IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, October 2026
Impact Factor: 8.2

Explainable Artificial Intelligence in Medical Robotics: A Review of Surgical Decision Support, Robotic Precision, and Clinical Trustworthiness

Dr. Bulbul Sharma
Assistant Professor, Gautam Buddha University
School of Engineering, Gautam Buddha University, Greater Noida, Uttar Pradesh
bulbulsharma369@gmail.com

Abstract: *There is potential to enhance surgical decision support, robotic accuracy, and clinical dependability through the use of XAI into medical robots. It is difficult to comprehend suggestions, oversee robotic activities, and guarantee patient safety using traditional AI systems due to their lack of transparency. Trustworthy medical robotic architectures, precision-oriented robotic control, AI-enabled surgical intelligence, and XAI approaches are all included in this study. Surgical decision-making and prediction interpretation using methods like Gradient-weighted Class Activation Mapping (Grad-CAM), attention processes, Shapley Additive explanations (SHAP), and Local Interpretable Model-Agnostic Explanations (LIME) are covered. In addition, the paper delves into topics such as AI-based safety monitoring, microsurgical aid, adaptive trajectory planning, real-time imaging, deep learning, and reinforcement learning. The areas of risk prediction, human-robot interaction, automated mistake detection, and training based on simulations receive special emphasis. Lifecycle governance, interpretability, fairness, cybersecurity, privacy protection, layered architectures, and runtime safety assurance are all ways to check if something is trustworthy. The review concludes by stressing the significance of human oversight, open communication, and clinical validation for responsible implementation. This review organises the field of explainable medical robotics and highlights important factors to keep in mind while designing trustworthy, understandable, and clinically acceptable robots.*

Keywords: Explainable Artificial Intelligence, Medical Robotics, Surgical Decision Support, Robotic Precision, Clinical Trustworthiness, Human–Robot Interaction

I. INTRODUCTION

Health is an essential component of human well-being, and advances in healthcare technology have dramatically enhanced the quality, precision, and efficiency of medical treatments. Improvements in patient care, reductions in surgical risks, and funding for cutting-edge medical treatments have all resulted from healthcare's embrace of new technology [1]. The use of surgical robots and other medical procedures has been transformed by recent technical developments and advances in the field of medical robotics. Improvements in medical imaging systems, such as those with greater resolutions and magnetic imaging, as well as advances in actuators, sensors, control systems, and materials, have all contributed to tremendous success [2]. Medical treatments that have been effectively executed with the help of robotic systems have gained widespread recognition, and as a result, more and more people are prepared to endure procedures that are either partially or totally conducted by robots. as a result, medical robots is seeing tremendous growth [3].

The growing need for intelligent decision-making systems to be transparent has led to the rise of Explainable Artificial Intelligence (XAI) as a core area of AI research. Systems that are easy to understand and rely on are becoming more

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 7, October 2026 Impact Factor: 8.2

important as artificial intelligence continues to permeate robotics [4]. Engineers bygone days programd every possible reaction into traditional robots, which relied heavily on rules. Deep learning (DL), machine learning (ML), and artificial intelligence (AI) have transformed robots into systems that can learn, adapt, and optimise themselves [5]. The so-called black-box dilemma arises when the internal workings of neural networks and probabilistic reasoning models are opaque due to their complexity, even if substantial advancements have been made. Where explainability is not only desirable but also ethically and legally required, such a lack of transparency poses a serious threat in areas of sensitive applications such as healthcare robots, autonomous cars, and industrial robotics. The goal of XAI in robotics is to create AI systems that can not only function autonomously but also explain their reasoning to a human user [6].

Human considerations like trust, safety, and responsibility are integral to explainability, which extends beyond algorithmic clarity. Cooperation between humans and robots, such as in co-robots and assistive medical devices, relies on mutual understanding for optimal performance and safety. With explainability, human operators can better anticipate and understand a robot's activities, making it easier to rescue or fix any mistakes that may occur [7]. Building psychological trust between people and robots through open communication is crucial for releasing autonomous technology into the wild. In addition, the need for algorithmic accountability is being emphasised more and more by regulatory frameworks and ethical norms. This means that developers of AI systems are obligated to ensure that their operations are conducted on interpretable and auditable algorithms. Robotics and XAI convergence is thus an interdisciplinary issue spanning domains such as AI, ethics, cognitive science, and machine learning [8]. Also, keep in mind that safety, ethics, and regulations are all important factors to consider while evaluating an explainable robot. Clinical decision making, robot accuracy, and clinical trustworthiness should be the focus of explainable AI in medical robotics [9].

Structure of the paper

The following is the paper's structure: Explainable AI and surgical decision assistance are covered in Section II. Robotic accuracy and command are covered in Section III, whereas clinical reliability, validation, and implementation are covered in Section IV. The literature is reviewed and research needs are identified in Section V. Section VI concludes by summarising the results and outlining areas for future study.

I. EXPLAINABLE AI AND SURGICAL DECISION SUPPORT

XAI improves surgical decision assistance by providing an explanation and interpretation for the AI's recommendations. XAI is useful for surgeons in several ways, including pre-operative planning, post-operative risk assessment, and decision-making. Key feature identification methods include SHAP, LIME, and Grad-CAM.

A. Surgical Robotics and High-Fidelity Simulation Training

Surgical robots, such as the extensively used Da Vinci system, play a crucial role in minimally invasive treatments by providing surgeons with improved dexterity, precision, and high-definition 3D vision [10]. Comprehensive training on these systems is necessary due to the complexity and high risks of these procedures; simulation has developed as a critical component of surgical education.

Skill Acquisition and Mastery

Training using a simulation platform allows students to perfect essential skills in a risk-free setting before interacting with real patients. This includes both general knowledge of robotic surgery and detailed instructions for performing the procedure, paving the way for further study and improvement.

Early Exposure

Introducing medical students to robotic surgical settings during their preclinical and early clinical years can greatly increase their interest in the profession, demystify the technology, and provide the groundwork for more advanced training later in their careers [11]. West Virginia University School of Medicine and the University of California, San Francisco (UCSF) both have programs that show how this kind of early exposure makes students more comfortable with robotic technique and encourages them to be more involved in the operating room.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, October 2026
Impact Factor: 8.2

Reduced Learning Curve

The learning curve for robotic surgery procedures is notoriously high, but with the help of robotic surgery simulators—and especially VR ones—this may be significantly reduced [12]. Virtual reality (VR) instruction, as opposed to no extra training, considerably improved performance for junior trainees having laparoscopic cholecystectomy, as assessed by Objective Structured Assessment of Technical Skill (OSATS) scores and time to completion.

Transferability of Skills

There is proof that technical surgical skills learned in robotic virtual reality simulators can be successfully applied in the real operating theatre, and performance on these simulators can even forecast intraoperative performance in the real world. This demonstrates the value of simulation as a tool for clinical practice preparation.

B. AI technologies in robotics

Artificial intelligence (AI) is a great asset to medical robotics since it allows for smart observation, surgical decision-making, and instrument control and manipulation. For jobs requiring task autonomy, trajectory planning, or image processing, machine learning, deep learning, or reinforcement learning can be quite useful. A surgeon can better comprehend the robot's judgements with the help of explainable AI.

Machine Learning and Deep Learning

Robots can now learn from data to perform better, thanks to machine learning and deep learning [13]. Robots' perception and decision-making skills are enhanced by Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), which enable them to interpret visual and sequential input.

Reinforcement Lerning

Robots can develop optimum behaviour by interacting with their environment in a trial-and-error fashion, according to reinforcement learning (RL) [14]. RL algorithms like Proximal Policy Optimisation (PPO) and Deep Q-Networks (DQNs) have found extensive application in autonomous systems. Updates to model-based RL have made policy learning and sample efficiency even better.

AI-Powered Edge Computing

Robots can now make better decisions in real time because to edge computing, which allows them to perform AI algorithms locally. AI models deployed on edge devices are useful in mission-critical applications that require instantaneous responses, such as autonomous cars and industrial automation [15]. Edge computing and AI work together to make robotic applications more secure and private while decreasing their need on cloud-based services.

Federated Learning in Robotics

Federated learning ensures privacy and security by allowing several robots to learn together without exchanging raw data. Healthcare, logistics, and smart cities are just a few examples of how this decentralised AI strategy enhances efficiency [16]. Additionally, distributed robotic systems can improve their learning and jointly adapt to new obstacles through federated learning.

AI in Soft Robotics

Flexible robotic devices that imitate biological motions are controlled by artificial intelligence in soft robotics. Medicinal applications, fragile material handling, and human-robot interactions are some of the many uses for soft robots powered by artificial intelligence. Soft robotics and artificial intelligence work together to make robots more versatile and safer to work with people.

AI in Bio-Inspired Robotics

to improve efficiency and flexibility by imitating the behaviour of biological creatures. Among the many potential uses are underwater exploration robots inspired by insects, drones driven by artificial intelligence, and robotic fish. All thanks to AI algorithms, these robots are able to successfully traverse and engage with even the most complicated environments.

Quantum Computing for AI in Robotics

The ability to speed up optimisation and machine learning processes, which may completely transform artificial

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, October 2026
Impact Factor: 8.2

intelligence in robotics [17]. Autonomous systems can benefit from quantum-enhanced AI in several ways, including better route planning, simulation, and decision-making in real-time.

C. XAI Techniques for Surgical Intelligence

Artificial intelligence (AI) has been increasingly used in surgical intelligence as a result of advancements in medical imaging and deep learning [18]. However, doctors and surgeons may find it more difficult to understand AI-generated predictions if the algorithms are designed with hidden complexity. By highlighting key areas and characteristics, XAI approaches like SHAP, LIME, and Grad-CAM simplify algorithm explanations.

Local Interpretable Model-Agnostic Explanations (LIME)

The explainable AI method known as LIME (Explanation by Local Interpretability of Model's Explanations) approximates a complex machine learning model with a simpler model, such as linear regression, and then offers local explanations for the model's predictions. By identifying pertinent input variables, medical robotics may use LIME to clarify AI-based suggestions and risk forecasts for specific individuals.

Shapley Additive explanations (SHAP)

The SHAP technique clarifies the AI-generated forecast by calculating the influence of input characteristics on the model's output. When it comes to medical robotics, the SHAP technique is useful for finding clinical characteristics that impact post-operative risk assessment, complication prediction, and decision tool support [19]. To make deep learning models more transparent and to assist healthcare clinicians in validating judgements produced by AI models, SHAP is comparable to Grad-CAM and LIME.

Attention mechanisms

The ML algorithm may learn to prioritise and take into account the most significant information using the Attention Mechanism. In order for the attention mechanism algorithms to reveal which data was utilised for decision-making, it is necessary to be able to assign varying degrees of priority to various parts of pictures, sequences, or sensor data.

Gradient-weighted Class Activation Mapping (Grad-CAM)

The Grad-CAM is a method for explaining images that draws attention to the regions of the input picture that have an impact on the CNN model's predictions. Instead of depending on structural properties of feature maps, as is the case with Class Activation Mapping—a classic method for visualising deep learning models—Grad-CAM generates localisation maps using the input image's gradients in the final convolutional layer [20][21]. Surgeons using AI-assisted medical robots can benefit from explanations while analysing pictures.

III. EXPLAINABLE AI FOR ROBOTIC PRECISION AND CONTROL

The use of explainable feedback in artificial intelligence (AI) improves surgical precision by allowing for a better understanding of how the AI processes imaging data, formulates motion plans, and controls the robot's movements. The feedback provides surgeons with valuable insights into their control systems, allowing them to identify potential errors and maintain appropriate monitoring.

A. AI-Enhanced Surgical Precision

Maintaining healthy tissue, reducing collateral harm, and guaranteeing the best possible treatment results all depend on surgical precision [22]. Artificial intelligence helps with surgical accuracy in several ways.

Real-Time Imaging and Navigation

Anatomical landmarks may be identified in real time and 3D reconstructions can be generated using intraoperative data like video, CT, or MRI using imaging systems that include AI [23][24]. Neurosurgery and orthopaedic procedures, in particular, benefit from these devices' enhanced orienting capabilities. With DL models trained on annotated datasets, robots may operate within very limited anatomical boundaries, accurately segmenting organs or tumours to within a millimetre.

Adaptive Motion and Trajectory Planning

Artificial intelligence (AI) may modify the trajectory of robotic arms in response to changes in tissue resistance or unforeseen anatomical variances by utilising haptics and force feedback as sensor inputs. Robots may gradually

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, October 2026
Impact Factor: 8.2

become better over time without any help from humans thanks to algorithms that teach them new patterns of movement via trial and error. Case in point: cardiac ablation.

Microsurgical Assistance

Microsurgical techniques in neurology, ENT, and ophthalmology need extraordinary accuracy. The Preceyes Surgical device is one example of an AI-enhanced robotic device that may reduce surgeon jitters and carry out precise motions. Furthermore, AI aids in instrument control with submillimeter accuracy, which is more than human dexterity can manage for such sensitive tasks.

B. Safety Improvements through AI Integration

The use of artificial intelligence (AI) is essential in surgical robots for the purpose of patient safety [25]. The ability to continuously learn and monitor allows robots to identify any issues or changes from typical surgical patterns.

Intraoperative Monitoring and Risk Prediction

Artificial intelligence systems use information gathered during surgery, including blood loss, vital signs, and tool trajectories, to foresee potential problems before they have a chance to do serious harm. Surgeons can be alerted to possible complications like as bleeding, organ perforation, or anaesthesia concerns by using predictive algorithms trained on huge datasets [26]. For example, AI may identify unusual tool movements that may indicate dangerous handling or procedural problems in robotic-assisted laparoscopy.

Automated Error Detection and Correction

Artificial intelligence (AI) may identify irregularities such tool misalignment, excessive force application, or unusual trajectories through real-time video analysis and sensor input. Automatic or haptic feedback-guided corrections can be carried out once discovered. These features are particularly helpful in telesurgery because of the potential for inadvertent injury due to delay or visual misalignment.

Simulation and Training

AI is also revolutionising surgical education by creating lifelike simulations that monitor performance indicators and provide customised feedback. Intelligent simulators like this evaluate skills including tool handling, procedural correctness, and hand-eye coordination using machine learning. This makes them great resources for training purposes, since they shorten the learning curve and make students more prepared to perform robotic procedures in the actual world.

IV. CLINICAL TRUSTWORTHINESS, VALIDATION, AND DEPLOYMENT OF EXPLAINABLE MEDICAL ROBOTICS

Reliability, openness, safe patient care, and enough human oversight are the bedrocks of explainable medical robots' credibility in the clinic. The robustness, accuracy, clarity, and equity of the clinical explanations are evaluated throughout the validation procedure. Concerns like monitoring, data protection, ethical governance, and regulatory compliance all contribute to responsible integration.

A. Trustworthy Medical Robotics

The advent of data-centric robotic ecosystems powered by artificial intelligence and enhanced by new digital technologies has the potential to revolutionise healthcare by facilitating human-centered, efficient, adaptable, and seamless clinical processes [27]. Significant improvements in diagnostics, individualised therapies, and operational efficiency are on the horizon thanks to these breakthroughs.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, October 2026
Impact Factor: 8.2

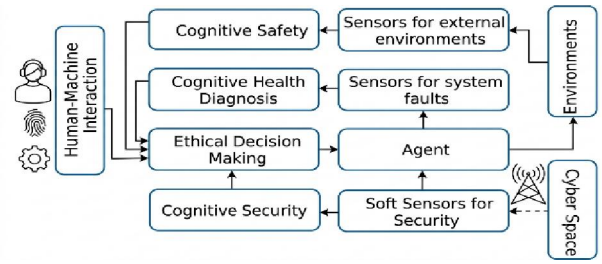


Fig. 1. Trustworthy Medical Robotics

The TAS paradigm emphasises the critical role of AI in improving cognitive processes associated with security, safety, system health, and human-robot interaction, as shown in Fig. 1 (modified from) [28]. Problems with system complexity and ethical decision-making are exacerbated by edge AI deployment, especially in robotic systems that engage directly with patients.

Layer-based Approach for MR Ecosystems

A layered architectural approach is suggested for the purpose of building trustworthy MR ecosystems, which will systematically handle regulatory compliance, explainability, safety, and dependability across all subsystems [29]. The five interconnected levels of the suggested design are as follows, and they all work together to provide robotic healthcare solutions that are human-centered and ethically aligned:

- **Hardware Layer:** Actuators, sensors, embedded systems, and flexible mechanical structures make up the hardware layer, which is the physical basis of the MR system [30].
- **Perception and Cognition Layer:** Environmental awareness, monitoring of system health, cognitive cybersecurity, and cognitive reasoning are all responsibilities of this AI-driven layer [31].
- **Control Layer:** The control system of a robot operates within guaranteed limits in terms of stability, accuracy, responsiveness, and robustness, even under uncertainties or disturbances. It also manages real-time responsiveness, coordination, and low-level motion control with bounded performance. Recovery protocols are also in place to maintain system stability under all conditions.
- **Human–Machine Interaction (HMI) Layer:** Interactions between patients and doctors may be secure, easy, and transparent thanks to this shield. Collaboration is enhanced through the use of shared autonomy frameworks and multimodal interface support [32].
- **Data and Communication Layer:** The interoperable, safe, and efficient transfer of data is guaranteed by this layer. Through encryption, access management, and conformity with GDPR and HIPAA regulations, it enables secure real-time communication between cloud services and edge devices [33].

Five Pillars of Trustworthy MR Ecosystems

A reliable medical robot goes beyond being an intelligent machine; it is a healthcare-specific system that reliably carries out its tasks in accordance with the utmost standards of security, honesty, ethics, and human-machine alignment. These five principles are these:

- **Safety and Runtime Assurance:** One of the biggest obstacles in medical robots is making sure it can rely on patients in high-stakes situations [34]. Sensor drift, hardware failures, and patient mobility are some of the dynamic and unpredictable clinical circumstances in which these systems must function.
- **Security and privacy:** Reliable medical robots that handle private patient information in intricate healthcare settings must adhere to stringent cybersecurity and privacy standards [35]. Protecting information against hostile assaults that alter perception or decision-making, as well as safeguarding data across cloud and edge systems, are important obstacles.
- **Accuracy, Bias, and Interpretability:** Transparency, objectivity, and accuracy are ongoing concerns with AI-powered healthcare robots. Unfortunately, under-represented groups may suffer disproportionately from unequal treatment due to biases in algorithm design or training data [36].

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, October 2026
Impact Factor: 8.2

- **Human-Robotics Interaction and Ethical Alignment:** The most reliable human-robot interactions in healthcare will be those that are intuitive, welcoming, and morally congruent. Adapting to crises and impaired cognition is a constant challenge for systems that must accommodate a wide range of physical and cognitive capacities [37].
- **Lifecycle Governance:** The ability to have faith in medical robotic systems is an ongoing process, not an endpoint, and it must be nurtured as the system evolves [38].

B. Professional Brain

Clinical HRC's cognitive layer is symbolised by the professional brain. Using its interpretation capabilities, it is able to transform intricate medical data into actionable instructions for robotic systems [39]. This capability is backed by three fundamental technologies: clinical reasoning, context modelling, and multimodal perception. Being able to function in healthcare environments that are both dynamic and unpredictable.

- **Multimodal Perception:** The ability to see through several channels rather than just one enables professionals to combine disparate pieces of clinical data [40]. Annotation, curation, and data acquisition of the highest quality are crucial to its progress.
- **Context Modeling:** In order to comprehend the patient's situation thoroughly, it is necessary to examine their medical history, current clinical status, and previous medical knowledge [41].
- **Clinical Reasoning:** The trained mind to function in hostile and unpredictable healthcare settings [42]. Outputs should be visible, context-aware, and acceptable for human confirmation, resulting from a combination of data-driven inference and explicit clinical limitations.

V. LITERATURE REVIEW

Enhancing transparency, dependability, and confidence in medical robots is the goal of explainable artificial intelligence. Current research delves into topics such as interpretable machine learning, brain-computer interfaces, artificial intelligence-assisted surgery, and robotic healthcare applications. Supporting therapeutic decision-making, surgical accuracy, and human-robot collaboration are these techniques.

B. Brahma et al. (2026) investigated the practicality and dependability of applying interpretable artificial intelligence (IAI) techniques to medical robots in order to increase the pace of adoption of these technologies in situations when resources are scarce. Automated medical treatments can be more open and trustworthy with the use of XML and IAI approaches, which detail how ML and AI systems make decisions. The results may then be confirmed by medical professionals and doctors because of this. Medical robots empower healthcare providers to make better judgements by bridging the gap between technology and human intelligence through the use of XML and IAI [43].A. Abdullah *et al.* (2026) examined MI-BCIs (Brain Computer Interfaces based on Motor Imagery) for its dependability, robustness, and explainability. MI-BCIs show great promise in the fields of neurorehabilitation, assistive robotics, and smart environments because to their ability to control external devices through imagined motions. Signal reliability, evaluation variables, and the absence of established assessment frameworks all work against their practical use. We combed through 43 papers published in scholarly journals and found that they fell into six categories: AI that can be explained, healthcare, neurorehabilitation, smart settings, trustworthy and secure MI-BCIs, and learning frameworks [44].

M. Shaikh et al. (2025) Multi-phase optimisation is the suggested strategy, which includes preoperative planning, intraoperative support, and postoperative monitoring. A crucial component is adaptive learning using federated and XAI models, but the gathering of data from different biological sensors with AI-driven diagnostic inference is also important. In addition, the approach optimises surgical precision while decreasing human error by using state-of-the-art imaging and haptics and integrating with the IoMT. The suggested method for conducting performance benchmarking under clinical restrictions is to use real-time validation and simulation-based testing [45].A. Nazar (2025) discusses in depth the theoretical underpinnings and consequences of XAI for robots in relation to the transparency-performance trade-off. It delves further into the topic by investigating the latest methods of explainability in deep learning-based

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, October 2026
Impact Factor: 8.2

robots, which place an emphasis on the significance of user interpretability, ethical assurance, and immediate feedback. In order to achieve trust, legal compliance, and responsible innovation in next-generation autonomous robots, the article emphasises that explainability is an essential component, not only a technological one [46].

C. Zhang et al. (2024) provide fresh approaches that utilise AI to merge preoperative imaging with surgery in order to circumvent these constraints and progress robotic procedures even further. Machine learning is known for its speed and its ability to handle massive volumes of messy, unstructured data. Throughout the perioperative phase, its widespread usage will significantly impact practice. By gaining access to pathology data that has historically only been accessible postoperatively, evaluation of preoperative imaging can enhance preoperative surgeon expertise [47].Y. Liu et al (2024) provided a background on the evolution of AI-supported SRS, reviewed the fundamental architecture of SRS, and examined the ways in which AI may enhance SRS functionality. Case studies of AI-assisted SRS, its effects in therapeutic contexts, and related legal and ethical issues are examined. Lastly, the difficulties with AI-assisted SRS are examined, including algorithm creation, surgical data science, human-robot collaboration, and fostering human-robot trust [48].

N. Deo *et al.* (2023) AI integration with healthcare robots has been on an upward trajectory for some time now. Using AI, hospitals in both urban and rural areas might greatly improve the efficiency of many functions, such as assisting with surgeries, improving hospital logistics, and performing routine examinations. One area where AI might be useful in healthcare is in guaranteeing individualised attention for patients, creating a risk-free workplace for medical professionals, and streamlining administrative processes [49].

A number of knowledge gaps persist despite the fact that explainable AI and medical robotics have made great strides in recent years. Research on explainability, AI-assisted surgical systems, brain-computer interfaces, and robotic healthcare applications is sparse, and there has been no effort to integrate these topics into a coherent, clinically verified framework. Additional research is needed to identify the trade-offs between transparency and performance, as well as the practical dependability of explanations and their interpretability in real-time. Concerning the feasibility, scalability, and uptake of explainable robotic systems, there is little data for healthcare settings with minimal resources. There is also a lack of well-established standard measures for evaluating explanation quality, robustness, safety, and human-robot collaboration. Additionally, there is a lack of trust when it comes to adopting these technologies in different medical settings due to the limited clinical validation. The necessity for reliable, understandable, and therapeutically useful medical robots is emphasised in Table I, which summarises the comparative research gaps and future directions

TABLE I: SUMMARY OF KEY STUDIES ON EXPLAINABLE ARTIFICIAL INTELLIGENCE IN MEDICAL ROBOTICS

Author & Year	Objective / Approach	Method / Technology	Key Findings	Advantages	Limitations
B. Brahma et al. (2026)	Examined the viability of explainable AI in medical robotics, particularly in resource-limited settings.	Explainable Machine Learning (XML) and Interpretable AI (IAI).	Explainability improves transparency and helps healthcare professionals understand and verify robotic decisions.	Supports trust, interpretability, and informed clinical decisions.	Practical validation and adoption in resource-limited healthcare settings require further investigation.
A. Abdullah et al. (2026)	Investigated reliability, robustness, and explainability in motor imagery-based brain–computer	Systematic analysis of 43 studies covering MI-BCIs, robotics, neurorehabilitation, XAI, and learning frameworks.	Identified challenges in signal reliability, evaluation consistency, and standardized assessment.	Supports trustworthy brain-controlled assistive and rehabilitation systems.	Standardized evaluation frameworks and robust real-world validation remain challenging.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 7, October 2026 Impact Factor: 8.2

	interfaces.				
M. Shaikh et al. (2025)	Proposed a multi-phase AI-driven framework for improving surgical precision and monitoring.	Biomedical sensors, federated learning, XAI, medical imaging, haptics, and IoMT.	Integrated diagnostic inference and adaptive learning across preoperative, intraoperative, and postoperative stages.	May reduce human error and support robotic-assisted orthopedic and microsurgical procedures.	Clinical performance requires simulation-based testing and real-time validation.
A. Nazar (2025)	Examined explainability in robotics and the trade-off between transparency and performance.	Deep learning explainability, user-centered interpretation, and real-time feedback.	Highlighted explainability as a factor in trust, ethical confidence, and responsible robotic autonomy.	Encourages transparent decisions and human–robot trust.	Balancing interpretability, computational performance, and real-time operation remains difficult.
C. Zhang et al. (2024)	Investigated AI integration into preoperative imaging and surgical workflows.	AI-based imaging analysis and perioperative decision support.	AI can extract clinically useful information from preoperative images and improve surgical planning.	Enhances preoperative assessment and supports informed surgical decisions.	Explainability-specific evaluation and generalization across clinical settings need further study.
Y. Liu et al. (2024)	Reviewed the development, architecture, applications, and challenges of AI-assisted surgical robotic systems.	AI-assisted surgical robotics, surgical data science, and human–robot coordination.	Identified algorithm development, clinical evidence, ethics, legal concerns, and trust as key considerations.	Provides an overview of system design and clinical integration challenges.	Human–robot coordination and trust building remain unresolved challenges.
N. Deo et al. (2023)	Reviewed the growing use of AI and robotics in healthcare services.	AI-enabled surgical assistance, hospital logistics, and routine patient checkups.	Highlighted potential benefits for patient care, healthcare-worker safety, and hospital operations.	Supports efficient healthcare delivery in urban and rural hospitals.	Limited focus on XAI techniques and the interpretability of robotic decisions.

VI. CONCLUSION AND FUTURE WORK

Medical robotics is rapidly evolving to incorporate explainable AI, which enhances robotic precision, clinical trustworthiness, and surgical decision support. Surgery suggestions can be better understood with the use of techniques like Grad-CAM, attention processes, SHAP, and LIME, which help comprehend predictions provided by AI. Robotic manipulation, intraoperative monitoring, adaptive trajectory planning, and anatomical identification are all made possible with the combination of ML, DL, and RL. Improvements in surgical competence and patient safety can be achieved via the use of simulation-based training and automated mistake detection systems. Integrating explainability,

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, October 2026
Impact Factor: 8.2

cybersecurity, data privacy, justice, ethical alignment, and human-robot interaction are crucial components of trustworthy medical robotic systems. Reliability of explanations, efficiency of calculation in real time, clinical validation, and compliance with regulations are still areas that need work despite these advancements. To guarantee the safe and transparent operation of AI-driven robotic systems in complicated healthcare settings, it is crucial to address these restrictions. Supporting informed surgical decisions and promoting responsible adoption in contemporary healthcare may be achieved by integrating explainable intelligence with dependable robotic control.Research in the future should focus on improving surgical robots using XAI approaches that are both practical and reliable in real-time, and then testing these methods across a variety of procedures and patient populations. Robust robotic control, human-centered interfaces, and uncertainty estimates should all be part of future research. For reliable implementation, multicenter clinical validation, standardised assessment measures, and ongoing lifecycle monitoring are crucial.

REFERENCES

[1] P. Kumar, "Edge Computing and IoT for Real-Time Healthcare Data Processing and Integration," in *2025 4th International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, IEEE, Dec. 2025, pp. 105–110. doi: 10.1109/ICAAIC64647.2025.11331211.

[2] V. Bagga and D. Bhattacharyya, "Robotics in neurosurgery," *Ann. R. Coll. Surg. Engl.*, vol. 100, no. 6, pp. 23–26, 2018, doi: 10.1308/rcsann.supp1.19.

[3] T. Ginoya, Y. Maddahi, and K. Zareinia, "A Historical Review of Medical Robotic Platforms," *J. Robot.*, vol. 2021, no. 1, pp. 1–13, Jan. 2021, doi: 10.1155/2021/6640031.

[4] A. Adadi and M. Berrada, "Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018, doi: 10.1109/ACCESS.2018.2870052.

[5] N. Shao, Q. Zhou, C. Shao, and Y. Zhao, "Adaptive Control of Robot Series Elastic Drive Joint Based on Optimized Radial Basis Function Neural Network," *Int. J. Soc. Robot.*, vol. 13, no. 7, pp. 1823–1832, 2021, doi: 10.1007/s12369-021-00762-0.

[6] L. Wen, Y. Liu, and H. Li, "CL-MAPF: Multi-Agent Path Finding for Car-Like robots with kinematic and spatiotemporal constraints," *Rob. Auton. Syst.*, vol. 150, p. 103997, Apr. 2022, doi: 10.1016/j.robot.2021.103997.

[7] W. Samek, G. Montavon, S. Lapuschkin, C. J. Anders, and K.-R. Müller, "Explaining Deep Neural Networks and Beyond: A Review of Methods and Applications," *Proc. IEEE*, vol. 109, no. 3, pp. 247–278, 2021, doi: 10.1109/JPROC.2021.3060483.

[8] A. Barredo Arrieta *et al.*, "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion*, vol. 58, pp. 82–115, Jun. 2020, doi: 10.1016/j.inffus.2019.12.012.

[9] N. K. S. Patel, "Enterprise AI Operations in Healthcare IT Infrastructure," *J. Inf. Syst. Eng. Manag.*, vol. 11, no. 3, pp. 2508–2522, 2026, doi: 10.5281/zenodo.22773821.

[10] M. Jeyaraman, N. Jeyaraman, S. Ramasubramanian, S. Murugan, and V. Vijayaragavan, "Review of Robotics in Healthcare Education Innovations, Challenges, and Future Directions," in *2025 IEEE 6th International Conference in Robotics and Manufacturing Automation (ROMA)*, IEEE, Aug. 2025, pp. 107–112. doi: 10.1109/ROMA66616.2025.11155836.

[11] A. L. Greenberg, S. M. Syed, A. Alseidi, P. S. O'Sullivan, and H. Chern, "Robotic training for medical students: feasibility of a pilot simulation curriculum.," *J. Robot. Surg.*, vol. 17, no. 3, pp. 1029–1038, Jun. 2023, doi: 10.1007/s11701-022-01508-4.

[12] G. Humm *et al.*, "The impact of virtual reality simulation training on operative performance in laparoscopic cholecystectomy: meta-analysis of randomized clinical trials," *BJS Open*, vol. 6, no. 4, pp. 1–15, Jul. 2022, doi: 10.1093/bjsopen/zrac086.

[13] K. Gowri, R. U. Rishaba, R. Nikkitha, and M. Swathi, "Artificial Intelligence-Driven Robotics and Advanced Automation Systems," *Int. J. Sci. Eng. Sci.*, vol. 9, no. 2, pp. 10–13, 2025.

[14] J. Kober, J. A. Bagnell, and J. Peters, "Reinforcement learning in robotics: A survey," *Int. J. Rob. Res.*, vol. 32,

no. 11, pp. 1238–1274, 2013, doi: 10.1177/0278364913495721.

[15] J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement," Apr. 2018. doi: 10.48550/arXiv.1804.02767.

[16] G. M. Gutierrez, J. A. Rincon, and V. Julian, "Federated Learning for Collaborative Robotics: A ROS 2-Based Approach," *Electronics*, vol. 14, no. 7, p. 1323, Mar. 2025, doi: 10.3390/electronics14071323.

[17] V. Chaturvedi, M. Junaid, T. Yadav, and Y. Perwej, "Quantum Computing: AI Integration for Advanced Robotic Learning and Decision-Making," *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol.*, vol. 12, no. 2, pp. 385–393, Apr. 2026, doi: 10.32628/CSEIT26121362.

[18] M. Fayaz, K. Hagsong, S. Danish, L. M. Dang, A. Sadeghi-Niaraki, and H. Moon, "Demystifying Artificial Intelligence: A Systematic Review of Explainable Artificial Intelligence in Medical Imaging," *Sensors*, vol. 26, no. 7, pp. 1–29, Mar. 2026, doi: 10.3390/s26072131.

[19] I. Haq *et al.*, "A review of breast cancer histopathology image analysis with deep learning: Challenges, innovations, and clinical integration," *Image Vis. Comput.*, vol. 162, no. July, p. 105708, 2025, doi: 10.1016/j.imavis.2025.105708.

[20] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning Deep Features for Discriminative Localization," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2016, pp. 2921–2929. doi: 10.1109/CVPR.2016.319.

[21] P. Ramaraj and M. Mittal, "Deep Learning Models for Automated Medical Imaging: Accuracy, Ethics, And Deployment Challenges," *Int. J. Artif. Intell. Deep Learn.*, vol. 2, no. 1, pp. 11–31, Jun. 2025, doi: 10.34218/IJAIDL_02_01_002.

[22] J. Mathew and S. Sheikh, "AI in Surgical Robotics: A Review of Precision, Safety, and Postoperative Outcomes Sultan Sheikh," 2025, pp. 75–81, June.

[23] J. N. K. Wah, "Revolutionizing surgery: AI and robotics for precision, risk reduction, and innovation.," *J. Robot. Surg.*, vol. 19, no. 1, p. 47, Jan. 2025, doi: 10.1007/s11701-024-02205-0.

[24] S. Mukherjee, "AI-Driven Deep Learning System for Alzheimer's Disease (AD) Detection Via MRI Imaging," in *2026 International Conference on Intelligent Computing, Networks, and Security (IC-ICNS)*, Bhubaneswar, India: IEEE, 2026, pp. 1–6, March. doi: 10.1109/IC-ICNS68863.2026.11537765.

[25] R. S. Kedar and C. G. Puri, "Enhanced Precision and Safety in Minimally Invasive Surgery: The Role of AI and Robotics Integration," in *2024 4th International Conference on Ubiquitous Computing and Intelligent Information Systems (ICUIS)*, IEEE, Dec. 2024, pp. 231–237. doi: 10.1109/ICUIS64676.2024.10866464.

[26] K. Reddy, P. Gharde, H. Tayade, M. Patil, L. S. Reddy, and D. Surya, "Advancements in Robotic Surgery: A Comprehensive Overview of Current Utilizations and Upcoming Frontiers.," *Cureus*, vol. 15, no. 12, p. e50415, Dec. 2023, doi: 10.7759/cureus.50415.

[27] H. He *et al.*, "Towards trustworthy medical robotics ecosystems," *IET Conf. Proc.*, vol. 20, no. 50, pp. 79–93, Apr. 2026, doi: 10.1049/icp.2026.1127.

[28] H. He, J. Gray, A. Cangelosi, Q. Meng, T. M. McGinnity, and J. Mehnen, "The Challenges and Opportunities of Human-Centered AI for Trustworthy Robots and Autonomous Systems," in *IEEE Transactions on Cognitive and Developmental Systems*, IEEE, Dec. 2022, pp. 1398–1412. doi: 10.1109/TCDS.2021.3132282.

[29] H. J. Marcus *et al.*, "The IDEAL framework for surgical robotics: development, comparative evaluation and long-term monitoring," *Nat. Med.*, vol. 30, no. 1, pp. 61–75, 2024, doi: 10.1038/s41591-023-02732-7.

[30] U. Martinez-Hernandez *et al.*, "Human-in-the-loop layered architecture for control of a wearable ankle–foot robot," *Rob. Auton. Syst.*, vol. 161, p. 104353, Mar. 2023, doi: 10.1016/j.robot.2022.104353.

[31] M. Chen, S. Nikolaidis, H. Soh, D. Hsu, and S. Srinivasa, "Trust-Aware Decision Making for Human-Robot Collaboration: Model Learning and Planning," *J. Hum.-Robot Interact.*, vol. 9, no. 2, Jan. 2020, doi: 10.1145/3359616.

[32] R. H. Wortham, A. Theodorou, and J. J. Bryson, "Improving robot transparency: Real-time visualisation of robot

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 7, October 2026
Impact Factor: 8.2

AI substantially improves understanding in naive observers," in *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, IEEE, Aug. 2017, pp. 1424–1431. doi: 10.1109/ROMAN.2017.8172491.

[33] A. D. Dwivedi, G. Srivastava, S. Dhar, and R. Singh, "A Decentralized Privacy-Preserving Healthcare Blockchain for IoT," *Sensors*, vol. 19, no. 2, p. 326, Jan. 2019, doi: 10.3390/s19020326.

[34] G.-Z. Yang *et al.*, "Medical robotics—Regulatory, ethical, and legal considerations for increasing levels of autonomy," *Sci. Robot.*, vol. 2, no. 4, Mar. 2017, doi: 10.1126/scirobotics.aam8638.

[35] A. J. London, "Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability.," *Hastings Cent. Rep.*, vol. 49, no. 1, pp. 15–21, Jan. 2019, doi: 10.1002/hast.973.

[36] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations.," *Science*, vol. 366, no. 6464, pp. 447–453, Oct. 2019, doi: 10.1126/science.aax2342.

[37] S. Tonekaboni, S. Joshi, M. D. McCradden, and A. Goldenberg, "What Clinicians Want: Contextualizing Explainable Machine Learning for Clinical End Use," *Proc. Mach. Learn. Res.*, vol. 106, pp. 1–21, Aug, 2019, doi: 10.48550/arXiv.1905.05134.

[38] M. Ribeiro, S. Singh, and C. Guestrin, "'Why Should I Trust You?': Explaining the Predictions of Any Classifier," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2016, pp. 97–101. doi: 10.18653/v1/N16-3020.

[39] X. Wu *et al.*, "Toward Autonomous Clinics: Human–Robot Collaboration in Clinical Care," *MedComm*, vol. 7, no. 8, 2026, doi: 10.1002/mco2.70885.

[40] J. Yang *et al.*, "MedMNIST v2 - A large-scale lightweight benchmark for 2D and 3D biomedical image classification," *Sci. Data*, vol. 10, no. 1, p. 41, Jan. 2023, doi: 10.1038/s41597-022-01721-8.

[41] M. Wen, Q. Zhou, B. Tao, P. Shcherbakov, Y. Xu, and X. Zhang, "Short-term and long-term memory self-attention network for segmentation of tumours in 3D medical images," *CAAI Trans. Intell. Technol.*, vol. 8, no. 4, pp. 1524–1537, 2023, doi: https://doi.org/10.1049/cit2.12179.

[42] X. Meng *et al.*, "The application of large language models in medicine: A scoping review," *iScience*, vol. 27, no. 5, p. 109713, May 2024, doi: 10.1016/j.isci.2024.109713.

[43] B. Brahma, N. Kumar Kamila, S. Kumar Pani, and P. P. Kar, "Chapter 13 - Use of explainable machine learning and interpretable artificial intelligence in medical robotics," in *Medical Robotics and Intelligent Healthcare Technologies*, 2026, pp. 339–369. doi: 10.1016/B978-0-443-24766-8.00015-4.

[44] A. A. Abdullah, K. A. Zidan, and A. S. Albahri, "A Systematic Review of Trustworthy and Explainable AI Frameworks for Motor ImageryBased Brain-Computer Interfaces in Robotic Control Systems," *Eng. Technol. Appl. Sci. Res.*, vol. 16, no. 1, pp. 30725–30736, 2026, doi: 10.48084/etasr.14778.

[45] M. S. Shaikh, N. K. Jain, N. Patel, D. Das, A. Mahesh, and S. Thakor, "Enhancing Surgical Precision and Recovery: A Review of Medical Robotics and AI Integration in Modern Operating Rooms," *EPJ Web Conf.*, vol. 328, pp. 1–12, 2025, doi: 10.1051/epjconf/202532801016.

[46] A. Nazar, "Explainable Artificial Intelligence in Robotics: Building Transparent and Trustworthy Autonomous Systems," *Int. J. Adv. Innov. Res.*, vol. 1, no. 2, pp. 35–42, 2025, doi: 10.63056/ijair.1.2.2025.94.

[47] C. Zhang, M. S. Hallbeck, H. Salehinejad, and C. Thiels, "The integration of artificial intelligence in robotic surgery: A narrative review," *Surgery*, vol. 176, no. 3, pp. 552–557, Sep. 2024, doi: 10.1016/j.surg.2024.02.005.

[48] Y. Liu *et al.*, "Evolution of Surgical Robot Systems Enhanced by Artificial Intelligence: A Review," *Adv. Intell. Syst.*, vol. 6, no. 5, pp. 1–25, May 2024, doi: 10.1002/aisy.202300268.

[49] N. Deo and A. Anjankar, "Artificial Intelligence With Robotics in Healthcare: A Narrative Review of Its Viability in India," *Cureus*, vol. 15, no. 5, pp. 1–9, May 2023, doi: 10.7759/cureus.39416.