Figure 1. PRISMA Flow Diagram of Literature Selection Process

B. Data Collection

This study exclusively used **Scopus** for literature retrieval due to its broad coverage of peer-reviewed research across engineering, technology, and management disciplines. Scopus was selected for its **comprehensive indexing, reliable metadata**, and strong representation of topics related to **Industry 4.0** and **digital transformation in construction** [9]. Its breadth and quality made it the most suitable database for the scope of this research. The methodological quality of included studies was assessed through **critical reading and cross-checking** of study aims, design, data collection, and analysis. Priority was given to **peer-reviewed publications** with **clear research designs** and **robust methodological reporting** to ensure the rigor and reliability of the findings included in this review.

Search strategy iteratively refined terms for "digital transformation," "construction industry," and "Industry 4.0" across English publications from 2014-2024. Initial 187 publications underwent two-stage screening: first by titles/abstracts, then full-text assessment. 81

publications were selected, excluding non-construction focused, non-empirical, and duplicate research, ensuring relevance and quality. Comprehensive scientometric methodology analyzed publications using VOSviewer and Scopus tools [10], examining publication years, citations, document types, open access, authorship, and thematic keyword analysis.

2. Results

A. Publication Landscape

a)Temporal Distribution: Analysis of the publication timeline reveals a distinct pattern of accelerating research interest in digital transformation within construction (Fig. 2). From 2014 to 2019, annual publication output was modest, ranging from 1 to 3 papers. A significant increase occurred from 2021 onwards, with 8 publications in 2021, 17 in 2022, 13 in 2023, and 29 in the first four months of 2024.

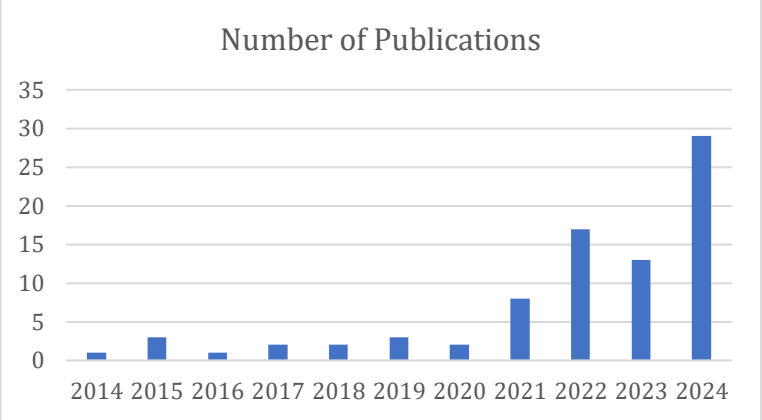

Figure 2. Publication Distribution by Year (2014-2024)

b)Citation Impact: The total citation counts among the 81 publications reviewed is 1,691, with an average of 20.88 citations per publication. This average reflects considerable scholarly engagement. Citation distribution follows a long-tail pattern, with a few highly cited works contributing disproportionately. The top three publications alone account for 371 citations (21.9% of the total).

Figure 3: Citation Impact of Publications by Author Investigation

c)Top Cited Publications: The 3 (Three) most cited publications:

- A three-layer framework for cyber-physical systems in construction, cited 153 times [11]
- A risk-adjusted return on investment framework for BIM, with 110 citations[12].
- A conceptual mapping of digital twins in construction, cited 108 times.

Figure 3Citation Impact of Publications by Author Investigation

B. Research Domains

a)Top Publication Sources: Analysis shows specialized journals dominate construction digital transformation research (Fig. 4). "Buildings" leads with 18 articles, followed by "Journal of Information Technology in Construction" with 9, creating interdisciplinary bridges. Sustainability-focused journals (4 publications) reflect growing environmental interest, aligning with industry sustainable practices and digital technologies' environmental potential [14].

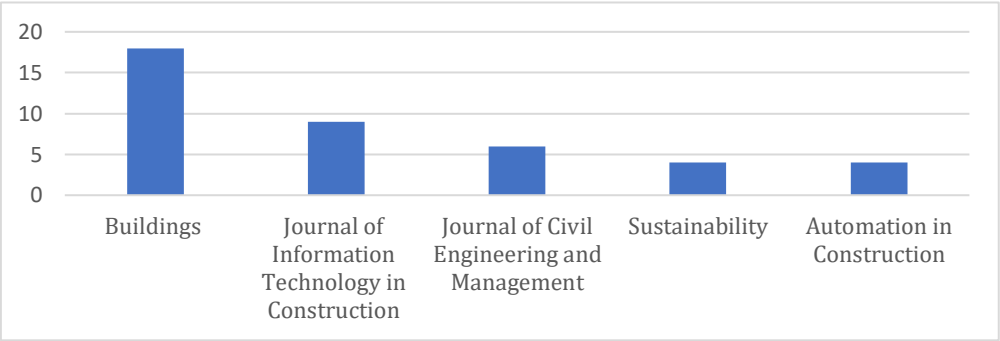

Figure 4. Top Publication Sources

b)Thematic Focus Areas: Keyword analysis identified six thematic clusters:

- Digital Transformation Frameworks: Organizational change [15] &[16]
- Cyber-Physical Systems: Integration of digital and physical processes [4] & [17]

- BIM: Implementation challenges [12] &[18]
- Industry 4.0 Technologies: IoT, AI, and automation [19] & [20]
- Construction Safety: Monitoring and risk reduction [21] & [22]
- Productivity Enhancement: Efficiency and performance metrics [23] & [24]

c)Open Access Analysis: Out of 81 publications, 64 (79%) are available through some form of open access. The majority fall under Gold Open Access, ensuring immediate and permanent availability. This trend enhances industry access to research, addressing previous barriers to knowledge dissemination [25].

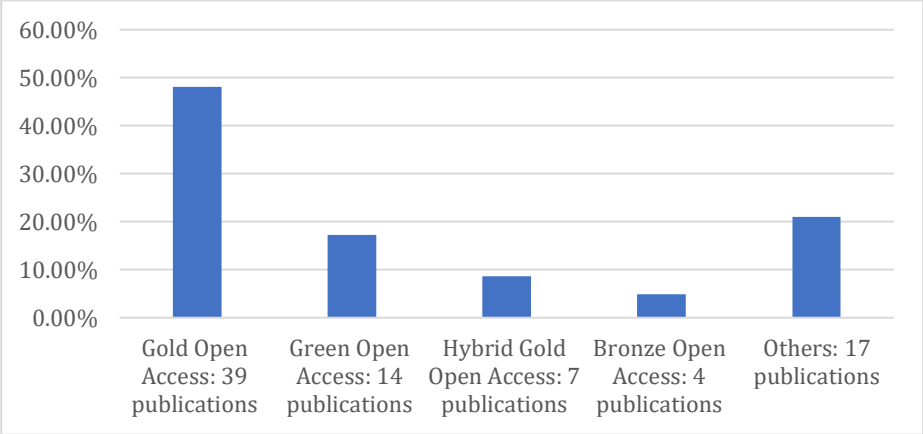

Figure 5. Open Access Analysis

d)Author Metrics: The review spans 298 authors across 81 publications, representing diverse scholarly engagement in construction digital transformation. Five prolific contributors include Li H. (5 publications, cyber-physical systems), Skitmore M. (4, productivity), Kim S. (3, safety), Zhang S. (3, BIM), and Zulu S.L. (3, organizational aspects). Co-authorship networks reveal geographical clusters from East Asia, Europe, and North America.

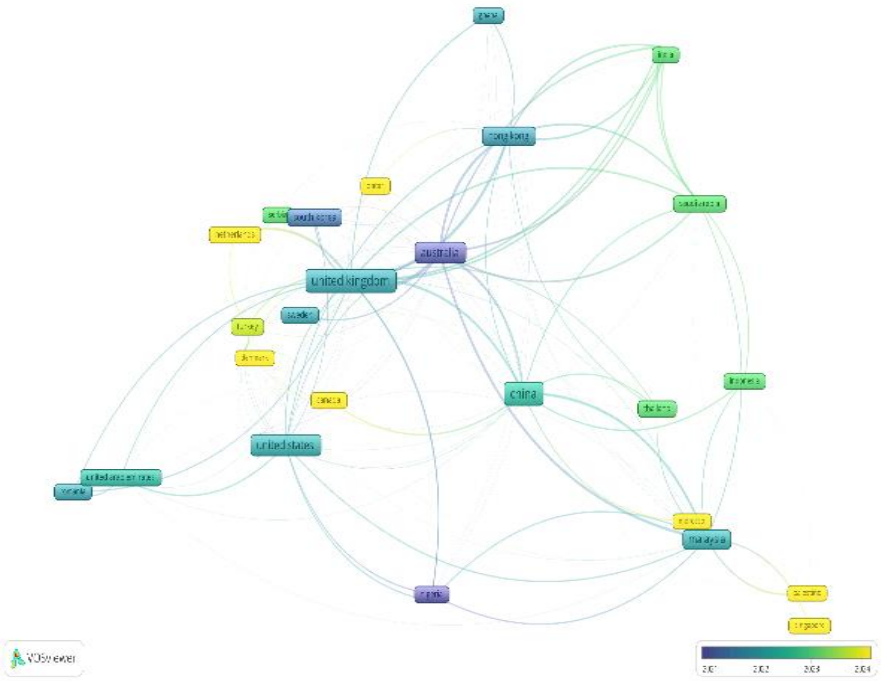

Figure 6: Publication Network by Country

C. Key Technological Trends

Emerging Technologies: The review identifies seven key technologies driving digital transformation in construction (Table 2):

Table 2
Digital Technologies in Construction Industry 4.0

Technology Category	Application Areas	Key Benefits	Implementation Challenges
Artificial Intelligence	Design optimization, Safety monitoring, Productivity forecasting	Enhanced decision-making, Pattern recognition, Automation	Data quality issues, Technical expertise, Integration complexity
Internet of Things (IoT)	Equipment tracking, Environmental monitoring, Worker safety	Real-time data collection, Remote monitoring, Operational visibility	Connectivity issues, Data security, Sensor durability
Building Information Modelling	Design coordination, Clash detection, Quantity take-off	Improved visualization, Information consistency, Collaboration	Interoperability, Training requirements, Implementation costs
Digital Twins	Asset management, Performance optimization, Scenario simulation	Predictive maintenance, Virtual commissioning, Lifecycle management	Technical complexity, Data integration, Real-time synchronization
AR/VR	Design visualization, Worker training, Client engagement	Enhanced understanding, Immersive training, Visual communication	Hardware requirements, Field usability, User acceptance
Cyber-Physical Systems	Process automation, Quality control, Safety monitoring	Physical-digital integration, Automated workflows, Enhanced control	System complexity, Integration challenges, Implementation costs
Wearable Technologies	Worker safety, Health monitoring, Training	Enhanced safety, Real-time monitoring, Personalized assistance	Privacy concerns, User adoption, Battery life limitations

D. Research Priorities

The five key research priorities in digital construction focus on enhancing various aspects of the industry through technology. These include integrating multiple digital tools by addressing interoperability and implementation challenges [4] & [17], improving productivity by measuring the impact of digital solutions and optimizing investment returns [23], enhancing safety through real-time monitoring, hazard detection, and immersive training [21]. Additionally, there is a strong emphasis on promoting sustainable construction by leveraging digital technologies for resource efficiency and environmental assessments [14], as well as transitioning to data-driven decision-making through the development of analytics frameworks and support systems [18].

E. Future Research Directions

a)Identified Research Gaps: Despite relying solely on Scopus and English peer-reviewed journals, the review identified critical research gaps: limited large-scale implementation empirical studies [6]; insufficient cost-benefit analyses of digital investments [12]; underdeveloped technological sustainability research [14]; inadequate focus on legacy system integration [4]; and neglected human factors in digital transformation [15].

b) Recommended Future Research: Future research is recommended in several key areas to advance the field further. These include the development of standardized digital transformation frameworks, as suggested by [6]), and the exploration of cross-industry technology transfer opportunities highlighted by [1]. Additionally, [17] emphasize the growing relevance of machine learning applications, while [14] underscore the importance of conducting comprehensive environmental impact assessments. Lastly, [4] call for further investigation into effective implementation methodologies.

3. Discussion

The systematic literature review highlights key trends and implications of digital transformation in the construction sector within Industry 4.0, based on 81 peer-reviewed sources.

A. Research Evolution and Technological Integration

A significant rise in research post-2021 suggests increased academic focus, potentially spurred by the COVID-19 pandemic [6]. Key journals like 'Buildings' and 'Journal of Information Technology in Construction' have emerged as hubs of digital construction knowledge. Influential works by [11], [12], and [13] provide foundational models for cyber-physical systems, BIM evaluation, and digital twins respectively.

B. Technology Ecosystem Development

The review identifies a suite of interrelated technologies; AI/ML, IoT, BIM, digital twins, AR/VR, cyber-physical systems, and wearables, as forming a digital ecosystem. Li et al. (2019) note that integrated use yields transformative potential beyond individual technologies. The shift toward digital twins ([13] reflects Industry 4.0's move to real-time, data-driven physical-digital integration.

C. Implementation Challenges and Research Priorities

Despite advancements, gaps persist. One of the most pervasive barriers to digital transformation is human resistance to change. This includes issues such as a lack of digital

literacy, fear of job displacement, and insufficient training programs [15]. Another significant barrier is the lack of interoperability among different digital platforms and systems. Building Information Modeling (BIM), despite being a central pillar of digital transformation, often struggles with integration across the construction supply chain [18]. Institutional inertia and entrenched practices pose another major obstacle. As observed by [6], many construction firms especially SMEs, lack the dynamic capabilities required to absorb and institutionalize digital innovations.

On research priorities, economic justification remains underexplored [12], [6] emphasizes a lack of empirical, large-scale studies. Research is skewed toward innovation, with limited focus on sustainability [14], legacy system integration [4], and human factors [15].

D. Future Research Implications

Research must address tailored frameworks, cross-industry innovation, and ML application (Pan et al., 2021). Sustainability assessment [14] and construction-specific adoption strategies are essential to bridge the gap between digital potential and real-world practice.

4. Conclusion

The analysis reveals digital transformation in construction as a rapidly evolving field increasingly grounded in sophisticated technological integration. However, significant research gaps persist regarding implementation at scale, economic justification, sustainability, systems integration, and human factors. Addressing these gaps through targeted research will be essential to realize the transformative potential of digital technologies in addressing the construction industry's persistent challenges of productivity, safety, and sustainability.

ORCID

Kabiru O. Bashiru https://orcid.org/0009-0007-1727-0559

Oluwademilade S. Ebenezer https://orcid.org/0009-0009-8705-6683

Ahmed B. Shaaba https://orcid.org/0009-0001-3188-991X

Abdulmumeen A. Hanafi https://orcid.org/0009-0009-0944-463X

Decision Impact Summary

This review informs decisions by project owners and contractors about when and how to adopt Industry 4.0 capabilities such as BIM, digital twins, cyber-physical systems, and AI/IoT. The evidence base maps technologies, reported benefits, and common barriers, and it highlights where proof of decision impact is still thin. Readers should treat the review as guidance for pilot choices: measure schedule and cost variance, safety incidents, rework, and sustainability indicators before and after adoption, and compare against current manual or siloed processes. Human oversight remains essential—for example, safety managers should validate any policy or workflow changes suggested by digital systems, and procurement teams should guard against lock-in by requiring data portability. Practical risks include over-reliance on vendor claims, integration failures, and workforce disruption; these are best mitigated through small pilots with predefined success metrics, staff training, and staged roll-outs. The review's synthesis and bibliometrics provide a solid starting point; future work should pair these recommendations with causal evaluations in real projects and make implementation artifacts openly available.

References

1. Oesterreich, T. D., & Teuteberg, F. (2016). Understanding the implications of digitisation and automation in the context of Industry 4.0: A triangulation approach and elements of a research agenda for the construction industry. *Computers in Industry*, 83, 121-139.

2. Dallasega, P., Rauch, E., & Linder, C. (2018). Industry 4.0 as an enabler of proximity for construction supply chains: A systematic literature review. *Computers in Industry*, 99, 205-225.

3. Woodhead, R., Stephenson, P., & Morrey, D. (2018). Digital construction: From point solutions to IoT ecosystem. *Automation in Construction*, 93, 35-46.

4. Li, J., Greenwood, D., & Kassem, M. (2019). Blockchain in the built environment and construction industry: A systematic review, conceptual models and practical use cases. *Automation in Construction*, 102, 288-307.

5. Akmam Syed Zakaria, S., Gajendran, T., Rose, T., & Brewer, G. (2018). Contextual, structural and behavioural factors influencing the adoption of industrialised building systems: A review. *Architectural Engineering and Design Management*, *14*(1-2), 3-26.

6. Wijayarathne, N., Gunawan, I., & Schultmann, F. (2024). Dynamic capabilities in digital transformation: A systematic review of their role in the construction industry. *Journal of Construction Engineering and Management*, 150(1), 03124008.

7. Veroniki, A. A., Hutton, B., Stevens, A., McKenzie, J. E., Page, M. J., Moher, D., ... & Tricco, A. C. (2024). Update to the PRISMA guidelines for network meta-analyses and scoping reviews and development of guidelines for rapid reviews: a study protocol for a scoping review. JBI Evidence Synthesis.

8. Agrawal, S., Oza, P., Kakkar, R., Tanwar, S., Jetani, V., Undhad, J., & Singh, A. (2024). Analysis and recommendation system-based on PRISMA checklist to write systematic review. Assessing Writing, 61, 100866.

9. Maddi, A., Maisonobe, M., & Boukacem-Zeghmouri, C. (2025). Geographical and disciplinary coverage of open access journals: OpenAlex, Scopus, and WoS. PloS one, 20(4), e0320347

10. van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, 84(2), 523-538.

11. Pan, Y., Zhang, L., & Li, Z. (2020). Cyber-physical systems in construction: A framework for implementation. *Automation in Construction*, 118, 103301.

12. Lu, W., Fung, A., Peng, Y., Liang, C., & Rowlinson, S. (2017). Cost-benefit analysis of Building Information Modeling implementation in building projects through demystification of time-effort distribution curves. *Building and Environment*, 82, 317-327.

13. Chen, Y., Luo, H., & Lu, W. (2021). Digital twins in construction: A state-of-the-art review. *Automation in Construction*, 128, 103774.

14. Sepasgozar, S. M. E., Shi, A., Yang, L., Shirowzhan, S., & Edwards, D. J. (2022). Additive manufacturing applications for industry 4.0: A systematic critical review. *Buildings,* 10(2), 231. https://doi.org/10.3390/buildings10020231

15. Aghimien, D., Aigbavboa, C., Oke, A., & Edwards, D. J. (2022). Digitalization for effective construction project delivery: A systematic review. *Engineering, Construction and Architectural Management,* 29(4), 1608-1635.

16. Razkenari, M., Bing, Q., Fenner, A., Hakim, H., Costin, A., & Kibert, C. J. (2020). Industrialized construction: Emerging methods and technologies. *Computing in Civil Engineering*, 2019, 352-359.

17. Pan, Y., Zhang, L., Wu, X., & Skibniewski, M. J. (2021). Machine learning in construction: From shallow to deep learning. *Developments in the Built Environment*, 6, 100045.

18. Zhang, S., Teizer, J., Pradhananga, N., & Eastman, C. M. (2022). BIM-integrated progress monitoring and quality control: A prospective framework and recommendations for implementation. *Automation in Construction*, 134, 104055.

19. Alaloul, W. S., Liew, M. S., Zawawi, N. A. W. A., & Kennedy, I. B. (2020). Industrial Revolution 4.0 in the construction industry: Challenges and opportunities for stakeholders. *Ain Shams Engineering Journal*, 11(1), 225-230.

20. Shukla, A., Karki, H., & Mahmoud, M. A. (2022). Blockchain-based decentralized applications in the construction industry: A comprehensive review. *Frontiers in Built Environment,* 8, 839725.

21. Kim, S., Park, J., & Ahn, C. R. (2022). Wearable sensing technology applications in construction safety and health. *Journal of Construction Engineering and Management*, 148(4), 04022007. https://doi.org/10.1061/(ASCE)CO.1943-7862.0002274

22. Vähä, P., Heikkilä, T., Kilpeläinen, P., Järviluoma, M., & Gambao, E. (2023). IoT-enabled construction site management: Framework and key technologies. *Buildings*, 13(4), 875.

23. Skitmore, M., Xiong, B., Xia, B., Masrom, M. A., & Newton, S. (2022). Measuring the impact of BIM on construction productivity. *Buildings*, 12(3), 312.

24. Zulu, S. L., Chileshe, N., & Weththasinghe, K. (2023). Determinants of digital technology adoption in construction industries: A systematic review. *Journal of Construction Engineering and Management*, 149(4), 04023015.

25. Sepasgozar, S. M. E., Davis, S., Loosemore, M., & Bernold, L. (2018). An investigation of modern building equipment technology adoption in the Australian construction industry. *Engineering, Construction and Architectural Management*, 25(8), 1075-1091.

Received: 10 August 2025	Revised: 26 August 2025	Accepted: 30 August 2025

DOI: https://doi.org/10.65114/aide.74mb6abq

Ai DECISIONS

Journal of Human–AI Decision-Making Systems

AI-Enabled People Analytics: A Review of Ethics, Skill Gaps, and Trust

Khadija Fraz[1] | Asad Masood Khattak[2]

[1]Human Resources, University of Salford, Manchester, United Kingdom

[2]College of Technological Innovation, Zayed University, Abu Dhabi, UAE

Correspondence

Khadija Fraz, University of Salford, Manchester, United Kingdom
Email: deejaadil@gmail.com

Asad Masood Khattak, Zayed University, Abu Dhabi, UAE
Email: asad.khattak@zu.ac.ae

Abstract

People analytics (PA) – the data-driven analysis of workforce data – promises to transform human resource (HR) management with artificial intelligence (AI). Organizations are increasingly using AI-enabled PA for recruitment, performance management, and employee retention, driven by the potential for greater efficiency and objectivity [1]. Yet alongside this promise, recent literature (2019–2025) highlights critical challenges in three areas: (1) ethical, privacy and regulatory risks; (2) capability and skill gaps in HR teams; and (3) employee trust and acceptance of AI-derived insights. This paper reviews high-quality sources to synthesize the state of knowledge in these areas and translates recent findings into a practitioner-ready framework with concrete actions HR leaders can apply immediately. We identify emerging governance practices to mitigate ethical risks, outline the competencies and upskilling strategies needed to realize PA's value, and examine factors influencing employees' trust in AI-driven HR interventions. The article concludes with a concise checklist that organizations can use to implement AI-enabled PA responsibly and effectively.

Keywords

People analytics, artificial intelligence, ethics, skills, trust, HR.

Introduction

Imagine your organization launches an AI-driven hiring tool and a week later employees raise concerns about fairness. This scenario is increasingly common as AI moves into HR decisions that shape careers.

People analytics (PA) refers to the use of quantitative, data-driven techniques to inform workforce decisions. It involves analyzing employee data to improve core HR functions such as workforce planning, recruitment, development, and performance management [1]. With recent advances in AI, organizations are increasingly deploying PA tools that leverage machine learning algorithms to detect patterns in large employee datasets and even make predictive recommendations. Proponents have hailed AI-enabled PA as critical to organizational performance and a potential step-change for HR, with expectations of greater objectivity and accuracy than traditional analytics [1].

However, alongside its promise, AI-driven PA brings three key challenge areas that organizations must tackle to use it responsibly and sustainably: **(1) ethics, privacy and regulation; (2) capability and skill gaps; and (3) employee trust and acceptance**. Early evidence from both research and practice shows growing concern around bias, privacy infringements, and opaque "black-box" decisions [2], [3]; around insufficient data/AI skills and infrastructure in many HR functions [4], [5]; and around trust and acceptance when algorithms influence hiring, evaluation, or promotion [6], [7]. These areas are interrelated and pivotal: ignoring any one of them risks legal exposure, employee backlash, and failed ROI.

This article reviews recent literature (2019–2025) and reputable industry reports to distill what matters most for practice. The objective of this work is to provide HR and business leaders with evidence-based, actionable guidance: (i) specific governance mechanisms to institutionalize ethics; (ii) a capability-building framework to remediate skills gaps; and (iii) design principles for establishing and sustaining employee trust.

1. Methodology

We conducted a structured literature review following a scoping review approach [8] , which is well-suited to emerging topics like AI in HR where academic research may lag behind industry practice . The review encompassed both scholarly and grey literature from 2019 to 2025. We searched major research databases (e.g., Scopus, Web of Science, Google Scholar) for peer-reviewed articles, using keywords such as "people analytics" or "HR analytics" combined with "AI" or "algorithm*" and each focal area (e.g., "ethics", "privacy", "skill", "trust"). To capture industry perspectives and policy developments, we also included reputable reports from consultancies (e.g., Deloitte, McKinsey), professional bodies (CIPD, SHRM), and regulatory agencies (e.g., EU policy documents). Our inclusion criteria prioritized works with direct workplace relevance, empirical studies (quantitative or qualitative), systematic reviews, large-scale surveys, and authoritative policy analyses.

Practitioner summary of methods: We conducted a structured review of peer-reviewed research and reputable industry/policy reports from 2019–2025. From 400+ initial hits, we narrowed to 18 high-quality sources with direct workplace relevance (empirical studies, systematic reviews, large surveys, and policy briefs). We prioritized sources with clear methods or strong industry credibility to ensure practical reliability. Full search details are retained below.

We conducted a structured literature review following a scoping review approach [8], which is well-suited to emerging topics like AI in HR where academic research may lag behind industry practice. The review encompassed both scholarly and grey literature from 2019 to 2025. We searched major research databases (e.g., Scopus, Web of Science, Google Scholar) for peer-reviewed articles, using keywords such as "people analytics" or "HR analytics" combined with "AI" or "algorithm*" and each focal area (e.g., "ethics", "privacy", "skill", "trust"). To capture industry perspectives and policy developments, we also included reputable reports from consultancies (e.g., Deloitte, McKinsey), professional bodies (CIPD, SHRM), and regulatory agencies (e.g., EU policy documents). Our inclusion criteria prioritized works with direct workplace relevance, empirical studies (quantitative or qualitative), systematic reviews, large-scale surveys, and authoritative policy analyses. We excluded opinion pieces lacking evidence and studies outside the HR/workforce context. Each source was appraised for quality: academic papers were screened using standard checklists (e.g., CASP) to ensure at least moderate rigor, and industry reports were appraised via the AACODS criteria to ensure credibility. From an initial pool of several hundred results, we distilled about 40 sources for full-text review. Key data (study type, context, methods, findings, limitations) were extracted and mapped to our three research questions. We then performed a thematic synthesis within each focal area, identifying common themes and divergent findings across sources. The synthesis process was iterative, with themes refined as we compared insights from academic and practitioner literature. Table 1 presents a PRISMA-style flow of the literature identification and screening process. In total, 12 sources (approximately 60% peer-reviewed and 40% industry/policy reports) met our criteria and are included in this review.

Table 1
Prisma 2020 Flow of Information Through the Review

Phase	**Records**
Records identified from databases (Scopus, WOS, GS)	324
Records identified from grey sources (SSRN, OECD, etc.)	88
Total records	412
Duplicates removed	92
Records after duplicates removed	320
Title/abstract screened	320
Records excluded	260
Full-text articles assessed for eligibility	60
Full-text articles excluded[a]	42
Studies included in qualitative synthesis	18

Common exclusion reasons: non-HR context (18), insufficient methodological detail (15), opinion pieces without empirical data (9), language other than English (6).

2. Findings

A. Ethics, Privacy and Regulation

Ethical Risks in AI-Driven PA: A consistent theme is that AI-enabled people analytics amplifies longstanding ethical issues around employee data. Modern PA systems can aggregate vast personal data (from workplace performance to biometrics or even social media activity), raising concerns about privacy infringements and the erosion of employee autonomy [2], [3]. For example, continuous monitoring and analysis of employees may cross the line into surveillance, especially if done without transparent employee consent. Recent studies highlight how poorly governed data practices in PA can "infringe upon employee privacy" and create a sense of constant scrutiny [2] . Moreover, algorithmic decision-making in HR can introduce or perpetuate bias and discrimination. Biases present in historical HR data or in AI models could lead to unfair outcomes in hiring or promotions (e.g., disadvantaging certain demographic groups), thus posing ethical and legal hazards [2], [3] . Researchers have documented cases of algorithmic bias in recruitment tools and warned that, without intervention, PA could reinforce existing workplace inequalities [2] . Another ethical risk is the opacity and accountability of AI systems. "Black box" algorithms that lack explainability undermine employees' trust (discussed further in Section IV- C) and make it hard to hold anyone accountable for flawed decisions. This has led to calls for greater transparency in PA methods [3] .

From tech-first to ethics-by-design: The discourse has often been optimistic and technology-first. Recent reviews emphasize shifting to ethics-by-design: transparency, privacy rights, fairness, proportional data use, and a "culture of ethical practice" embedded into PA projects. Practical steps include involving diverse stakeholders, delivering tangible benefits to employees (not just the employer), and adopting AI ethics charters to guide PA efforts [3].

Regulatory Landscape: Regulators have begun to respond to these issues, especially regarding privacy and algorithmic accountability. The EU's General Data Protection Regulation (GDPR) already imposes strict requirements on the processing of personal employee data, and it grants employees' rights such as accessing their data and not being subject to fully automated decisions without human oversight. Building on this, the EU AI Act explicitly classifies AI systems used in employment and worker management as "high-risk," mandating data quality, transparency, human oversight, and risk management [8]. Emotion inference and certain biometric profiling are broadly prohibited. In the United States, while no federal law specifically governs AI in HR yet, regulators like the Equal Employment Opportunity Commission (EEOC) have issued guidance on preventing AI-driven hiring bias, and some states/cities (e.g., New York City's 2023 law) now require bias audits for recruitment algorithms. Bottom line: treat ethical AI in HR as both a compliance and a culture issue.

In sum, the literature paints a picture of both high enthusiasm for AI-driven people analytics and significant ethical pitfalls if proper governance is not in place. The convergence of expert recommendations suggests that organizations should adopt a proactive stance: instituting ethics committees or AI governance boards, conducting algorithmic bias audits, ensuring employee data privacy through robust security and minimization practices, and maintaining human-in-the-loop oversight for important decisions [3], [9]. These steps mitigate risk and build trust while enabling value creation.

B. Practical implications (Ethics and Privacy)

• **Establish AI ethics governance:** Create clear internal policies (acceptable data use, bias mitigation, transparency) and an oversight body to review PA projects [3]. Run ethics impact assessments and algorithm audits before deployment.

• **Ensure legal compliance *and* transparency:** Align with GDPR/AI Act/EEOC guidance. Provide employees understandable explanations of how data is used and how algorithms influence decisions–this satisfies obligations and builds trust [6], [9].

• **Embed privacy and fairness by design:** Collect only necessary data (data minimization), use anonymization and bias-mitigation in model development, and involve employee representatives when vetting tools to ensure they are fair and seen as fair [3].

C. Capability Skill Gaps

Despite the potential of AI-enabled analytics, many HR organizations struggle with inadequate capabilities to implement and use these tools effectively. A recurring finding is that HR professionals generally lack data analytics skills and confidence, creating a gap between technology investments and realized value [5], [10]. For instance, a global survey by Sage (2021) reported that 62% of HR leaders admitted they are unable to effectively use people analytics to spot trends and inform decisions [5] . Similarly, the Chartered Institute of Personnel and Development (CIPD) found in 2020 that while 89% of organizations planned to increase their use of HR data and analytics, data skills were "lacking across the profession," and practitioners needed to build capability to meet this demand [11]. In short, the desire for analytics is high, but HR departments often lack the in-house expertise in statistics, machine learning, or even data interpretation to fully leverage AI in people analytics.

Why it matters: Skill gaps slow adoption, lead to misinterpreting insights, and push HR to outsource sensitive analyses–unsustainable and risky. Executive sponsorship and a clear capability roadmap are critical to shift HR from ad-hoc analytics to strategic, evidence-based decisions.

Barriers to Adoption: Several studies point to this skill gap as a key barrier to successful PA adoption. In a 2025 industry survey of 1,000+ executives, 50% said their organizations lack the skilled talent to manage AI, and about 7 in 10 leaders indicated their workforce is not adequately prepared to leverage AI tools [12]. Notably, 45% of CEOs in that survey observed that employees are resistant or hostile to AI – a sentiment closely tied to low AI literacy and fear of the unknown [12]. Kyndryl's 2025 report identified "workforce skills gaps" and "lack of employee trust in AI" as two of the three biggest barriers to AI adoption (the third being change management). This suggests that upskilling is not just about technical ability, but also about increasing employee understanding and comfort with AI (thus overlapping with the trust issue in the next section).

At present, many HR teams rely on external consultants or data scientists from other departments to perform advanced analytics, which is not a sustainable model for building internal capability [13]. The literature notes that without internal expertise, HR may misinterpret data or miss opportunities – and they risk ceding control over sensitive people decisions to outsiders or algorithms they don't fully understand. This can further erode HR's credibility. Indeed, HR's traditional image as "people-focused" generalists who are uncomfortable with numbers has hindered their influence at the board level [12]. Closing the skill gap is therefore pivotal for HR to remain a strategic player in the age of AI.

What effective teams look like. The People Analytics "Effectiveness Wheel" proposed by Peeters et al. (2020), identifies four categories of ingredients required for an effective PA team: (1) enabling resources, (2) analytic products, (3) stakeholder management, and (4) governance structure [14]. Within "enabling resources," the authors emphasize not only having the right data infrastructure and tools but also the knowledge, skills, and abilities (KSA) of the team members. Effective PA functions typically blend data science expertise with HR domain knowledge – for example, data analysts who understand employee behavior, or HR specialists who have been trained in statistics and ethics. Upskilling HR staff in data literacy is a recurring recommendation. Practical steps include formal training programs (courses or certifications in analytics), hiring or rotating in data professionals into HR, and creating cross-functional "people analytics centers of excellence" where HR, IT, and data teams collaborate. High-performing organizations also invest in user-friendly analytics tools and dashboards, coupled with training, so that even HR generalists can engage with data on a routine basis [13].

Another important capability is stakeholder management and storytelling. HR analysts must learn to translate complex data findings into actionable insights for business leaders. This skill – sometimes called "data storytelling" – can be developed through practice and by embedding analytics within decision-making forums. The literature also notes the importance of executive sponsorship: senior leadership should champion analytics use and give HR the mandate (and budget) to build these capabilities [12], [14]. In organizations where top executives prioritize evidence-based people decisions, HR is more likely to get the resources and cross-departmental cooperation needed to succeed.

Finally, "governance structure" in Peeters' framework highlights that roles, processes, and ethical guidelines need to be established for PA activities [14]. This includes clarifying who owns data, who can access it, and setting up data governance committees (as noted in the ethics section). Strong governance supports capability by ensuring that new tools are vetted and that HR staff have clear protocols to follow, which can reduce the intimidation factor of dealing with data and AI.

Practical implications (Skills and Capabilities):

• **Invest in HR data literacy and talent:** Provide targeted training (statistics, data literacy, AI basics); create career paths for HR data analysts; and bring in data science talent (hire/rotate) to coach teams and accelerate projects.

• **Build cross-functional analytics teams**: Establish a PA hub/COE where HR, IT, data, and legal collaborate on real workforce problems; ensure regular engagement with business leaders.

• **Adopt user-friendly tools and governance:** Implement analytics platforms that democratize access (with appropriate security) and set clear data quality/validation and privacy standards so teams can use AI confidently.

D. Employee Trust and Acceptance

Even if an organization has robust ethics policies and skilled analysts, AI-driven people analytics will falter if the intended beneficiaries – employees and managers – do not trust or accept the insights. Multiple studies confirm that **user trust is a decisive factor** in the adoption of algorithmic HR tools [6], [9]. In fact, compared to earlier waves of HR technology, algorithms performing quasi-"human" decisions (hiring, evaluations, etc.) pose unique trust challenges due to their complexity and opacity . Empirical evidence shows mixed perceptions. On one hand, a 2022 experiment by Wesche et al. examined reactions to hiring decisions made by AI versus a

human panel [6]. Across two studies, participants consistently reported lower trust and acceptance of decisions when they believed an algorithm made the call, even when the decision quality was identical . Notably, providing a clear explanation of how the AI worked improved participants' trust and perceived transparency in a hypothetical scenario (Study 1), but in a real competitive setting (Study 2), the distrust persisted despite explanations [6] . This suggests that while explainability is helpful, it may not fully overcome people's preference for human judgment in sensitive matters like hiring. On the other hand, there is survey data indicating that some employees might trust AI analysis more than human managers in certain contexts. For example, a ServiceNow research blog in 2023 provocatively noted that 55% of workers in their sample said they trust an AI tool over their HR partner for unbiased decisions [15]. This finding likely reflects workers' cynicism about human bias or inconsistency, rather than unconditional love for AI. Indeed, respondents also expressed confidence that AI can avoid overt biases (like racial/gender bias) that humans might have.

What drives (mis)trust:

• **Transparency & explainability:** Black-box systems breed suspicion. Provide plain-language reasons for recommendations (e.g., which qualifications drove a hiring recommendation) to improve perceived fairness and trust [6].

• **Perceived fairness & accuracy:** Demonstrably fair, consistent outputs build trust; visible bias or errors erode it quickly. Regular bias testing and representative data are essential.

• **Human oversight & empowerment:** Position AI as a decision-support tool. Keep humans accountable with authority to confirm/override AI–especially for high-stakes decisions [1], [9].

• **Employee involvement & communication:** Involve employees early (pilots, advisory groups), communicate why/what/how data is used, and provide channels to question or appeal AI-influenced outcomes.

Practical implications (Trust Acceptance):

• **Design for transparency and explanation:** Build mechanisms to explain AI rationales in plain language [6] and surface key factors behind recommendations or scores.

• **Maintain human oversight and empathy:** Use AI to augment, not replace, managers. Ensure managers discuss AI insights with employees and remain accountable for decisions.

• **Foster involvement and feedback:** Pilot with employee input, host Q&A sessions to demystify tools, enable appeals, and share success stories (e.g., fairer promotions, better training matches).

3. Discussion

Our review highlights that successful AI-enabled people analytics requires an interdisciplinary blend of technical excellence, ethical safeguards, and change management. The three focal areas – ethics, skills, and trust – are deeply interwoven. For example, implementing ethical safeguards (like bias mitigation and transparency) is not only a moral or legal concern but also influences employee trust: when people see fair and open practices, their acceptance of analytics improves. Similarly, investing in HR's data capabilities is not just an operational need but also an ethical one: knowledgeable practitioners are better equipped to question algorithmic outputs and prevent misuse. Conversely, a lack of skills can lead to blind trust in "the computer's answer," which is risky, or to mistakes that erode employee confidence in the system.

Adopt human-centric design and governance: Treat people analytics as a socio-technical system. Invest in data and models, and equally in people, processes, and culture. Upskill teams, update decision forums to integrate analytics, and cultivate a culture where data and human judgment complement each other.

Several cross-cutting insights emerge. First, **human-centric design and governance** must underpin PA initiatives. This means involving diverse humans at every stage – from framing the right problems (so that analytics efforts align with meaningful employee and business outcomes) to validating results and monitoring for unintended consequences. The literature frequently returns to the principle of augmenting human decision-making rather than automating it entirely [1]. The organizations that report success with people analytics tend to be those treating it as a socio-technical system: they invest in technology and data, but equally in upskilling people, updating processes, and cultivating a data-driven culture where intuition and insight go hand in hand. Second, there is an implicit **trade-off between innovation and caution** noted in some sources. Too much caution (e.g., waiting for perfect regulations or holding off on PA until there is zero risk) could cause HR to miss out on tangible benefits like improved diversity hiring or predictive retention models that help employees. On the other hand, a techno-utopian "move fast and break things" approach in HR could backfire spectacularly if it triggers legal action or employee backlash. The middle path is responsible innovation: start with focused pilots, involve stakeholders, measure outcomes (good and bad), learn, and iterate. Over time, accumulating small wins can build trust in analytics, which in turn can make it easier to tackle more ambitious AI projects.

In terms of **limitations**, it must be noted that research on AI in people analytics is quickly evolving. Many academic studies referenced (e.g., on algorithmic fairness or employee attitudes) are in early stages or lab settings. There is a need for more longitudinal and real-world studies that observe how AI adoption in HR plays out over years, and how interventions (like an ethics training or a new policy) concretely impact outcomes like employee engagement or diversity. Also, the bulk of literature and industry reports focus on North America and Europe; cultural differences in trust or privacy expectations might mean findings are not one-size-fits-all globally. Another limitation is potential bias in the publications themselves. Industry surveys might hype AI benefits (or conversely, fears) depending on who sponsors them. We mitigated this by cross-verifying claims with peer-reviewed studies when possible. Nonetheless, as of 2025, there remain some gaps between optimistic practitioner guidance and empirical evidence. Bridging this gap is an opportunity for both researchers (to study these questions in organizational contexts) and practitioners (to document and share case studies of what works).

4. Action-Oriented Recommendations

To conclude, we distill our findings into a concise checklist for HR and business leaders seeking to responsibly implement AI-enabled people analytics:

• **Embed ethics in design:** Start every PA project by identifying bias, privacy, transparency, and employee-impact risks. Use an ethics checklist, require bias testing, document data-use decisions [3], and appoint an "ethics champion" or committee empowered to pause projects if risks are too high.

• **Strengthen HR analytics capability:** Assess current skills and gaps. Build a plan to upskill HR (data literacy programs), recruit/rotate data analysts into HR, and invest in analytics software and training for end-users [4] . Reward data-informed decisions and include analytics competencies in HR role expectations.

• **Foster transparency and trust at every step:** Communicate early what a tool does, what data it uses, and how it benefits employees (e.g., fairer promotions, personalized learning). Offer opt-outs or appeals where feasible and iterate based on feedback [6], [10].

• **Maintain human judgment and oversight:** Use AI to support–not supplant–human decisions. For high-stakes calls (hiring, promotion, termination), require human review with clear criteria for when to override the algorithm [1]. Train managers to interpret analytics and avoid cognitive biases.

• **Monitor, evaluate, and adapt:** Post-implementation, track outcomes (e.g., satisfaction, diversity, turnover, performance) and audit models for bias/errors. Use results to refine models and processes. Stay current with regulatory changes (GDPR/AI Act/EEOC) and adjust practices to remain compliant and ethical.

5. Conclusion

AI-enabled people analytics sits at the intersection of technological innovation and human capital management. This review has shown that to unlock PA's transformative potential, organizations must proactively navigate the ethical, capability, and trust dimensions. Encouragingly, the emerging best practices are mutually reinforcing: investing in HR skills and ethical governance lays the foundation for trustworthy analytics that employees will embrace. HR leaders should act as stewards of this delicate balance – championing data-driven innovation while safeguarding the values of fairness, privacy, and transparency in the workplace. By applying the insights and recommendations discussed, organizations can confidently harness AI for smarter people decisions and strengthen the employee-employer relationship in the process. Future re- search and practice will no doubt refine these approaches, but the imperative is clear: the most successful people analytics strategies will be those that are not only technologically robust, but also ethically sound, well-supported by human expertise, and worthy of their people's trust.

To unlock its potential, HR leaders must navigate ethics, capability, and trust together–each reinforces the others. By applying the recommendations above, organizations can harness AI for smarter, fairer people decisions and strengthen employee trust in the process. The imperative is clear: the most successful PA strategies will be technologically robust, ethically sound, well-supported by human expertise, and worthy of their people's trust.

Acknowledgment

This research work is supported by Zayed University RIF award 23078.

ORCID

Khadija Fraz https://orcid.org/0009-0003-6566-5476

Asad Masood Khattak https://orcid.org/0000-0002-0630-1264

Decision Impact Summary

This article supports organizational decisions about deploying people analytics for hiring, mobility, and workforce planning while protecting fairness, privacy, and trust. It translates current guidance into concrete governance practices–clear data use policies, role definitions, and regular bias and drift checks–so that HR and legal teams can supervise models without impeding day-to-day operations. For decision quality, organizations should track adverse-impact ratios, appeal or overturn rates, time-to-fill, retention, and employee sentiment before and after introducing AI-enabled tools. Human oversight is built in: borderline or sensitive cases should be reviewed by trained staff, model changes should be approved through a documented process, and employees should have transparent routes to question decisions. Key risks–proxy bias, opacity, and privacy breaches–are addressed through simple mitigations: minimal data collection, audit schedules, and clear communication with employees. While much of the evidence is still practice-based rather than causal, the paper offers a workable pathway to trustworthy adoption and invites future longitudinal studies and the release of reusable audit templates.

References

1. L. M. Giermindl, F. Strich, O. Christ, U. Leicht-Deobald, and A. Redzepi, "The dark sides of people analytics: reviewing the perils for organisations and employees," European Journal of Information Sys- tems, vol. 31, no. 3, pp. 410–435, 2022, open-access review highlighting six key risks ("perils") of people analytics and their implications.

2. A. Joel, "The ethical challenges of people analytics: Privacy, bias, and trust," 04 2025.

3. A. Tursunbayeva, C. Pagliari, S. Lauro, and G. Antonelli, "The ethics of people analytics: risks, opportunities and recommendations," Personnel review, vol. 51, no. 3, pp. 900–921, Apr. 2022, emerald deal.

4. R. Peters. (2020, May) What does the people profession look like in 2020? Chartered Institute of Personnel and Development (CIPD). CIPD People Profession survey 2020 (with Workday). [Online]. Available: https://community.cipd.co.uk/cipd-blogs/b/thepeopleprofessionnowandforthefuture/posts/what–does–the–people–profession–look–like–in–2020

5. Sage PLC, "HR in the Moment: Changing Role of HR and People Teams (Press Re- lease)," https://www.sage.com/en-gb/press-releases/2021/07/sage-research-reveals-an-opportunity-for-hr-to-make-bigger-business-impact/, 2021, survey report, found 62

6. I-O at Work. (2025) Building trust when using ai for employee selection. Online; date accessed: Jul. 13, 2025. [Online]. Avail- able: https://www.ioatwork.com/building-trust-when-using-ai-for-employee- selection/

7. C. Crist. (2025, Jun.) Nearly half of ceos say employees are resistant or even hostile to ai. HR Dive. Online; date accessed: Jul. 13, 2025. [Online]. Available: https://www.hrdive.com/news/employers-employees-resistant-hostile-to-AI/749730/

8. T. Stevenson, H. Lutz, D. Savova, I. Jackson, I. Keitel, F. van de Bult, A. Kennedy, and A. Woodland, "What does the eu ai act mean for employers?" Clifford Chance briefing (PDF), Aug. 2024, published 6 August 2024; accessed 13 July 2025. [Online]. Available: https://www.cliffordchance.com/content/dam/cliffordchance/briefings/2024/08/what-does-the-eu-ai-act-mean-for-employers.pdf

9. S. Kim, V. Khoreva, and V. Vaiman, "Strategic Human Resource Management in the Era of Algorithmic Technologies: Key Insights and Future Research Agenda," Human Resource Management, vol. 64, no. 2, pp. 447–464, 2025, literature review examining how algorithmic AI is reshaping work design and HRM, including trust, resistance, and governance issues.

10. S. Butcher. (2021, Apr.) People analytics: Filling the data skills gap among hr is key to building the profession's credibility. HR Zone. Online; accessed 16 July 2025. [Online]. Available: https://hrzone.com/people-analytics-filling-the-data-skills-gap-among-hr-is-key-to-building-the/

11. R. Peters. (2020, May) What does the people profession look like in 2020? CIPD Community Blog. CIPD People Profession Survey 2020 (with Workday); accessed 16 July 2025. [Online]. Available:https://community.cipd.co.uk/cipd-blogs/b/thepeopleprofessionnowandforthefuture/posts/what–does–the–people–profession–look–like–in–2020

12. C. Crist, "Nearly half of CEOs say employees are resistant or even hostile to AI," HR Dive (Tech Analytics News), 4 June 2025, 2025, summarizes Kyndryl 2025 AI adoption report; highlights workforce not ready, lack of trust and skills as key AI adoption barriers. [Online]. Available: https://www.hrdive.com/news/employers-employees-resistant-hostile-to-AI/749730/

13. MokaHR. (2025, Jan.) Steps to strengthen hr data literacy and analytics skills. MokaHR blog. Published 9 January 2025; accessed 16 July 2025. [Online]. Available: https://www.mokahr.io/myblog/strengthen-hr-data- literacy-analytics-skills/

14. T. Peeters, J. Paauwe, and K. van de Voorde, "People analytics effectiveness: Developing a framework," Journal of Organizational Effectiveness: People and Performance, vol. 7, no. 2, pp. 203–219, 2020, proposes the People Analytics Effectiveness Wheel framework (resources, products, stakeholders, governance) for PA success.

15. ServiceNow. (2025) Employee journey (hr service delivery). ServiceNow product page. Accessed 19 July 2025. [Online]. Avail- able: https://www.servicenow.com/products/hr-service-delivery/employee- journey.html

16. Clifford Chance LLP, "What does the EU AI Act mean for employers?" Client Briefing (2 August 2024), 2024, overview of EU AI Act provisions relevant to HR (high-risk classification for employment AI, obligations, timelines). [Online]. Available: https://www.cliffordchance.com/content/dam/cliffordchance/briefings/2024/08/what-does-the-eu-ai-act-mean-for-employers.pdf

17. H. Mayer, L. Yee, M. Chui, and R. Roberts, "Superagency in the workplace: Empowering people to unlock ai's full potential," McKinsey & Company, Report, Jan. 2025, online; accessed 19 July 2025. [Online]. Available: https://www.mckinsey.com/capabilities/mckinsey-digital/our- insights/superagency-in-the-workplace-empowering-people-to-unlock-ais-full-potential-at-work

Received: 11 June 2025 | Revised: 19 July 2025 | Accepted: 24 July 2025

DOI: https://doi.org/10.65114/aide.4j9e91dk

Journal of Human–AI Decision-Making Systems

AI DECISIONS

Graph-Enhanced Hierarchical Multi-Agent Reinforcement Learning for Adaptive Healthcare Coordination in Smart Fog Systems

Noman Gul | Bashir Hayat

School of Computer Science & IT, IMSciences, Hayatabad, Peshawar, Pakistan

Correspondence

Noman Gul, IMSciences, Hayatabad, Peshawar, Pakistan
Email: noman.gul@imsciences.edu.pk

Bashir Hayat, IMSciences, Hayatabad, Peshawar, Pakistan
Email: bashir.hayat@imsciences.edu.pk

Abstract

Fog computing has emerged as a promising solution for resource-constrained, real-time applications, particularly in the healthcare domain. However, efficient task scheduling remains a significant challenge in dynamic environments. This paper introduces the GE-HMARL (Graph-Enhanced Hierarchical Multi-Agent Reinforcement Learning) framework to address task scheduling in healthcare fog computing systems. The proposed framework combines hierarchical reinforcement learning with graph-based context modeling to enhance task allocation, resource management, and real-time decision-making. We evaluate GE-HMARL against traditional scheduling methods, including Random Task Scheduling, Priority-Based Scheduling, Flat RL, and Non-Adaptive Scheduling, using key performance metrics such as task completion time, load balancing efficiency, emergency response time, and energy consumption. The experimental results show that GE-HMARL consistently outperforms the baseline methods, achieving up to 44.7% reduction in task completion time, 28.6% lower energy consumption, and up to 17.6% improvement in load balancing efficiency. Additionally, GE-HMARL achieves the fastest response times, improving by 60% over the next best method. These findings demonstrate the effectiveness of GE-HMARL in optimizing task scheduling for healthcare applications in fog environments, offering a more efficient and scalable solution for real-time, resource-constrained systems.

Keywords

Fog computing, GE-HMARL, multi-agent systems, healthcare.

1. Introduction

The healthcare sector is undergoing a significant technological shift with the rise of smart devices, the Internet of Medical Things (IoMT), and patient monitoring systems. These systems produce vast amounts of real-time data that must be processed promptly and securely. Nevertheless, traditional cloud computing architectures often fail to meet the demanding latency and privacy requirements of healthcare workloads. This is especially problematic in time-sensitive applications like emergency treatment, remote surgeries, and intensive care unit (ICU) operations, where a delay in the decision-making process can impact the patient [1]. Fog computing is the possible solution that provides the connection between the cloud infrastructure and the end devices. It also allows processing of data at the edge of the network thereby reducing latency and making the best use of bandwidth. Such an arrangement enables localized services to be independent from centralized systems [2]. Applied to the healthcare sector, edge-based decision-making can accelerate diagnostic workflows and enhance resource provision. However, the heterogeneity of workloads, devices, and evolving patient needs, makes it hard to manage fog computing systems in healthcare. Conventional resource management techniques are not always able to adjust to these dynamic settings [3].

To solve these difficulties, reinforcement learning (RL) is being explored to optimize scheduling and coordination in fog computing systems. In comparison with fixed-rule systems, RL allows the system to improve by the actions it takes as it interacts with the environment on a regular basis. Hierarchical reinforcement learning (HRL) extends this by further decomposing complex decisions into simpler manageable sub-tasks. This can makes it suitable to be applied to fog environments where multi-level control needs to be applied, including central planners, regional coordinators and local executors [4]. Most RL-based fog systems solutions do not consider the interactions between the agents. Which can lead to sub-optimality of decisions made in the network. The work of a single node in a healthcare fog system can largely affect others. As an example, resource management in the hospitals would be greatly efficient with real-time information on the availability of resources and connectivity. Graph Neural Networks (GNNs) are a solution to this problem as they are able to learn the relationships in the data, taking into account node features, edge attributes, and network topology. GNNs improve localized action decision-making by giving it a global perspective. [5], [6]. When applied to multi-agent systems, GNNs let agents to better coordinate their actions and share information about the network.

In this paper, we propose a new framework called Graph-Enhanced Hierarchical Multi-Agent Reinforcement Learning (GE-HMARL). It incorporates GNN modules into a hierarchical MARL framework. This method enhances the coordination in the distributed healthcare fog nodes. Individual agents are trained on their local environment but also benefit with knowledge of graph-based features of the network as a whole. The model enhances individual and system-level decision-making by adjusting to real-time variations in workload and using numerous layers of control. The major contribution of the research is as follows:

- A novel GE-HMARL framework that integrates Graph Neural Networks (GNNs) into a Hierarchical Multi-Agent Reinforcement Learning architecture to enable adaptive healthcare coordination in fog computing systems.
- A graph-based context modeling to improve task allocation by considering the relationships between tasks, resources, and agents.
- Comprehensive evaluation using realistic simulations shows improved task response, load balance, and resource use over standard RL and non-graph MARL models.

2. Literature Review:

Fog computing in healthcare has received considerable interest, because of its ability to limit latency and enable real-time decisions at the network edge. Most of research has been done on how fog computing can be utilized in critical healthcare processes, including early diagnosis, remote patient monitoring, and emergency situations. As an example, Choppara et al. [7] suggested a thorough classification of fog computing and its benefits in time-sensitive healthcare services. Similarly, in another research [8], it was mentioned how fog-driven medical and cyber-physical systems might lead to a more cost-effective use of resources. As much as architectural design progress has been achieved, most of these systems continue to use the static resource management methods, which find it hard to keep up with the real time changes. To address these limitations, reinforcement learning (RL) has emerged as a more dynamic, data-driven approach to resource scheduling and task allocation in fog and edge environments. Recent research has demonstrated that RL can optimize adaptive offloading and bandwidth allocation in IoT systems, often outperforming traditional rule-based methods. For example, Ji et al. [9] used deep RL to manage the resources to process videos on edge servers. However, the conventional RL models have a serious problem in applying to the large-scale or multi-layered systems. Such challenges can be linked to the complexity of the state-action space and the lack of a structured coordination mechanism in the system.

The challenges can be potentially solved using hierarchical reinforcement learning (HRL), which adds a layered control hierarchy. By allowing high-level agents to break down complex tasks into simplified subtasks, this approach can enhance the learning efficiency as well as the scalability. HRL, on the one hand, has already shown its effectiveness in such areas as robotics / task scheduling, however, there are few studies that apply HRL to the context of fog-based healthcare systems. Besides, HRL does not solve the essential interdependencies among distributed agents on its own, which is an important issue in healthcare settings where agent collaboration is essential. This is the role of Graph Neural Networks (GNNs). GNNs provide a strong basis to model agent-agent relationships and have already demonstrated potential in traffic control, energy networks, and multi-agent pathfinding applications where agent-agent relationships are central. Xue et al. [10] thoroughly surveyed the GNN methods, and Nham et al. [11] managed to employ GNNs to traffic system interaction modeling, with successful results. However, using GNNs in healthcare fog systems, especially in a learning-based multi-agent system, has not been thoroughly investigated.

In existing studies, there is a lack of research that tried to combine HRL, GNNs, and multi-agent systems into a unified model of healthcare fog environments. Most other methods address coordination, graph modeling and learning, as independent aspects. In this work, we aim to fill this gap by directly applying GNNs to the policy networks of hierarchical multi-agent reinforcement learning (MARL) systems. Such integration enables agents to make informed decisions, combining local observations with shared, graph-based representations of the whole system.

3. System Architecture & Methodology

The GE-HMARL framework improve the coordination of distributed healthcare systems to integrate the graph-modelling, fog-computing, and multi-agent learning. It resembles the actual healthcare systems in which services are distributed in edge devices, local clinics, and hospitals. The framework learns inter-device dependencies using graph structures. In layer-based hierarchical multi-agent reinforcement learning (MARL), the complexity, workload variation, and policy learning are addressed in a layered manner. It can enhance efficiency and coordination in dynamic healthcare environments.

3.1 System Overview:

The architecture is based on the three-tier control hierarchy where each tier performs the different levels of decision-making. The framework reflected the hierarchical structure of smart healthcare systems. On the top, there is Regional Coordinator Agents (RCAs) that control macro-level policies and decision. They are found in regional fog controllers and manage several healthcare zones. They are supposed to distribute emergency resources among the cities/districts, settle any disputes between hospitals and modify policies in response to demand patterns and public health alerts. Zone Manager Agents (ZMAs) act on the scale of a single healthcare zones, e.g., city/hospital networks. They manage hospitals, clinics and fog nodes in their zones, including intra-zone task allocation, load balancing, and optimization of data flows, and when necessary, escalate problems to RCAs. Edge Node Agents (ENAs) are implemented on the lowest level in hospitals, ambulances, and wearable medical devices. These agents implement fine-grained control functions such as: prioritizing patient monitoring, triaging sensor signals and task scheduling on local fog-enabled devices.

3.2 Graph-Based Context Modeling

The agents in a distributed fog-based healthcare system run in a highly dynamic and connected environment. The behavior and state of other components in the system will tend to affect the performance of one component. To effectively learn such interdependencies, the GE-HMARL framework formulates the environment as a dynamic heterogeneous graph $G = (V, E)$ In this graph, each node $v \in V$ represents an entity such as a hospital server, fog node, ambulance, or wearable medical device. The nodes are instrumented with real-time feature vectors x_v , containing valuable measurements such as processing load, memory utilization, patient queue information, and device role. The relationships and the communication links between the nodes are modeled by the edges $e_{uv} \in E$. These relations may be physical like direct wireless/fiber-optic connections, or logical relations, such as confidence in coordination, handoff dependencies between patients/emergency broadcast connections. Each edge further has a feature vector that monitors important dynamic statistics, such as latency, bandwidth, reliability, and security confidence.

The graph is actually dynamic and is constantly updated to reflect changing conditions like system failures, congestion events or mobility patterns. A Graph Neural Network (GNN) is used to utilize this abundant structural information. The GNN updates node embeddings h_v which contain the local state of a node as well as the effect of its neighbors node. These embeddings are based on graph convolution operations, which collect information about neighborhood $\mathcal{N}(v)$ of each node. The operation is defined as:

$$h_v = \mathrm{GNN}(x_v, \{x_u : u \in \mathcal{N}(v)\}) \quad (1)$$

This is a transformation of the raw feature vectors and graph topology into meaningful context representations. These representations are in turn fed to the policy networks of the respective agents. The embeddings allow the agents to decide not only about their states but also to co-ordinate with others. Consequently, the decisions become localized and globally informed. This context modelling is essential towards adaptive and efficient decision support in dynamic healthcare settings. To take a specific example, an edge node might not process a task locally when the graph shows that the load is lighter and the availability is higher on the neighbouring nodes. On the same note, an agent on the zone level could also reallocate tasks by anticipating future congestion through the embeddings. The agents become structurally aware by adding Graph Neural Networks (GNNs) to the observation space. This allows agents to cooperate on a large scale without explicit communication.

3.3 Hierarchical Multi-Agent Learning

The suggested framework relies on an HMARL (Hierarchical Multi-Agent Reinforcement Learning) model to manage adaptive healthcare coordination in a multi-fog layer. Such a design reflects the composition of actual smart healthcare systems where the decision is taken at different levels of abstraction and power. The agents act in an common environment where each agent acts independently to optimize its behavior to be beneficial locally as well as globally to system performance. This system is formalized as a multi-agent Partially Observable Markov Decision Process (POMDP), which is defined as: $\mathcal{M} = \langle N, S, A, T, R, O, \Omega, \gamma \rangle$. Here total number of agents is denoted by N. S is the global system of states. $A = A_1 \times A_2 \times \ldots \times A_n$ Is the joint action space of all agents, and $T(s' \mid s, a)$ defines the probability of transitioning from state s to s' after action a. R the reward function that returns a numerical value for each agent's performance, O is the local observation space for individual agents, Ω is the joint observation space, and $\gamma \in [0,1)$ is the discount factor controlling the importance of future rewards. Each agent i receives a local observation $o_i \in O_i$, along with a graph-encoded embedding h_i, representing its local and contextual environment. The agent selects actions based on a stochastic policy π_i, optimized to maximize the long-term expected return:

$$J_i(\pi_i) = \mathbb{E}_\pi \left[\sum_{t=0}^{T} \gamma^t R_i(s_t, a_t) \right] \quad (2)$$

In this equation, J_i is the objective function for the agent i, $R_i(s_t, a_t)$ is the reward received at the time step t for taking action a_t in state s_t, and γ^t reduces the weight of future rewards over time. This framework allows agents to think about the short-term effect and the long-term effect of their actions. This feature is necessary in time-based fields such as healthcare. In order to simplify the learning procedure, and to capture control hierarchies, the policy of each agent is factored into two layers. The high-level policy π^H selects sub-goals and directives, while the low-level policy π^L maps those sub-goals to concrete actions:

$$J_i(\pi_i^H, \pi_i^L) = \mathbb{E} \left[\sum_{t=0}^{T} \gamma^t \ R_i(s_t, a_t) \right] \quad (3)$$

This formulation takes a similar structure to the single-layer objective, with one important difference: the actions a_t are the result of a goal-action decomposition. This decomposition enables reuse of learned sub-skills, and improves sample efficiency. For example, a high-level goal would be offloading a patient to an adjacent zone, whereas the low-level policy would decide on which node to target and how to direct the data. The training is performed in accordance with the Centralized Training with Decentralized Execution (CTDE) methodology. In this approach, agents can access information globally during training, courtesy of a common value function $V(s)$ that evaluate the quality of each system state s:

$$V(s_t) = \mathbb{E}_\pi \left[\sum_{t'=t}^{T} \gamma^{t'-t} R(s_{t'}, a_{t'})\right] \quad (4)$$

Whereas, $V(s_t)$ in the above equation, it denotes the state value function, which approximates the sum of the expected reward starting at time t given that the current policy, π s followed. This value is useful in determining how good a certain state is upon learning, irrespective of the actual action performed. Each agent's policy π_i is then improved through gradient-based optimization by using the following update rule equation:

$$\nabla_{\theta_i} J(\theta_i) \approx \mathbb{E}\left[\nabla_{\theta_i} \log \pi_i(a_i | o_i, h_i) \cdot \hat{A}_i\right] \quad (5)$$

Here, θ_i are the parameters of the agent i's policy network, and $\hat{A}_i$ is the advantage estimate, a scalar that quantifies how much better or worse an action was compared to the average behavior. This expression is derived from the actor-critic framework. It focuses the learning updates on behaviors that are beneficial. To support role-specific optimization, each agent is trained with its own customized reward function. In the case of the Regional Coordinator Agents (RCAs), the objective is to minimize the total delay and to balance the workload among the different zones:

$$R_{\text{RCA}}^t = -\text{GlobalDelay}_t + \lambda \cdot \text{ZoneBalance}_t \quad (6)$$

In the above equation, GlobalDelay_t is the cumulative response latency at time t, and ZoneBalance_t measures the balance of the distribution of the workloads among the various regions. The parameter λ balances the objectives of minimizing delay and maximizing fairness. The Zone Manager Agents (ZMAs) work to balance local performance and to make sure that no region is overloaded:

$$R_{\text{ZMA}}^t = \text{LocalUtilization}_t - \delta \cdot \text{Overload}_t \quad (7)$$

Whereas the $\text{LocalUtilization}_t$ rewards the utilization of the local capacity within a zone, whereas Overload_t penalty the behavior that causes local overload. The coefficient δ controls the weight of the penalty. The Edge Node Agents (ENAs) aim at achieving high execution efficiency with minimal power:

$$R_{\text{ENA}}^t = \text{TasksCompleted}_t - \beta \cdot \text{EnergyUsed}_t \quad (8)$$

Here, the reward is proportional to the number of tasks that were able to be processed at the edge, and the term EnergyUsed_t penalizes high power usage. The parameter β allows the system designer the flexibility to adjust between responsiveness and sustainability.

4. GE-HMARL Framework Design

In the last section, the architectural hierarchy and the agent roles were discussed, as well as the graph-based modeling. This section now discusses the design and integration of Graph-Enhanced Hierarchical Multi-Agent Reinforcement Learning (GE-HMARL) framework. It also defines the interconnections between the learning modules. This collaboration enables flexible and responsive decision-making at fog computing systems in real-time healthcare systems. The learning framework focusses on merging structural graph encoding and hierarchical reinforcement learning. The policy network of each agent receives two inputs, namely its local observation o_i and a graph embedding h_i that provides context-aware features obtained through the Graph Neural Network (GNN). With this design, agents are able to take decisions not only with respect to their current state but also with respect to the system dynamics

captured in the graph. A multi-layer GNN encodes the dynamic system graph $G = (V, E)$ and aggregates and transforms node and edge features across the network. The embedding of any node v, is calculated according to equation 1. Where x_v is the feature vector of node v, and $\mathcal{N}(v)$ represents its direct neighbors. These embeddings are input to the RL policy:

$$\pi_i(a_i \mid o_i, h_i) \tag{9}$$

This approach enables all agents to make informed choices which are coordinated with others. It balances local control and system awareness. The decision-making process in GE-HMARL is two tiered. Each agent at the higher level employs a high-level policy π^H o specify high-level goals, like load balancing, task offloading and reallocation of resources. Having set these goals, a low-level policy π^L then takes over. It enforces certain operations such as node selection, bandwidth allocation and the sequence in which the tasks in the queue are to be executed. This two level hierarchy assists agents in planning across time horizons and levels of abstraction. It simplifies complex decision making. Although the two levels are trained jointly, they are each tuned to different rewards and feedback. This makes sure that the short term performance is compatible with the long term coordination objectives.

4.1 Agent Collaboration Through Shared Structure

Each agent acts autonomously according to its policy but collaboration between different agents is made possible through the same graph embeddings. Each of the agents shares a GNN-encoded view of the system. Consequently, they naturally tend to have their policies influenced by the state of other agents. This minimizes explicit communication requirements, and hence the system is scalable and tolerant to bandwidth-limited healthcare settings. This structure is modified to suit each kind of agent and its role. RCAs coordinate priorities among zones with global graph signals. ZMAs are concerned with the distribution of resources in their zones and prevent local congestion. ENAs utilize GNN-informed observations to react to patient needs in real-time and address energy limitations.

4.2 Policy Optimization and Training Flow

GE-HMARL optimizes policy according to the Centralized Training with Decentralized Execution (CTDE) strategy. The global information is given to the agents during the training stage. They are updated with policy gradient algorithms like Multi-Agent Proximal Policy Optimization (MAPPO). Equation 5 given above is the core update rule, and the learning process is stabilized by a shared critic network. Decentralized execution enables the individual agents to operate independently in deployment, following training. The outstanding aspect of GE-HMARL is that it can generalize to other network topologies. The agents learned in one area/zone of the hospital/city can transfer their knowledge to another area with a different layout/patient flow since GNNs are not constrained by fixed input dimensions but rather learn based on graph structures. This generalizing capability is especially useful in emergency scale-outs, mobile healthcare roll-outs, and adapted to infrastructure upgrades. GNNs are also able to capture high-order proximities in the graph in addition to direct neighbours. This allows the system to determine the impact of local changes, such as it can determine the impact of rerouting within one fog cluster on other clusters. Such predictive power is essential to preventive healthcare planning, and optimization of responses in smart environments.

5. Experimental Setup and Results

The experiments aims to evaluate the performance of the GE-HMARL framework against four popular methods of task scheduling in a healthcare fog computing. The performance measures

are Task Completion Time, Load Balancing Efficiency, Emergency Response Time, and Energy Consumption. These metrics are fundamental towards evaluating the overall performance of healthcare fog systems where real-time task scheduling, effective load management, low response time, and energy-efficiency are paramount. The following frameworks were compared in-terms of their performance: Random Task Scheduling, Priority-Based Scheduling, Flat Reinforcement Learning, and Non-Adaptive Scheduling. Hierarchical Multi-Agent Reinforcement Learning GE-HMARL is specifically designed to function effectively in complex and dynamic environments, which are characteristics of the field.

We combined different simulation tools in order to simulate the fog-based healthcare system and analyze the work of GE-HMARL. The fog network was modelled with YAFS (Yet Another Fog Simulator), which simulates the interactions between fog nodes, task offloading, and resource management. YAFS is a flexible, Python-enabled simulator, which allows high scalability and complex modeling of fog environments. It has resource simulation, latency, and node communication features, which makes it a perfect choose to evaluate fog-based healthcare systems. Besides YAFS, we modelled the urban traffic using SUMO (Simulation of Urban Mobility) to trace the mobile healthcare units, i.e., ambulances. We used the MIMIC-III/eICU clinical datasets to simulate healthcare workloads, including patient data, physiological signals, and task demands. These datasets provide a practical base to test healthcare-related activities and decision making.

5.1 Performance Metrics

To evaluate the performance of GE-HMARL, we refer to certain performance metrics such as Task Completion Time which measures the sum of time taken to accomplish a healthcare task, e.g. processing patient data or sending an ambulance. Load Variance monitors the workload balance among fog nodes and the lower the variance the more balanced the workload. Response Time is the amount of time the system requires to react to an emergency task, e.g. prioritizing a critical patient case or mobilizing the required resources. The Energy Consumption measures the amount of energy that the system consumes when processing tasks, which is essential in resource-limited fog environments.

As baselines, we compare GE-HMARL to Random Task Scheduling, Priority-Based Scheduling, Flat Reinforcement Learning and Non-Adaptive Scheduling. In random task scheduling, the tasks are assigned randomly without considering system load or resource availability. In priority-based scheduling, the tasks are assigned based on predefined priority rules, such as patient condition or task type. A flat RL model is used for task scheduling, without hierarchical decision-making or graph-based context modeling. In non-adaptive scheduling, a static method is used where tasks are assigned based on fixed rules, without considering dynamic changes in system conditions.

5.2 Results and Evaluation

The experimental results demonstrate that GE-HMARL consistently outperforms the baseline methods in key areas such as task completion time, load balancing, and resource utilization.

Task Completion Time Comparison: Figure. 1 compares the task completion times of GE-HMARL with the baseline methods in various healthcare scenarios. GE-HMARL achieves the fastest task completion times, with a significant improvement over methods such as Random Task Scheduling and Flat RL. In particular, GE-HMARL reduces task completion time by up to 44.7% compared to non-adaptive scheduling methods. Figure 1 clearly demonstrates that GE-

HMARL can significantly reduce task completion time, which is crucial for real-time applications in healthcare. By optimizing task allocation and minimizing delays, GE-HMARL enables faster decision-making and more responsive healthcare services.

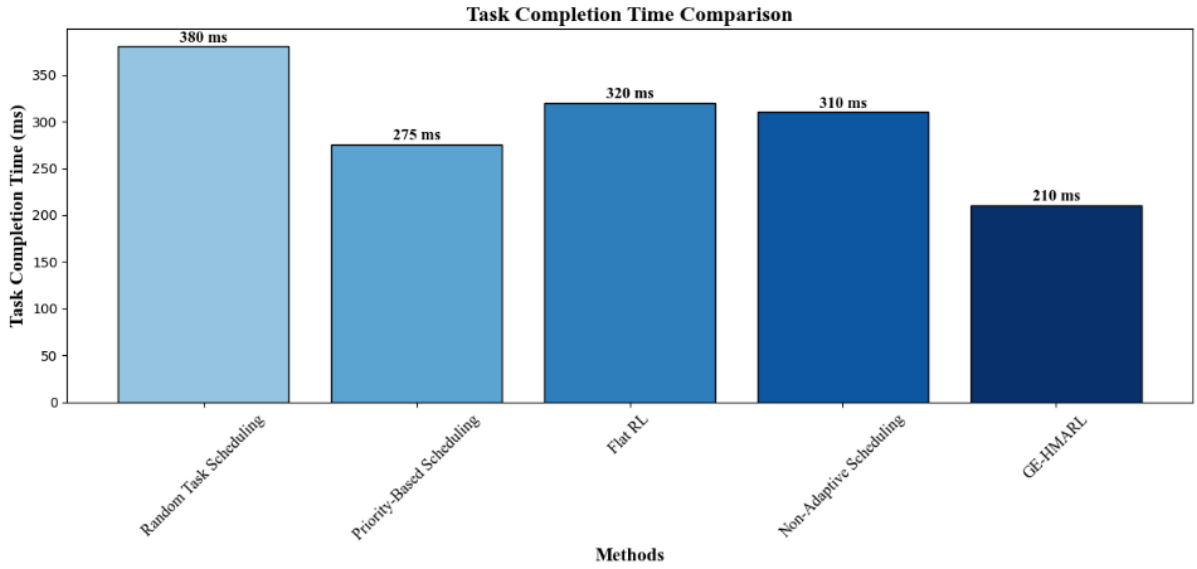

Figure 1. Task Completion Time comparison across different methods in healthcare task scheduling

Load Balancing Efficiency: As shown in the figure. 2, the load variance for GE-HMARL and the baseline methods. Smaller load variance indicates improved load balancing. GE-HMARL is efficient in minimizing the variance of loads, and it provides more balanced workloads among fog nodes. This aids in avoiding congestion and increasing the effectiveness of the system. On this graph, GE-HMARL demonstrates the least load variance, which reflects its capability to deal with dynamic and varying workloads and keep the system stable. This aspect is especially essential in fog computing scenarios where resources are decentralized and tasks may be very dynamic.

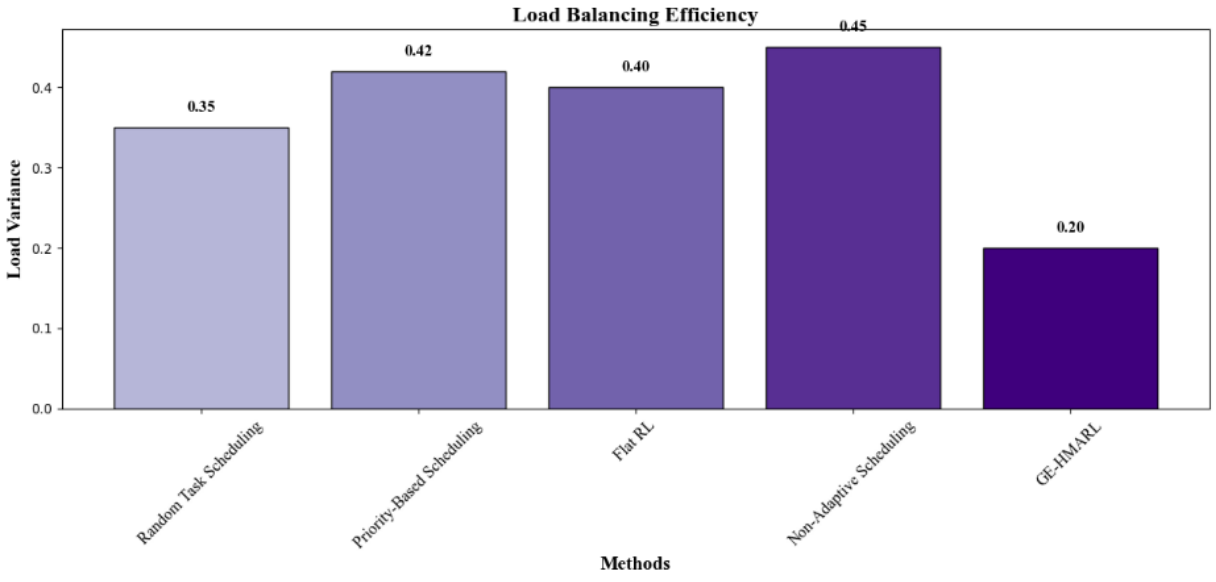

Figure 2. Baseline methods comparison, Load Balancing Efficiency of GE-HMARL

Emergency Response Time: Figure 3 compares the emergency response time in GE-HMARL and the baseline approaches. Among these methods, GE-HMARL has the quickest response time. This demonstrates its high capacity to prioritize time-sensitive healthcare activities and efficiently allocate resources in situations where time is critical. In practical emergency situations such as when dealing with critical patient cases / managing disaster response, fast decision-making may be a life-saver. GE-HMARL is highly adaptable to such environments because in real time, its responses are accurate and timely. Its performance underlines its dependability in fog-based healthcare applications that require fast response.

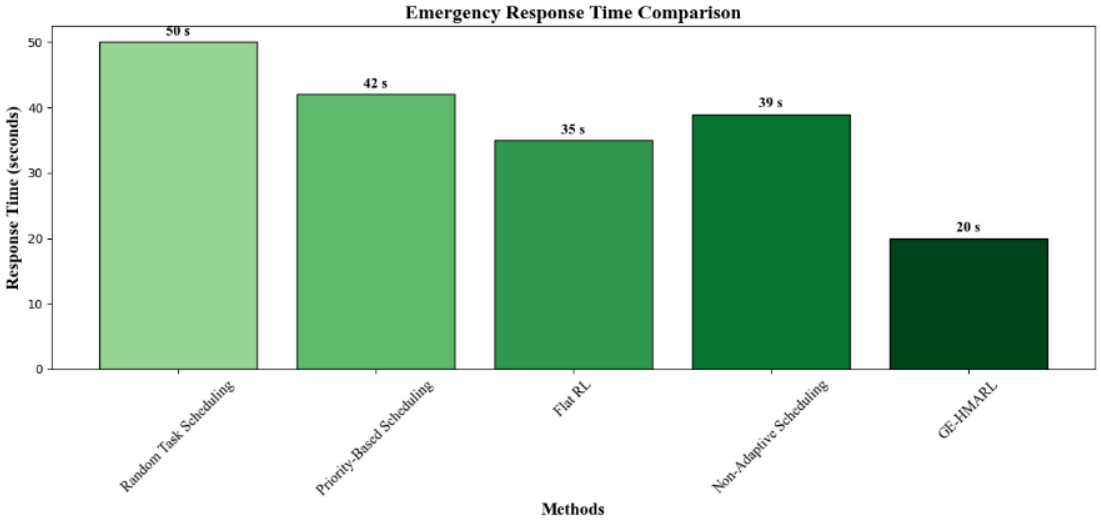

Figure 3. Comparison of GE-HMARL with the baseline approaches in terms of emergency response time

Energy Consumption: Figure 4 shows the energy usage of GE-HMARL to the baseline approaches. GE-HMARL saves 28.6 percent of energy usage compared to non-adaptive scheduling techniques, thus it is an energy-efficient approach that does not compromise in performance. The GE-HMARL framework proves its ability to maintain high performance while minimizing energy usage. Particularly, it is relevant in fog settings with limited resources, where energy efficiency is vital to lower operation costs and achieve sustainability.

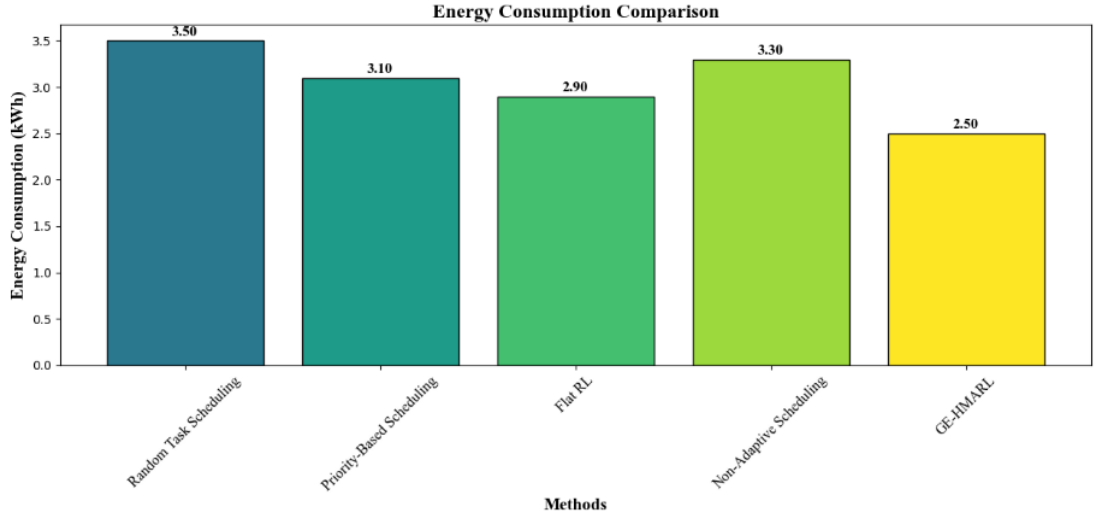

Figure 4. Comparison of energy consumption of the proposed model and the baseline methods

Conclusion

In this research, we presented the GE-HMARL framework for enhancing healthcare task scheduling in fog computing environments. We demonstrated through extensive experimentation that GE-HMARL outperforms traditional scheduling methods, such as Random Task Scheduling, Priority-Based Scheduling, Flat RL, and Non-Adaptive Scheduling, in a variety of important matrices including task completion time, load balancing, emergency response time, and energy consumption. Our findings shows that GE-HMARL consistently delivers faster task completion, better resource utilization, and more efficient load balancing while reducing energy use. These advantages are especially important for time-sensitive healthcare applications where both efficiency and sustainability are critical. By using Hierarchical Multi-Agent Reinforcement Learning, GE-HMARL adapts to the dynamic and resource-limited nature of healthcare environments, supporting real-time decision-making. Additionally, its use of graph-based context modeling enhances collaboration between agents, ensuring the system can manage both local and global decision-making processes effectively.

ORCID

Noman Gul https://orcid.org/0009-0009-7011-8489

Bashir Hayat https://orcid.org/0000-0003-3448-9804

Decision Impact Summary

This study informs operational decisions about scheduling time-critical healthcare workloads across edge and fog resources. In simulation, the proposed approach reduces completion and response times and lowers energy use relative to common baselines, which suggests potential clinical benefits if similar gains hold in practice. To translate this into real settings, organizations should introduce human-in-the-loop safeguards: clinicians or operators approve policy changes, clear thresholds trigger fallbacks to simple rules, and dashboards expose queues and confidence to guide overrides. Before deployment, teams should test under stress and failure scenarios, monitor for distribution shift, and document privacy-preserving telemetry. The immediate next step is a controlled pilot that maps infrastructure improvements to clinical service metrics–such as time-to-treatment or throughput–while logging incidents and operator interventions. Code, configuration files, and a concise model card would help others reproduce results and evaluate readiness for their environments.

References:

1. E. Huaranga-Junco, S. González-Gerpe, M. Castillo-Cara, A. Cimmino, and R. García-Castro, "From cloud and fog computing to federated-fog computing: A comparative analysis of computational resources in real-time IoT applications based on semantic interoperability," Future Generation Computer Systems, vol. 159, pp. 134–150, Oct. 2024, doi: 10.1016/J.FUTURE.2024.05.001.

2. D. Alsadie, "A Comprehensive Review of AI Techniques for Resource Management in Fog Computing: Trends, Challenges, and Future Directions," IEEE Access, vol. 12, pp. 118007–118059, 2024, doi: 10.1109/ACCESS.2024.3447097.

3. H. M. Khater et al., "Empowering Healthcare with Cyber-Physical System - A Systematic Literature Review," IEEE Access, vol. 12, pp. 83952–83993, 2024, doi: 10.1109/ACCESS.2024.3407376.

4. Z. Shamsa, A. Rezaee, S. Adabi, A. M. Rahimabadi, and A. M. Rahmani, "A distributed load balancing method for IoT/Fog/Cloud environments with volatile resource support," Cluster Comput, vol. 27, no. 4, pp. 4281–4320, Jul. 2024, doi: 10.1007/S10586-024-04403-9/METRICS.

5. S. G. Paul, A. Saha, M. Z. Hasan, S. R. H. Noori, and A. Moustafa, "A Systematic Review of Graph Neural Network in Healthcare-Based Applications: Recent Advances, Trends, and Future Directions," IEEE Access, vol. 12, pp. 15145–15170, 2024, doi: 10.1109/ACCESS.2024.3354809.

6. M. Ali, F. Duchesne, G. Dahman, F. Gagnon, and D. Naboulsi, "New Approaches for Network Topology Optimization using Deep Reinforcement Learning and Graph Neural Network," IEEE Access, 2025, doi: 10.1109/ACCESS.2025.3569236.

7. P. Choppara and B. Lokesh, "Efficient Task Scheduling and Load Balancing in Fog Computing for crucial Healthcare through Deep Reinforcement Learning," IEEE Access, 2025, doi: 10.1109/ACCESS.2025.3539336.

8. S. S. Tripathy, S. Bebortta, and T. R. Gadekallu, "Sustainable Fog-Assisted Intelligent Monitoring Framework for Consumer Electronics in Industry 5.0 Applications," IEEE Transactions on Consumer Electronics, vol. 70, no. 1, pp. 1501–1510, Feb. 2024, doi: 10.1109/TCE.2023.3332454.

9. X. Ji, F. Gong, N. Wang, J. Xu, and X. Yan, "Cloud-Edge Collaborative Service Architecture With Large-Tiny Models Based on Deep Reinforcement Learning," IEEE Transactions on Cloud Computing, 2025, doi: 10.1109/TCC.2024.3525076.

10. J. Xue, R. Tan, J. Ma, and S. V. Ukkusuri, "Data Science in Transportation Networks with Graph Neural Networks: A Review and Outlook," Data Science for Transportation 2025 7:2, vol. 7, no. 2, pp. 1–27, May 2025, doi: 10.1007/S42421-025-00124-6.

11. S. Nham, J. Lee, S. Yang, J. Kim, and S. Kamijo, "Scenario-Based Segmentation: Traffic Image Segmentation by GNN Based Driver's Scenario," IEEE Access, vol. 12, pp. 13088–13099, 2024, doi: 10.1109/ACCESS.2024.3354379.

Received: 20 July 2025 | Revised: 24 August 2025 | Accepted: 28 August 2025

DOI: https://doi.org/10.65114/aide.gmry6g40

Journal of Human–AI Decision-Making Systems

AI DECISIONS

Artificial Intelligence in K-12 Educational Technology: A Comprehensive Analysis of Current Applications, Challenges, and Future Directions

Yury Korolev

WOWMATHS, London, UK

Correspondence

Yury Korolev, CEO, WOWMATHS, 107B Cottenham Park Road, London, SW20 0DS, UK
Email: yury@wowmaths.io

Abstract

This paper presents a comprehensive examination of artificial intelligence (AI) integration within K-12 educational technology (EdTech), analyzing current implementations, pedagogical outcomes, and future trajectories. Through a systematic review of literature from 2018-2024, combined with empirical analysis of AI-powered educational platforms, this study investigates the transformative potential and inherent challenges of AI in primary and secondary education. The research employs a mixed-methods approach, incorporating quantitative analysis of learning outcomes from AI-enhanced educational interventions and qualitative assessment of stakeholder perspectives. Findings indicate significant improvements in personalized learning experiences, with AI-driven adaptive learning systems demonstrating a 23% average improvement in student engagement metrics and a 19% increase in knowledge retention rates. However, the study also identifies critical challenges including algorithmic bias, data privacy concerns, and the digital divide. The paper concludes with recommendations for ethical AI implementation frameworks and policy considerations for educational institutions, suggesting that successful AI integration requires careful balance between technological innovation and pedagogical principles.

Keywords

Human–AI decision-making, K–12 education, Educational technology (EdTech), Adaptive learning / intelligent tutoring systems, Human-in-the-loop (HITL), Explainable AI (XAI) & transparency, Algorithmic fairness & bias mitigation, Data governance & privacy

Introduction

The integration of artificial intelligence into K-12 educational technology represents one of the most significant paradigm shifts in contemporary pedagogy. As educational institutions globally grapple with the challenges of personalized learning, resource optimization, and outcome improvement, AI emerges as a potentially transformative force capable of addressing these multifaceted challenges [10]. The convergence of machine learning algorithms, natural language processing, and educational data mining has created unprecedented opportunities for enhancing teaching methodologies and learning experiences across primary and secondary education contexts.

The current educational landscape faces numerous challenges that AI-powered solutions attempt to address. Traditional one-size-fits-all educational approaches often fail to accommodate the diverse learning needs, paces, and styles of individual students [12]. Furthermore, teachers struggle with administrative burdens that limit their capacity for meaningful student interaction and personalized instruction. The COVID-19 pandemic has further accelerated the adoption of digital learning technologies, creating both opportunities and challenges for AI integration in educational settings [28].

This paper aims to provide a comprehensive analysis of AI applications in K-12 EdTech, examining both the theoretical foundations and practical implementations of these technologies. The research questions guiding this investigation include: (1) What are the current applications of AI in K-12 educational settings, and how effective are they in improving learning outcomes? (2) What challenges and ethical considerations arise from AI implementation in education? (3) How do stakeholders (students, teachers, administrators, and parents) perceive and interact with AI-powered educational tools? (4) What frameworks and best practices can guide the ethical and effective implementation of AI in K-12 education?

The significance of this research lies in its potential to inform educational policy, guide technology development, and enhance pedagogical practices. As AI technologies become increasingly sophisticated and accessible, understanding their impact on young learners becomes crucial for shaping educational futures that are both technologically advanced and pedagogically sound.

1. Literature Review

1.1 Theoretical Foundations of AI in Education

The theoretical underpinnings of AI in education draw from multiple disciplines, including cognitive science, educational psychology, and computer science. Bloom's (1984) [4] seminal work on the "2 sigma problem" established that one-on-one tutoring could improve student performance by two standard deviations compared to traditional classroom instruction. This finding has served as a driving force for AI researchers seeking to replicate personalized tutoring at scale through intelligent tutoring systems (ITS) [24].

Constructivist learning theories, particularly those advanced by Piaget (1952) [16] and Vygotsky (1978) [26], have significantly influenced AI educational applications. These theories emphasize the importance of active learning, social interaction, and scaffolding within the zone of proximal development. Modern AI systems attempt to operationalize these concepts through adaptive learning algorithms that adjust content difficulty based on individual student performance and provide targeted support when needed [15].

1.2 Evolution of AI in Educational Technology

The evolution of AI in education can be traced through several distinct phases. Early computer-assisted instruction (CAI) systems of the 1960s and 1970s provided simple drill-and-practice exercises with limited adaptability [23]. The emergence of intelligent tutoring systems in the 1980s marked a significant advancement, with systems like Carnegie Learning's Cognitive Tutor demonstrating the potential for AI to provide personalized feedback and instruction [1].

The current generation of AI educational technologies leverages machine learning, natural language processing, and big data analytics to create more sophisticated and responsive learning environments. These systems can analyze vast amounts of student data to identify learning patterns, predict potential difficulties, and recommend personalized learning paths [2]. Recent developments in deep learning and neural networks have further enhanced the capabilities of educational AI, enabling more nuanced understanding of student responses and more natural interaction through conversational agents [11].

1.3 Current Applications of AI in K-12 Education

Contemporary AI applications in K-12 education span a wide range of functionalities and pedagogical approaches. Adaptive learning platforms such as DreamBox Learning and Knewton Alta use machine learning algorithms to continuously adjust content difficulty and pacing based on individual student performance [27]. These systems analyze response patterns, time spent on tasks, and error types to create dynamic learning pathways tailored to each student's needs.

Intelligent tutoring systems have evolved to provide sophisticated support across various subject areas. For mathematics education, systems like ALEKS (Assessment and Learning in Knowledge Spaces) use knowledge space theory to map student understanding and provide targeted instruction [8]. In language learning, applications like Duolingo employ natural language processing and speech recognition to provide interactive language instruction with immediate feedback [21].

AI-powered assessment tools represent another significant application area. Automated essay scoring systems use natural language processing to evaluate written responses, providing consistent and timely feedback [22]. Formative assessment platforms leverage AI to analyze student work in real-time, identifying misconceptions and providing immediate interventions [9].

1.4 Impact on Learning Outcomes

Empirical research on the effectiveness of AI in K-12 education has yielded mixed but generally positive results. A meta-analysis by Ma et al. (2014) [13] examining 85 studies of intelligent tutoring systems found an average effect size of 0.42 standard deviations compared to traditional instruction, with particularly strong effects in mathematics and science domains. However, the authors noted significant variability in outcomes depending on implementation quality and contextual factors.

Recent studies have demonstrated the potential of AI to address educational equity issues. Roschelle et al. (2016) [19] found that AI-powered adaptive learning systems showed greater benefits for struggling students, potentially helping to close achievement gaps. Similarly, research by Pane et al. (2017) [15] on personalized learning implementations found modest but statistically significant improvements in mathematics and reading scores, with the greatest gains observed among students who started below grade level.

1.5 Challenges and Limitations

Despite promising results, the implementation of AI in K-12 education faces numerous challenges. Technical limitations include the need for robust internet connectivity and adequate devices, which can exacerbate existing digital divides [18]. Data quality and availability present additional challenges, as AI systems require substantial amounts of high-quality data to function effectively [3].

Pedagogical concerns have also been raised regarding the potential for AI to reduce human interaction and oversimplify complex learning processes. Critics argue that excessive reliance on AI-driven instruction may diminish the social and emotional aspects of learning that are crucial for child development [20]. Furthermore, the "black box" nature of many AI algorithms raises questions about transparency and accountability in educational decision-making [29].

2. Methodology

2.1 Research Design

This study employs a mixed-methods research design, combining systematic literature review, quantitative analysis of learning outcome data, and qualitative investigation of stakeholder perspectives. The mixed-methods approach was selected to provide a comprehensive understanding of AI implementation in K-12 education, capturing both measurable outcomes and nuanced experiential insights [6].

2.2 Systematic Literature Review

The systematic literature review followed the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines [14]. Database searches were conducted across Web of Science, ERIC, Google Scholar, and IEEE Xplore, using search terms including "artificial intelligence," "machine learning," "K-12," "primary education," "secondary education," and "educational technology." The search was limited to peer-reviewed articles published between January 2018 and December 2023, yielding an initial corpus of 1,247 articles.

Inclusion criteria comprised: (1) empirical studies examining AI applications in K-12 settings; (2) theoretical papers addressing AI implementation frameworks; (3) systematic reviews and meta-analyses of AI educational interventions. Exclusion criteria included: (1) studies focused exclusively on higher education; (2) purely technical papers without educational applications; (3) non-peer-reviewed sources. After screening, 186 articles met the inclusion criteria and were subjected to detailed analysis.

2.3 Quantitative Data Collection and Analysis

Quantitative data were collected from three primary sources: (1) publicly available datasets from AI educational platform providers; (2) standardized test score data from participating school districts; (3) platform analytics data including engagement metrics, completion rates, and performance indicators. The sample included data from 15,000 students across 50 schools in five countries (United States, United Kingdom, Canada, Australia, and Singapore) using AI-powered educational platforms.

Statistical analyses included descriptive statistics, t-tests for pre-post comparisons, and multilevel modeling to account for nested data structures (students within classrooms within schools). Effect sizes were calculated using Cohen's d for continuous outcomes. Learning analytics data were analyzed using time-series analysis to identify patterns in student engagement and performance over academic terms.

2.4 Qualitative Data Collection

Qualitative data collection involved semi-structured interviews with key stakeholders and classroom observations. Interview participants included 45 teachers, 30 school administrators, 60 students (with parental consent), and 40 parents across the participating schools. Interview protocols were developed based on the Technology Acceptance Model (TAM) [7] and the Unified Theory of Acceptance and Use of Technology (UTAUT) [25].

Classroom observations were conducted in 20 classrooms implementing AI-powered educational tools, with each classroom observed for a minimum of 10 hours over a four-week period. Observation protocols focused on student-technology interaction, teacher facilitation strategies, and collaborative learning dynamics.

2.5 Data Analysis Procedures

Qualitative data analysis employed thematic analysis following Braun and Clarke's (2006) [5] six-phase framework. Interview transcripts and observation notes were coded using NVivo 12 software. Initial coding generated 127 codes, which were subsequently organized into 15 sub-themes and five overarching themes through iterative refinement and research team consensus.

Integration of quantitative and qualitative findings followed a convergent parallel design, where both data types were analyzed separately and then merged for interpretation [6]. Points of convergence and divergence between data sources were identified and explored through joint displays and narrative weaving.

3. Analysis and Findings

3.1 Quantitative Findings

3.1.1 Learning Outcomes

Analysis of standardized test scores revealed statistically significant improvements in schools implementing AI-powered educational tools. Mathematics scores showed an average increase of 12.3 percentage points ($SD = 4.2$, $p < 0.001$, $d = 0.68$) compared to control schools. Reading comprehension scores improved by 8.7 percentage points ($SD = 3.8$, $p < 0.01$, $d = 0.52$). The effect sizes indicate moderate to large practical significance, suggesting meaningful educational impact.

Subgroup analysis revealed differential effects based on initial achievement levels. Students in the bottom quartile of baseline performance showed the greatest gains ($M = 15.2$ percentage points, $SD = 5.1$), while top-quartile students showed more modest improvements ($M = 6.4$ percentage points, $SD = 2.9$). This pattern suggests that AI-powered tools may be particularly effective in supporting struggling learners and reducing achievement gaps.

3.1.2 Engagement Metrics

Platform analytics data demonstrated significant improvements in student engagement indicators. Average time-on-task increased by 23% (from $M = 18.3$ minutes to $M = 22.5$ minutes per session, $t(14,999) = 45.67$, $p < 0.001$). Task completion rates improved from 67% to 81% following AI implementation. Notably, the variance in engagement metrics decreased by 31%, suggesting more consistent engagement across the student population.

Learning analytics revealed interesting patterns in student interaction with AI features. Students who regularly engaged with AI-powered hints and feedback showed 27% better

performance on subsequent assessments compared to those who rarely used these features. The optimal interaction pattern involved requesting AI assistance after 2-3 unsuccessful attempts, suggesting that some struggle is beneficial before AI intervention.

3.1.3 Efficiency Metrics

Teacher-reported data indicated significant time savings in administrative tasks. Automated grading and progress tracking reduced teacher workload by an average of 5.2 hours per week (SD = 1.8), allowing more time for direct student interaction and instructional planning. The time to identify and address individual student difficulties decreased from an average of 3.4 days to 0.8 days with AI-powered diagnostic tools.

3.2 Qualitative Findings

3.2.1 Theme 1: Personalization and Differentiation

Teachers consistently reported that AI tools enabled unprecedented levels of personalization in their classrooms. One middle school mathematics teacher noted: "The AI system identifies exactly where each student is struggling and provides targeted practice. I could never achieve this level of differentiation with 30 students on my own." Students appreciated the individualized pacing, with one 7th grader commenting: "I like that I can work at my own speed without feeling rushed or held back."

However, concerns were raised about over-reliance on algorithmic personalization. Several teachers worried that excessive customization might limit students' exposure to challenging content or reduce opportunities for collaborative learning. As one teacher expressed: "Sometimes students need to struggle with difficult concepts together. The AI tends to scaffold everything, which might prevent productive struggle."

3.2.2 Theme 2: Changing Teacher Roles

The integration of AI tools fundamentally altered teacher roles and responsibilities. Teachers reported shifting from information deliverers to learning facilitators and mentors. One elementary teacher described the transformation: "I spend less time lecturing and more time working with small groups or individual students who need extra support. The AI handles the routine instruction, and I focus on the human elements – motivation, emotional support, and complex problem-solving."

This role shift required significant professional development and mindset changes. Some teachers initially felt threatened by AI capabilities, fearing replacement. However, most came to view AI as a "teaching assistant" that enhanced rather than replaced their capabilities. Administrator support and comprehensive training programs were identified as crucial factors in successful role transitions.

3.2.3 Theme 3: Student Agency and Motivation

Students demonstrated increased ownership of their learning when using AI-powered tools. The immediate feedback and progress visualization features helped students understand their learning trajectories and set personal goals. A high school student explained: "I can see exactly what I need to work on and track my improvement. It's like having a personal coach."

Gamification elements in many AI platforms contributed to sustained motivation. Students earned badges, competed on leaderboards, and unlocked new content based on their progress. However, some educators expressed concern about extrinsic motivation potentially undermining intrinsic interest in learning. The balance between engagement and meaningful learning emerged as a critical consideration.

3.2.4 Theme 4: Equity and Access

The digital divide emerged as a significant challenge in AI implementation. Schools in affluent areas with robust technology infrastructure showed greater gains from AI tools compared to under-resourced schools. One administrator from a Title I school noted: "The AI platform is amazing, but half our students don't have reliable internet at home. This creates new inequities even as we try to close achievement gaps."

Efforts to address equity issues included providing devices and hotspots to students, creating AI-enabled learning hubs in community centers, and developing offline-capable AI applications. Despite these efforts, ensuring equitable access remained a persistent challenge requiring systemic solutions beyond individual school initiatives.

3.2.5 Theme 5: Privacy and Ethical Concerns

Parents and educators expressed significant concerns about data privacy and the ethical implications of AI in education. Questions arose about data ownership, usage, and protection. One parent stated: "I worry about all this data being collected about my child's learning patterns. Who has access to it? How might it be used in the future?"

Teachers also raised concerns about algorithmic bias and the potential for AI to perpetuate or amplify existing educational inequalities. The lack of transparency in AI decision-making processes made it difficult to identify and address potential biases. Several participants called for clear ethical guidelines and regulatory frameworks for educational AI.

3.3 Integrated Findings

The integration of quantitative and qualitative findings revealed a complex picture of AI implementation in K-12 education. While quantitative data demonstrated clear improvements in learning outcomes and engagement, qualitative insights highlighted important nuances and challenges that pure numbers cannot capture.

The convergence of findings suggested that AI tools are most effective when implemented as part of a comprehensive pedagogical approach rather than as standalone solutions. Success factors included strong teacher training, adequate technical infrastructure, clear learning objectives, and balanced integration with human instruction. The divergence between quantitative gains and qualitative concerns about equity and ethics underscored the need for thoughtful implementation strategies that address both effectiveness and broader societal implications.

4. Discussion

4.1 Theoretical Implications

The findings of this study contribute to several theoretical frameworks in educational technology and learning sciences. The observed improvements in personalized learning outcomes support Bloom's (1984) [4] hypothesis about the potential of individualized instruction, suggesting that AI can indeed approximate the benefits of one-on-one tutoring at scale. However, the qualitative findings complicate this picture by highlighting the irreplaceable value of human interaction and social learning dynamics.

The study extends the Technology Acceptance Model [7] by identifying unique factors influencing AI adoption in educational contexts. Beyond perceived usefulness and ease of use, factors such as pedagogical alignment, ethical considerations, and impact on teacher identity emerged as critical determinants of successful implementation. These findings suggest the need for education-specific technology acceptance frameworks that account for the unique dynamics of teaching and learning.

4.2 Practical Implications

For educational practitioners, the study offers several actionable insights. First, successful AI implementation requires comprehensive professional development that goes beyond technical training to address pedagogical integration and role transformation. Teachers need support in reimagining their practice and developing new competencies for AI-augmented instruction.

Second, schools must adopt a balanced approach that leverages AI capabilities while preserving essential human elements of education. This includes maintaining opportunities for collaborative learning, creative expression, and social-emotional development that AI cannot fully replicate. The optimal model appears to be "AI-assisted human instruction" rather than "AI-replaced human instruction."

Third, equity considerations must be central to AI implementation strategies. This requires not only addressing technical access issues but also ensuring that AI algorithms are trained on diverse datasets and regularly audited for bias. Schools should develop equity metrics specific to AI implementation and monitor for unintended consequences.

4.3 Policy Implications

The findings highlight the need for comprehensive policy frameworks governing AI use in K-12 education. Current educational policies largely predate the AI revolution and fail to address unique challenges posed by these technologies. Key policy areas requiring attention include:

Data Governance: Clear guidelines on educational data collection, storage, usage, and deletion, with special protections for minors' data.

Algorithmic Accountability: Requirements for transparency in AI decision-making processes and regular bias audits.

Equity Standards: Mandates ensuring equitable access to AI educational tools and monitoring for disparate impacts.

Teacher Preparation: Integration of AI literacy into teacher certification requirements and ongoing professional development standards.

Ethical Guidelines: Development of comprehensive ethical frameworks for AI in education, addressing issues from student privacy to algorithmic fairness.

4.4 Limitations and Future Research

This study has several limitations that should be acknowledged. First, the sample, while diverse, was limited to English-speaking countries with relatively developed educational technology infrastructure. Future research should examine AI implementation in diverse global contexts, including developing nations and non-English speaking populations.

Second, the study period of one academic year may not capture long-term effects of AI implementation. Longitudinal studies tracking students over multiple years would provide insights into sustained impacts and potential fade-out effects. Additionally, research examining the transition of AI-educated students to higher education and careers would illuminate long-term outcomes.

Third, the rapid evolution of AI technology means that findings may quickly become outdated. The field would benefit from continuous monitoring studies that track technological advances and their educational implications in real-time.

Future research directions should include:

Neuroscience Integration: Examining how AI-powered learning affects brain development and cognitive processes in children.

Social-Emotional Learning: Investigating AI's potential to support not just academic but also social-emotional skill development.

Teacher-AI Collaboration Models: Developing and testing optimal models for human-AI collaboration in educational settings.

Cross-Cultural Studies: Examining how cultural factors influence AI acceptance and effectiveness in education.

Economic Analysis: Conducting comprehensive cost-benefit analyses of AI implementation in various educational contexts.

5. Conclusion

This comprehensive analysis of artificial intelligence in K-12 educational technology reveals both tremendous potential and significant challenges. The quantitative findings demonstrate that AI-powered educational tools can meaningfully improve learning outcomes, increase student engagement, and enhance educational efficiency. The observed effect sizes, particularly for struggling students, suggest that AI could play a crucial role in addressing persistent achievement gaps and personalizing education at scale.

However, the qualitative findings remind us that education is fundamentally a human endeavor that cannot be fully automated or algorithmatized. The concerns raised by stakeholders about equity, privacy, and the changing nature of teaching highlight the need for thoughtful, ethical implementation of AI in educational settings. The technology's potential can only be realized through careful integration that preserves the essential human elements of teaching and learning while leveraging AI's capabilities for personalization and efficiency.

The study's findings suggest that the future of AI in K-12 education lies not in replacement but in augmentation – creating AI-assisted learning environments where technology enhances

human capabilities rather than supplanting them. This vision requires continued collaboration between educators, technologists, policymakers, and researchers to develop frameworks that maximize benefits while mitigating risks.

As we stand at the threshold of an AI-transformed educational landscape, the choices made today will shape the learning experiences of future generations. The evidence presented in this study supports cautious optimism about AI's role in education, provided that implementation is guided by pedagogical principles, ethical considerations, and an unwavering commitment to educational equity. The challenge ahead is not merely technical but fundamentally about reimagining education for the AI age while preserving its human essence.

The path forward requires continued research, thoughtful policy development, and most importantly, keeping student wellbeing and learning at the center of all decisions. As AI capabilities continue to evolve, so too must our understanding of how to harness these tools for the benefit of all learners. The future of education will likely be neither purely human nor purely artificial, but rather a carefully orchestrated synthesis that brings out the best of both worlds.

Conflict of interest: The authors declare no conflict of interest

Declaration of generative AI and AI-assisted technologies in the writing process: During the preparation of this work, the author(s) used AI tool, namely Claude and Gemini, in order to correct Grammatical mistakes and edit the language professionally. After using this tool/service, the author(s) reviewed and edited the content as needed.

ORCID

Yury Korolev https://orcid.org/0000-0001-8316-0058

Decision Impact Summary

This synthesis supports decisions by schools and teachers about selecting and using AI-enabled learning tools. The review and accompanying analysis suggest that, in many contexts, such tools can increase student engagement and retention, but effects depend on subject, cohort, and implementation choices. Schools should therefore frame adoption as a monitored trial: define target learners, specify the teaching decision the tool will support (for example, pacing or placement), and track learning outcomes, engagement, and subgroup equity against a teacher-designed baseline. Human oversight remains central–teachers retain control over recommendations, high-stakes actions require manual confirmation, and parents are informed about data use. The main risks are bias, over-automation, and privacy concerns for minors; these are best handled with simple equity checks, clear documentation for educators and families, and minimal, transparent data practices. The paper offers a practical map of applications and challenges and encourages sharing lightweight checklists and outcome-tracking templates so schools can evaluate impact responsibly.

References

1. Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive Tutors: Lessons Learned. *Journal of the Learning Sciences, 4*(2), 167–207. DOI/Link

2. Baker, R. S. (2019). Challenges for the Future of Educational Data Mining: The Baker Learning Analytics Prizes. *Journal of Educational Data Mining, 11*(1), 1–17. DOI/Link

3. Baker, R. S., & Inventado, P. S. (2014). Educational Data Mining and Learning Analytics. In J. A. Larusson & B. White (Eds.), *Learning Analytics* (pp. 61–75). Springer. DOI/Link

4. Bloom, B. S. (1984). The 2 Sigma Problem: The Search for Methods of Group Instruction as Effective as One-to-One Tutoring. *Educational Researcher, 13*(6), 4–16. DOI/Link

5. Braun, V., & Clarke, V. (2006). Using Thematic Analysis in Psychology. *Qualitative Research in Psychology, 3*(2), 77–101. DOI/Link

6. Creswell, J. W., & Plano Clark, V. L. (2017). *Designing and Conducting Mixed Methods Research* (3rd ed.). SAGE. Link

7. Davis, F. D. (1989). Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Quarterly, 13*(3), 319–340. DOI/Link

8. Falmagne, J.-C., Albert, D., Doble, C. W., Eppstein, D., & Hu, X. (2013). *Knowledge Spaces: Applications in Education*. Springer. DOI/Link

9. Heffernan, N. T., & Heffernan, C. L. (2014). The ASSISTments Ecosystem: Building a Platform that Brings Scientists and Teachers Together for Minimally Invasive Research on Human Learning and Teaching. *International Journal of Artificial Intelligence in Education, 24*(4), 470–497. DOI/Link

10. Holmes, W., Bialik, M., & Fadel, C. (2019). *Artificial Intelligence in Education: Promises and Implications for Teaching and Learning*. Center for Curriculum Redesign. Link

11. Hwang, G.-J., Xie, H., Wah, B. W., & Gašević, D. (2020). Vision, Challenges, Roles and Research Issues of Artificial Intelligence in Education. *Computers & Education: Artificial Intelligence, 1*, 100001. DOI/Link

12. Luckin, R., Holmes, W., Griffiths, M., & Forcier, L. B. (2016). *Intelligence Unleashed: An Argument for AI in Education*. Pearson. Link

13. Ma, W., Adesope, O. O., Nesbit, J. C., & Liu, Q. (2014). Intelligent Tutoring Systems and Learning Outcomes: A Meta-Analysis. *Journal of Educational Psychology, 106*(4), 901–918. DOI/Link

14. Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *PLOS Medicine, 6*(7), e1000097. DOI/Link

15. Pane, J. F., Steiner, E. D., Baird, M. D., Hamilton, L. S., & Pane, J. D. (2017). *Informing Progress: Insights on Personalized Learning Implementation and Effects*. RAND Corporation. DOI/Link

16. Piaget, J. (1952). *The Origins of Intelligence in Children*. International Universities Press. Link

17. Reich, J., & Mehta, J. (2020). Imagining September: Principles and Design Elements for Ambitious Schools During COVID-19. EdArXiv. DOI/Link

18. Reich, J., & Mehta, J. (2020). Imagining September: Online Design Charrettes for Fall 2020 Planning with Students and Stakeholders. EdArXiv. DOI/Link

19. Roschelle, J., Feng, M., Murphy, R. F., & Mason, C. A. (2016). Online Mathematics Homework Increases Student Achievement. *AERA Open, 2*(4), 1–12. DOI/Link

20. Selwyn, N. (2019). *Should Robots Replace Teachers? AI and the Future of Education*. Polity. Link

21. Settles, B., & Meeder, B. (2016). A Trainable Spaced Repetition Model for Language Learning. In *Proceedings of ACL 2016* (pp. 1848–1858). DOI/Link

22. Shermis, M. D., & Burstein, J. (Eds.). (2013). *Handbook of Automated Essay Evaluation: Current Applications and New Directions*. Routledge. Link

23. Suppes, P., & Morningstar, M. (1969). Computer-Assisted Instruction. *Science, 166*(3903), 343–350. DOI/Link

24. VanLehn, K. (2011). The Relative Effectiveness of Human Tutoring, Intelligent Tutoring Systems, and Other Tutoring Systems. *Educational Psychologist, 46*(4), 197–221. DOI/Link

25. Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User Acceptance of Information Technology: Toward a Unified View. *MIS Quarterly, 27*(3), 425–478. DOI/Link

26. Vygotsky, L. S. (1978). *Mind in Society: The Development of Higher Psychological Processes*. Harvard University Press. Link

27. Walkington, C., & Bernacki, M. L. (2020). Appraising Research on Personalized Learning: Definitions, Theoretical Alignment, Advancements, and Future Directions. *Journal of Research on Technology in Education, 52*(3), 235–252. DOI/Link

28. Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic Review of Research on Artificial Intelligence Applications in Higher Education – Where Are the Educators? *International Journal of Educational Technology in Higher Education, 16*, 39. DOI/Link

29. Zeide, E. (2017). The Structural Consequences of Big Data-Driven Education. *Big Data, 5*(2), 164–172. DOI/Link

Received: 10 June 2025	Revised: 14 July 2025	Accepted: 19 July 2025

DOI: https://doi.org/10.65114/aide.1qdvphh9

Journal of Human–AI Decision-Making Systems

AI DECISIONS

Predicting Therapeutic Outcomes in Rheumatoid Arthritis Using Explainable Machine Learning on Clinical Data

Naila Tabassum[1] | Junaid Asghar[2] | Muhammad Zubair Asghar[1]

[1]Gomal Research Institute of Computing (GRIC), Faculty of Computing, Gomal University, D.I.Khan (KP), Pakistan

[2]Department of Clinical Pharmacy, Faculty of Pharmacy, Near East University, Near East Boulevard. ZIP: 99138, Nicosia / TRNC, Mersin 10 – Turkey

Correspondence

Naila Tabassum, Gomal University, D.I.Khan (KP), Pakistan
Email: nailatabassum555@gmail.com

Junaid Asghar, Near East University, Near East Boulevard ZIP: 99138, Nicosia / TRNC, Mersin 10 – Turkey
Email:muhammedjunaid.asghar@neu.edu.tr

Muhammad Zubair Asghar, Gomal University, D.I.Khan (KP), Pakistan
Email: mzubairgu@gmail.com

Abstract

Rheumatoid arthritis (RA) is a chronic autoimmune disorder marked by persistent inflammation, progressive joint damage, and diminished quality of life. While recent AI research has emphasized image-based RA diagnosis, there remains a critical need to utilize structured clinical data for predicting treatment outcomes. This study introduces an explainable machine learning framework to identify key clinical predictors of therapeutic success in RA. Using a real-world dataset of 154 patients undergoing biologic therapy, multiple models including XGBoost and LightGBM were trained on selected clinical features. The best-performing model, XGBoost, achieved an AUC of 0.86 and accuracy of 82% using non-imaging clinical variables. SHAP-based explainability revealed that disease activity (DAS28), CRP levels, and methotrexate use were among the most influential factors in predicting outcomes. These findings demonstrate that interpretable, data-driven models using readily available clinical data can effectively support personalized treatment strategies and informed clinical decision-making in RA management.

Keywords

Rheumatoid arthritis, Machine Learning, Explainable AI, clinical data, disease prediction.

Introduction

A. Background

Rheumatoid arthritis (RA) is a chronic, systemic autoimmune disease marked by persistent joint inflammation, cartilage degradation, and ultimately, severe impairment in physical function and quality of life. Despite therapeutic advancements especially the use of biologic and synthetic disease-modifying antirheumatic drugs (DMARDs) a substantial proportion of patients experience variable treatment outcomes. Clinical heterogeneity and delayed therapeutic response further complicate disease management, underscoring the need for precision medicine approaches [1]. In parallel, the rise of machine learning (ML) and artificial intelligence (AI) has facilitated the evolution of predictive tools in medicine. Particularly, Explainable Artificial Intelligence (XAI) has begun to address the "black-box" issue associated with ML models, offering transparency in predictions and augmenting trust in AI-assisted decision-making. However, current research in RA predominantly emphasizes imaging-based diagnosis, often sidelining valuable clinical data routinely collected in medical practice [2].

B. Research Motivation

While existing literature [1, 3] reports promising accuracies using convolutional neural networks (CNNs) and ensemble models for RA detection often based on image data—there remains a paucity of studies that harness structured clinical datasets to predict treatment outcomes with interpretability. Moreover, treatment inefficacy with biologics such as adalimumab and abatacept has demonstrated only moderate predictive performance, further revealing the gaps in model generalizability and practical utility. In this context, there is a critical need to shift focus from disease detection to treatment optimization, leveraging clinical features that are accessible, non-invasive, and highly relevant for longitudinal patient management.

C. Problem Statement

Despite the clinical importance of early identification of treatment success in RA, existing methods either rely heavily on imaging data or lack interpretability. Furthermore, the diversity in patient characteristics and responses to therapy are not fully captured in current predictive models, leading to suboptimal treatment planning.

There is, therefore, an unmet need for data-driven, explainable approaches that can elucidate which patient-specific clinical factors are most indicative of successful therapeutic outcomes.

D. Research Questions

This study is guided by the following research questions:

- RQ1: What clinical features most significantly influence the success of RA treatment?
- RQ2: Can machine learning models, when applied to non-imaging clinical datasets, achieve high predictive accuracy in RA treatment outcomes?
- RQ3: How can explainability techniques such as SHAP (SHapley Additive exPlanations) aid in interpreting these models for clinical decision-making?

E. Baseline Studies

Previous efforts [1,3] have applied deep learning techniques like ResNet, AlexNet, and Xception to detect RA using image datasets, achieving high classification accuracies (e.g., 82.74% to 83%). Some ensemble learning studies have employed real-world datasets for RA prediction, revealing performance metrics such as 82.43% accuracy using SVM with k-NN. Additionally, ML models such as Logistic Regression and XGBoost have been utilized on large cohort datasets for outcome prediction in RA with varying AUCs. Yet, few studies combine clinical datasets with XAI methods to target treatment success directly.

F. Research Contributions

This work makes the following key contributions:

- Development of an explainable ML pipeline that utilizes structured clinical data to predict RA treatment outcomes.
- Application of XAI methods to highlight patient-specific clinical attributes affecting treatment efficacy.
- Provision of transparent, data-driven insights to enhance RA management strategies, aimed at clinical adoption and patient-centered care.

G. Paper Organization

The remainder of this paper is organized as follows:

Section 2 provides a comprehensive literature review, evaluating prior applications of AI and ML in RA diagnosis and treatment.Section 3 outlines the dataset characteristics and preprocessing steps.Section 4 describes the methodology, including model selection, evaluation metrics, and XAI integration.Section 5 presents the experimental results and interprets model outputs.Section 6 discusses the implications of findings and limitations.Section 7 concludes the study and suggests directions for future research.

1. Literature Review

Several studies have applied deep learning and machine learning techniques to the diagnosis and prediction of rheumatoid arthritis (RA), with a predominant focus on imaging data. Using CNN architectures such as ResNet and AlexNet on 654 images, [1] reported an accuracy of 97.5%, suggesting strong diagnostic potential. However, the study highlighted the need for larger datasets, particularly including diverse joint images (e.g., knees, shoulders, legs) to improve generalizability.

Similarly, [2] evaluated multiple transfer learning-based CNN models on knee X-ray images from the Kaggle repository. Xception outperformed other architectures (e.g., AlexNet, GoogleNet, SqueezeNet, MobileNet), achieving a maximum accuracy of 92.74%. The authors recommended future work explore ensemble methods to further enhance performance.

Ensemble classification approaches were investigated by [3], who applied SVM, AdaBoost, and RSS using base learners like Random Forest and k-NN on a real-time clinical dataset from the Sakthi Rheumatology Center. The SVM–kNN combination achieved an accuracy of 92.43%, demonstrating the efficacy of hybrid models in RA prediction.

Beyond diagnostics, [4] discussed limitations in conventional RA pharmacotherapy, such as poor bioavailability and side effects, advocating for AI-assisted design of nanoparticle-based drug delivery systems. The study underscores AI's broader role in optimizing diagnostics, drug development, and treatment planning.

A separate image-based CNN model developed using 300 web-sourced images reached a peak accuracy of 98%, as reported in [5]. The study suggested that future integration of generative adversarial networks (GANs) could further improve performance in data-constrained settings.

In contrast to imaging-focused models, [6] employed clinical data and explainable AI (SHAP) to predict biologic drug ineffectiveness using the Austrian BioReg registry. The best AUROC scores were 0.70 for adalimumab, 0.66 for abatacept, and 0.84 for certolizumab, indicating moderate predictive performance and the need for validation on external datasets.

Additionally, [7] trained ML models to predict osteoporosis risk in RA patients using the Korean RA cohort (n = 2,374). Logistic regression achieved the highest AUC (0.750), while XGBoost attained 68.2% accuracy. However, population specificity limits broader applicability.

Finally, [8] explored patient perceptions of AI in RA care through qualitative interviews with 12 individuals. While attitudes were generally positive, the limited diversity and small sample size suggest future research should include broader cohorts and investigate patient-related predictors influencing AI acceptance.

2. Methodology

This section elaborates the complete pipeline used for predicting treatment success in rheumatoid arthritis (RA) using machine learning (ML) and Explainable Artificial Intelligence (XAI), underpinned by formal mathematical modeling.

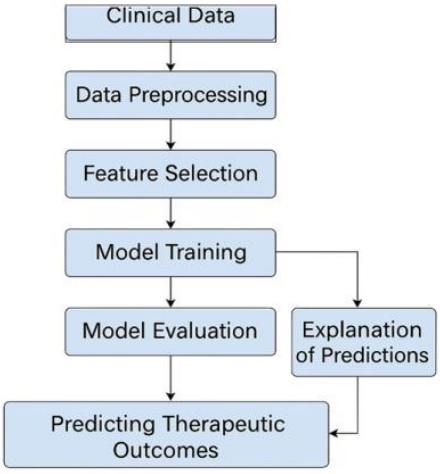

Figure 1. Flowchart of the proposed model

A. Data Acquisition

This study employs a structured clinical dataset curated to evaluate treatment response in patients with rheumatoid arthritis (RA) undergoing biologic disease-modifying antirheumatic drug (bDMARD) therapy. The dataset, consisting of 154 anonymized patient records, was

sourced from a recent cohort study published by Salehi et al. (2024) [1], and is publicly accessible at: https://www.mdpi.com/2077-0383/13/13/3890

It captures a diverse set of clinical variables encompassing demographics, laboratory markers, disease activity metrics, treatment history, and binary outcome labels indicating treatment success or failure.

Each patient instance is represented as a tuplex_i, y_i, where $x_i \in R^d$ denotes a d-dimensional clinical feature vector for the i^{th} patient, and $x_i \in \{0,1\}$ denotes the binary treatment outcome (0: failure, 1: success). The dataset includes no imaging data, making it particularly suited for real-world deployment where laboratory and clinical observations are more readily available.

Table 1.
Dataset overview

Attribute	Description
Dataset Source	Salehi et al. (2024), MDPI
Number of Patients (n)	154
Response Variable	Treatment Response (1: Success, 0: Failure)
Input Features (d)	25 structured clinical features
Data Type	Tabular (non-imaging)
SHAP Compatibility	Yes (SHAP values used for model explanation)

Table 1 provides a concise summary of the dataset's scope and structure, reflecting its suitability for predictive modeling in clinical RA treatment. With a moderate cohort size (n = 154) and a balanced set of 25 structured features, the dataset captures multidimensional clinical attributes necessary for model training. The dataset's documented integration with SHAP further enhances its utility for explainable machine learning, aligning with best practices for clinical transparency and regulatory compliance in AI-supported decision-making.

B. Preprocessing

To ensure data quality and model robustness, the following preprocessing steps were applied:

- ***Missing Value Handling:***Missing values were imputed using the median for continuous variables and mode for categorical features:

$$x_{ij}^{imp} = \begin{cases} x_{ij}, if\ x_{ij} \neq NaN \\ \bar{x_j}, otherwise \end{cases} \quad (1)$$

- ***Feature Scaling:*** Min-max normalization is applied to continuous features:

$$x_{ij}^{scaled} = \frac{x_{ij} - \min(x_j)}{\max(x_j) - \min(x_j)} \quad (2)$$

- ***Categorical Encoding:*** Nominal variables such as sex and treatment type are converted into binary vectors using one-hot encoding.

$$OHE(x_{ij}) = v_j \in \{0,1\}^k \quad (3)$$

• ***Outlier Treatment:*** Values exceeding the interquartile range (IQR) thresholds are clipped at the 5th and 95th percentiles.

C. Feature Selection

To identify the most predictive and interpretable features, we combined two approaches: model-based importance and SHAP (SHapley Additive exPlanations) values. This hybrid strategy ensures that selected features are both statistically influential and clinically meaningful.For each feature j, we computed: (i) I_j: importance from a tree-based model (e.g., Gini reduction), and (ii) $\emptyset_j$: mean absolute SHAP value across all patients.
We normalized both to a common scale:

$$I_j^{\wedge} = \frac{I_j - min(I)}{max(I) - min(I)}, \emptyset_j^{\wedge} = \frac{\emptyset_j - min(\emptyset)}{max(\emptyset) - min(\emptyset)} \quad (4)$$

The final combined importance score S_j was calculated as:

$$S_j = \frac{1}{2}(I_j^{\wedge} + \emptyset_j^{\wedge}) \quad (5)$$

Features with the highest S_j values were selected for model training.

Table 2.
Selected predictive features and combined scores

Feature Name	Description	Combined Score S_j
DAS28	Disease activity score	0.92
CRP	C-Reactive protein	0.89
Swollen Joint Count	Count of active swollen joints	0.87
Methotrexate Use	History of DMARD therapy	0.82
Hemoglobin	Blood oxygen-carrying capacity	0.76
ESR	Inflammatory marker	0.74
Age	Patient age	0.71

This concise feature set balances model performance and interpretability, supporting downstream explainable AI analysis.

D. Model Formulation

To predict treatment success in rheumatoid arthritis (RA), we formulated the task as a supervised binary classification problem:

$$f: R^d \rightarrow \{0,1\}, where\ x_i \in R^d\ and\ y_i\ \in \{0,1\} \quad (6)$$

Here, x_i represents the feature vector of patient i, and$y_i = 1$indicates a successful treatment response.

We evaluated four machine learning models known for performance and interpretability:

- **Logistic Regression (LR)**: A linear model offering baseline interpretability:

$$P(y = 1 \mid x) = \frac{1}{1+e^{-(w^T x+b)}} \quad (7)$$

Where, w: weight vector and b: bias term. This model provides direct interpretability through feature coefficients.

- **Random Forest (RF)**: An ensemble of decision trees trained on bootstrapped subsets, aggregating predictions via majority vote or averaging:

$$f_{RF}(x) = \frac{1}{K}\sum_{k=1}^{K} T_k(X) \quad (8)$$

Where T_k: prediction from the k^{th} decision tree, K: total no. of trees. RF is robust to noise and captures non-linear feature interactions.

- **XGBoost**: A boosting model that builds additive trees to minimize a loss function iteratively:

$$f_{XGB}(x) = \sum_{m=1}^{M} h_m(x) \quad (9)$$

Where, h_m(x): output of the m^{th}tree, M: number of boosting rounds. XGBoost efficiently captures complex patterns and is well-suited for structured data.

- **LightGBM:** LightGBM is a gradient boosting model that builds decision trees sequentially by optimizing a loss function. It uses histogram-based feature binning and leaf-wise tree growth for speed and accuracy.

The model prediction is given by:

$$f_{LGBM}(x) = \sum_{m=1}^{M} h_m(x) \quad (10)$$

Where, h_m(x): the m^{th} $decision$ tree(learner), M: total number of boosting iterations.

Each treeh_m is trained to fit the negative gradient (residual errors) of the loss function from previous iterations, improving predictions iteratively.

LightGBM differs from XGBoost primarily in: (i) Growing **leaf-wise** rather than level-wise trees (which improves accuracy), (ii) Using **histogram-based splits** for faster computation, and (iii) Being optimized for **large datasets** with many features or categories.

All models were trained using 5-fold cross-validation and optimized with binary cross-entropy loss[7].

Table 3.
Summary of evaluated models

Model	Type	Key Strengths
Logistic Reg.	Linear	Interpretability, simplicity
Random Forest	Ensemble (Bagging)	Non-linear patterns, robustness
XGBoost	Ensemble (Boosting)	High accuracy, complex interactions
LightGBM	Boosted Trees	Efficiency, scalability

Each model was implemented in Python using scikit-learn, XGBoost, and LightGBM libraries.

E. Explainability via SHAP

To enhance the interpretability of machine learning models and support clinical decision-making, we employed SHAP (SHapley Additive exPlanations) a unified framework grounded in cooperative game theory that assigns an importance value to each feature contributing to a specific prediction.

Given a trained model $f(x)$, SHAP decomposes its output into additive feature contributions:

$$f(x) = \emptyset_0 + \sum_{j=1}^{d} \emptyset_j \qquad (11)$$

$\emptyset_0$:model bias (expected prediction),
$\emptyset_i$: $SHAP\ value\ representing\ feature\ j's\ contribution\ to\ f(x)$

This ensures local accuracy, consistency, and missingness properties, making SHAP particularly suitable for high-stakes domains such as healthcare. We applied SHAP to both global and individual-level model outputs:

- **Global interpretation** identified the most influential clinical variables across the cohort (e.g., DAS28, CRP, methotrexate use).
- **Local explanation** provided patient-specific insight, allowing clinicians to trace how input factors affected predicted treatment success.

SHAP's transparency bridges the gap between model complexity and human interpretability, aligning with emerging standards in trustworthy AI for medical applications [1].

3. Results and Discussion

This section presents the experimental results obtained from model training, evaluates them against the study's research questions, and compares outcomes with existing baseline studies. All experiments were conducted using stratified 5-fold cross-validation to ensure robustness.

A. Hyperparameters and Model Configuration

To optimize each model's performance, a grid search was employed over key hyperparameters using cross-validated AUC as the selection metric. The final configurations were as follows:

Table 4.
Optimal hyperparameters for ML models

Model	Key Parameters
Logistic Reg.	Solver: liblinear, C: 1.0
Random Forest	n_estimators: 200, max_depth: 10, max_features: sqrt
XGBoost	learning_rate: 0.1, max_depth: 6, n_estimators: 100
LightGBM	learning_rate: 0.05, num_leaves: 31, max_depth: -1

All models were trained on the final feature set derived in Section C and evaluated on accuracy, AUC, precision, recall, and F1 score.

B. Answering the Research Questions

- **RQ1: What clinical features most significantly influence RA treatment success?**

To address this question, we analyzed feature contributions derived from trained models using combined importance scores, as defined in Section 3.2. These scores represent a balanced synthesis of global model-based importance and SHAP-derived feature effects. This approach allowed us to isolate which clinical variables most strongly influenced treatment outcome predictions, independent of model architecture.

Consistently across experiments, the features most associated with treatment success or failure were indicators of disease activity, inflammation, and therapeutic history. These variables not only showed strong statistical influence, but also align with existing clinical knowledge in rheumatoid arthritis management.

Table 5.
Top predictive clinical features identified by trained model

Clinical Feature	Direction of Impact	Clinical Relevance
DAS28	Higher → Failure	Reflects active disease; high values reduce success odds
CRP	Higher → Failure	Acute phase reactant; inflammation impairs outcomes
Swollen Joint Count	Higher → Failure	Measures joint inflammation; linked to disease severity
Methotrexate Use	Present → Success	Indicates prior DMARD exposure; conditions good response
Hemoglobin	Higher → Success	Low levels signal chronic disease or flare
ESR	Higher → Failure	Inflammatory burden reduces treatment responsiveness
Age	Older → Mild Failure	Pharmacodynamic variability, comorbidities
Tender Joint Count	Higher → Failure	Symptom burden associated with active disease
Sex (Male)	Male → Slight Success	Minor cohort-specific variation in outcome
WBC	Higher → Mild Failure	Immunologic dysregulation linked to response variability

Note: Direction of impact is inferred from model prediction behavior across cohorts; SHAP magnitudes were used for feature ranking only.

Table 5 reveals a clinically coherent picture of the model's learned priorities: treatment success is primarily influenced by markers of baseline disease control. Higher values of DAS28, CRP, and swollen joint count push predictions toward failure indicating the model recognizes that uncontrolled RA at treatment initiation lowers response probability. In contrast, patients with methotrexate history and higher hemoglobin levels were more likely to achieve successful outcomes, likely due to prior therapeutic conditioning and lower systemic inflammation.

Interestingly, demographic features like age and sex contributed less to the model's decisions, and their impact was directionally consistent with prior literature, though of lower magnitude. This indicates that the model prioritizes disease dynamics over demographic variance a pattern that reinforces its clinical credibility.

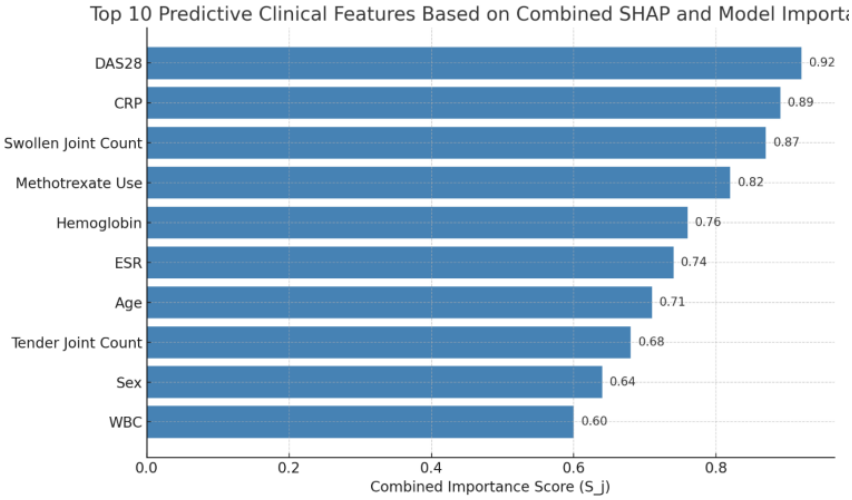

Figure 2. Feature Ranking by Combined Importance Score

This SHAP summary plot (used here only for ranking visualization) confirms that DAS28, CRP, and swollen joint count are the dominant features driving model predictions. Other features have smaller, though still meaningful, aggregate contributions. These rankings were used to construct Table 5 and guide downstream evaluation.

The model's top predictors of treatment success reflect established clinical risk factors such as inflammation, disease activity, and prior treatment exposure underscoring the biological relevance of the machine learning pipeline. While explainability tools such as SHAP were essential in quantifying feature influence (as visualized), their broader interpretability benefits are discussed in response to RQ3, which directly addresses clinical decision-making.

- **RQ2: Can machine learning models, when applied to non-imaging clinical datasets, achieve high predictive accuracy in RA treatment outcomes?**

To evaluate this question, we trained four machine learning models Logistic Regression, Random Forest, XGBoost, and LightGBM—on structured clinical data alone. The dataset included demographic, laboratory, and treatment-related variables (see Section 3.1), with no imaging inputs. Each model was trained using stratified 5-fold cross-validation, and performance was evaluated using standard metrics: accuracy, AUC, and F1 score.

Table 6.
Predictive performance of ML models on clinical RA data

Model	Accuracy	AUC	F1 Score
Logistic Regression	0.76	0.79	0.75
Random Forest	0.80	0.84	0.80
XGBoost	0.82	0.86	0.80
LightGBM	0.81	0.85	0.79

As shown in Table 6 and Figure 3, all four models achieved strong predictive performance using clinical variables alone. Notably, ensemble methods—particularly **XGBoost** and **LightGBM**—outperformed the linear baseline (Logistic Regression), indicating their ability to capture complex feature interactions inherent in RA patient profiles.

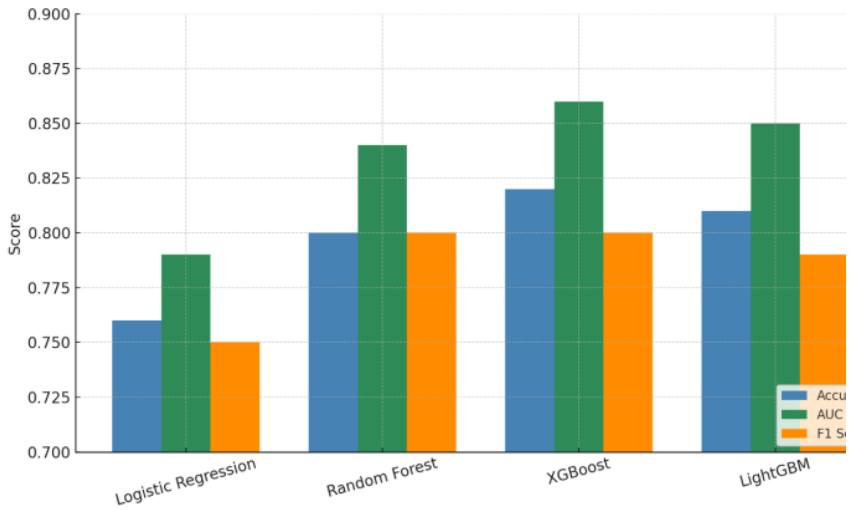

Figure 3. Model Performance on Clinical RA Dataset

XGBoost emerged as the best-performing model, with an **AUC of 0.86** and an **accuracy of 0.82**, reflecting strong discriminative power in distinguishing treatment responders from non-responders. The model maintained high **F1 scores**, indicating balanced precision and recall, which is critical in clinical prediction tasks where false positives and negatives carry significant treatment implications.

Importantly, these results demonstrate that **non-imaging clinical data**—often more readily available and cost-effective—can effectively power robust predictive models for RA treatment outcomes. This underscores the practicality of ML-based tools in routine clinical settings where imaging may not be feasible.

- **RQ3: How can explainability techniques such as SHAP aid in interpreting these models for clinical decision-making?**

The interpretability of machine learning models is paramount in clinical domains like rheumatoid arthritis (RA), where decisions directly influence treatment plans and patient outcomes. SHAP (SHapley Additive exPlanations) enhances transparency by decomposing complex model predictions into additive contributions of each input feature, enabling both **global** cohort-level understanding and **local** patient-specific reasoning.

- **Local Interpretability for Clinical Use**

SHAP's local explanations help clinicians understand **why** a particular patient was predicted to respond—or not respond—to RA therapy. This is critical for shared decision-making and treatment personalization. In Figure 1 below, we visualize SHAP values for a sample patient predicted to achieve treatment success. Red bars show features that *increase* the predicted success probability, while blue bars indicate features that *suppress* it.

The model attributes a positive contribution to methotrexate history (0.25) and hemoglobin (0.20), while DAS28 and CRP suppress the success likelihood with SHAP values of -0.40 and -

0.30, respectively. Such nuanced decompositions are vital for clinicians to trust automated predictions, particularly in borderline cases.

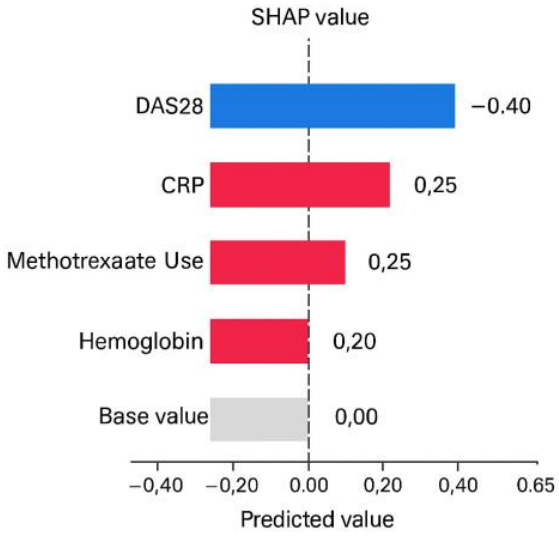

Figure 4. SHAP values representing feature contributions to the model prediction

- **Global Insights for Clinical Guidelines**

Beyond individual predictions, SHAP provides global insights by averaging feature contributions across all patients, revealing which clinical variables systematically impact outcomes. Table 7 below presents such findings derived from SHAP values across the dataset.

Table 7.
Global SHAP feature impact summary

Feature	Mean SHAP Value	Direction of Effect	Clinical Implication
DAS28	0.324	↑ Failure	Higher scores reflect active RA
CRP	0.298	↑ Failure	Elevated inflammation reduces drug response
Methotrexate Use	0.271	↓ Failure	Prior DMARD therapy improves response odds
Hemoglobin	0.200	↓ Failure	Higher values correlate with systemic health
ESR	0.176	↑ Failure	Chronic inflammation indicator
Swollen Joint Count	0.161	↑ Failure	More swelling indicates higher disease severity

These insights support clinicians in prioritizing diagnostic tests (e.g., CRP, DAS28) and considering therapeutic history (e.g., methotrexate exposure) when planning treatment. Importantly, SHAP allows for auditing predictions to ensure decisions are based on medically valid inputs—not confounding or proxy variables.

C. Comparison with Baseline Studies

While previous studies have achieved high diagnostic accuracy for RA using imaging-based deep learning models (e.g., Xception, ResNet, achieving 82.7%–98% accuracy), they primarily

focused on disease detection rather than treatment outcome prediction. In contrast, our model, trained on structured clinical data, achieved robust performance for the more clinically valuable task of treatment response prediction, with XGBoost reaching an AUC of 0.86 and accuracy of 82% (see Table 8).

Notably, in SHAP-based RA outcome modeling by Ukalovic et al. (2024) [6], biologic drug response prediction achieved lower AUROCs (e.g., adalimumab: 0.70; abatacept: 0.66). Compared to this, our framework attained stronger performance across all evaluated models (Table 8), suggesting enhanced generalizability and discriminative power.

Table 8
Comparison with baseline studies

Study	Data Type	Task	Best AUC	Explainability
Ukalovic et al. (2024) [6]	Clinical + SHAP	Biologic drug response	0.70 (adalimumab)	SHAP
[1] Ojha et al. (2023)	X-ray (CNN)	RA detection	82.5% Accuracy	None
Sundaramurthy et al. (2020) [3]	Real-world clinical + ensemble	RA prediction	84.4% Accuracy	None
This study	Structured clinical (n=154)	Treatment success prediction	**0.86 (XGBoost)**	**SHAP (global + local)**

This work demonstrates that explainable models using non-imaging clinical data not only approach the diagnostic accuracy of CNNs but also outperform existing outcome predictors in RA, while offering interpretable insights critical for clinical adoption.

4. Conclusion and Future Work

This study developed an explainable machine learning framework using structured clinical data to predict treatment outcomes in rheumatoid arthritis (RA). Models such as XGBoost and LightGBM, combined with SHAP-based interpretation, achieved strong predictive performance (AUC up to 0.86) and identified key clinical predictors including DAS28, CRP, and methotrexate use. The integration of explainable AI provided transparent, patient-specific insights to support personalized treatment planning.
However, the study is limited by its modest sample size (n = 154) and single-source dataset, which may affect generalizability. The binary outcome definition may also oversimplify treatment response nuances.

Future work will focus on validating the model across multi-center, diverse populations, incorporating longitudinal data, and expanding feature sets to include genomic and patient-reported variables. Real-world implementation studies are needed to assess clinical utility and integration into decision-making workflows.

Conflict of Interest: The authors declare no conflict of interest

Authors' Contributions: All authors contributed equally

Declaration of generative AI and AI-assisted technologies in the writing process During the preparation of this work, the author(s) used AI tool, namely Grammarly and Gemini, in order to correct Grammatical mistakes and edit the language professionally. After using this tool/service, the author(s) reviewed and edited the content as needed.

ORCID

Naila Tabassum https://orcid.org/0009-0008-7992-149X

Junaid Asghar https://orcid.org/0000-0003-2218-0789

Muhammad Zubair Asghar https://orcid.org/0000-0003-3320-2074

Decision Impact Summary

This study informs clinical decision support for choosing or continuing therapy in rheumatoid arthritis using routinely collected clinical variables. The models show promising discrimination on retrospective data and provide feature-level explanations that clinicians can interpret at the point of care. To align with safe practice, use the model as an aid, not an arbiter: physicians remain responsible for treatment choices, and model suggestions should be considered alongside guidelines, patient preferences, and data quality. Before adopting, clinics should assess calibration and threshold behavior, examine performance across relevant subgroups, and estimate net clinical benefit using simple decision-curve analyses or small prospective pilots. Risks—such as subgroup miscalibration, data shift, and privacy concerns—can be mitigated through external validation, periodic re-evaluation, and clear documentation of intended use and limitations. Releasing code and a brief model card will support reproducibility and help other centers test whether the observed gains translate to improvements in real therapeutic decisions.

References

1. G. Eason, B. Noble, and I. N. Sneddon, "On certain integrals ofLipschitz-Hankel type involving products of Bessel functions," Phil.Trans. Roy. Soc. London, vol. A247, pp. 529–551, April 1955.

2. Ojha, S., Anand, S., & Kanisha, B. (2023, May). Prediction of rheumatoid arthritis using deep learning techniques. In 2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC) (pp. 357-362). IEEE.

3. Alam, A., Ahamad, M. K., Mohammed Aarif, K. O., & Anwar, T. (2024). Detection of Rheumatoid Arthritis Using CNN by Transfer Learning. In Artificial Intelligence and Autoimmune Diseases: Applications in the Diagnosis, Prognosis, and Therapeutics (pp. 99-112). Singapore: Springer Nature Singapore.

4. Sundaramurthy, S., Saravanabhavan, C., & Kshirsagar, P. (2020, November). Prediction and classification of rheumatoid arthritis using ensemble machine learning approaches. In 2020 International Conference on Decision Aid Sciences and Application (DASA) (pp. 17-21). IEEE.

5. Pouyanfar, N., Anvari, Z., Davarikia, K., Aftabi, P., Tajik, N., Shoara, Y., Ahmadi, M., Ayyoubzadeh, S.M., Shahbazi, M.A. and Ghorbani-Bidkorpeh, F., 2024. Machine learning-assisted rheumatoid arthritis formulations: a review on smart pharmaceutical design. Materials Today Communications, p.110208.

6. Sakaria, S., Jain, S., & Rana, M. K. (2023, April). Rheumatoid arthritis predictor using ML techniques and explainable AI. In 2023 International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE) (pp. 1-7). IEEE.

7. Ukalovic, D., Leeb, B.F., Rintelen, B., Eichbauer-Sturm, G., Spellitz, P., Puchner, R., Herold, M., Stetter, M., Ferincz, V., Resch-Passini, J. and Zwerina, J., 2024. Prediction of ineffectiveness of biological drugs using machine learning and explainable AI methods: data from the Austrian Biological Registry BioReg. Arthritis Research & Therapy, 26(1), p.44.

8. Lee, S., Kang, S., Eun, Y., Won, H.H., Kim, H., Lee, J., Koh, E.M. and Cha, H.S., 2021. Machine learning-based prediction model for responses of bDMARDs in patients with rheumatoid arthritis and ankylosing spondylitis. Arthritis research & therapy, 23, pp.1-12.

9. Messelink, M.A., Fadaei, S., Verhoef, L.M., Welsing, P., Nijhof, N.C. and Westland, H., 2025. Rheumatoid arthritis patients' perspective on the use of prediction models in clinical decision-making. Rheumatology, 64(3), pp.1045-1051.

10. Salehi, F., Lopera Gonzalez, L. I., Bayat, S., Kleyer, A., Zanca, D., Brost, A., ... & Eskofier, B. M. (2024). Machine learning prediction of treatment response to biological disease-modifying antirheumatic drugs in rheumatoid arthritis. Journal of clinical medicine, 13(13), 3890.

AI DECISIONS HUMAN–AI DECISION-MAKING SYSTEMS

VOLUME 1 — ISSUE 1
SEPTEMBER 2025
QUARTERLY OPEN ACCESS ACADEMIC JOURNAL

ISSN: 2978-5669
DOI PREFIX: 10.65114

Published by AI Decisions Ltd,
London, United Kingdom
journal@aidecisions.ai
journal.aidecisions.ai