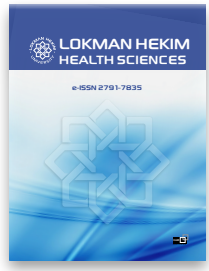
LOKMAN HEKIM HEALTH SCIENCES

DOI: 10.14744/lhhs.2026.20865
Lokman Hekim Health Sci 2026;6(2):00–00



lokmanhekimhs.com

SYSTEMATIC REVIEW

Artificial Intelligence in Neurosurgical Education: A Systematic Review of Technical Skills Training, Clinical Reasoning, and Surgical Planning

Ömer Selçuk Şahin, Samet Dinç

Department of Neurosurgery, Etlik City Hospital, Ankara, Türkiye

Abstract

Introduction: Artificial intelligence (AI) and machine learning (ML) are increasingly used in neurosurgical education to mitigate limitations of apprenticeship-based training (restricted operative exposure, duty-hour constraints) and to enable objective, scalable competency assessment. This systematic review synthesized and critically appraised evidence on AI/ML applications for technical skills training, clinical reasoning support, and surgical planning.

Methods: Following Preferred Reporting Items for Systematic Reviews and Meta-analyses 2020, we searched seven databases (SciSpace Deep Review, SciSpace Basic Search, SciSpace Full-Text Search, Web of Science Core Collection, PubMed, Google Scholar, and arXiv) for English-language, peer-reviewed studies published January 2010–January 2026. Two reviewers independently screened records, extracted data, and assessed risk of bias using design-appropriate appraisal tools. Given methodological heterogeneity, a narrative synthesis was conducted.

Results: From 789 records, 36 studies met the inclusion criteria. Most focused on technical skills training (69.4%), followed by surgical planning (27.8%); fewer evaluated clinical reasoning support. AI-based assessment systems differentiated expertise with 83–100% accuracy. AI-augmented tutoring and feedback systems yielded improvements comparable to expert instruction (effect sizes 0.20–0.66). Common limitations included small sample sizes, single-center designs, and limited external validation.

Discussion and Conclusion: AI/ML technologies demonstrate clinically meaningful benefits for neurosurgical technical skills training. Cognitive and decision-support applications remain less mature and require multi-institutional validation, standardized outcomes, and longitudinal evaluation to support broader curricular integration.

Keywords: Artificial intelligence; Clinical decision-making; Competency-based medical education; Machine learning; Neurosurgical education; Simulation training

Neurosurgery is regarded as one of the most technically demanding and cognitively complex medical specialties. It requires advanced psychomotor coordination, refined visuospatial skills, and high-stakes clinical decision-making under conditions of uncertainty. Neurosurgical education has long relied predominantly on apprenticeship-based training models, in which expertise is acquired through gradual clinical exposure under the

Cite this article as: *Şahin ÖS, Dinç S. Artificial Intelligence in Neurosurgical Education: A Systematic Review of Technical Skills Training, Clinical Reasoning, and Surgical Planning. Lokman Hekim Health Sci 2026;6(2):00–00.*

Correspondence: Ömer Selçuk Şahin, M.D. Etlik Şehir Hastanesi, Nöroşirürji Kliniği, Ankara, Türkiye
E-mail: oselcuks@gmail.com **Submitted:** 23.02.2026 **Revised:** 23.03.2026 **Accepted:** 31.03.2026 **Available Online:** 03.06.2026

guidance of experienced professionals. Despite its efficacy, this paradigm is increasingly constrained by several factors. Chief among these are duty-hour regulations, heightened patient-safety expectations, growing procedural complexity, and variability in operative case exposure, particularly for rare or high-risk procedures.[1,2]

In response to these structural challenges, artificial intelligence (AI) and machine learning (ML) technologies have emerged as promising tools to augment neurosurgical education. AI systems can process high-dimensional data, identify latent performance patterns, and generate adaptive feedback at scale. Within the domain of neurosurgical training, the potential of AI applications has been thoroughly investigated across a wide range of educational settings. These applications encompass a diverse array of functions, including simulation-based technical skills training, automated performance assessment, intelligent tutoring systems, surgical planning support, and early cognitive decision-support tools.[1–35]

Simulation-based education has emerged as a pivotal component of contemporary neurosurgical training, offering controlled, standardized environments conducive to skill acquisition. Virtual reality (VR) and mixed reality (MR) platforms have demonstrated particular utility in neurosurgical education, enabling immersive anatomical visualization and procedural rehearsal in complex scenarios such as skull base tumor resection.[36–38] The integration of AI into these VR and simulation platforms has led to the rapid evolution of these tools from formative practice instruments to objective competency assessment instruments. A multitude of studies have demonstrated the capacity of ML-driven assessment systems to accurately differentiate levels of surgical expertise by meticulously analyzing kinematic, temporal, force-based, and video-derived metrics.[2,6–9,11,16,30] These systems offer advantages over traditional subjective rating scales in consistency, reproducibility, and granularity of feedback, while AI-augmented VR simulators provide standardized training experiences that address variability in clinical case exposure.[36,37]

Beyond episodic scoring, deep learning-based approaches enable continuous monitoring of surgical performance and analysis of the learning curve, allowing objective tracking of skill progression over time.[3,6,16] These competencies are closely aligned with competency-based medical education frameworks, which emphasize progression based on demonstrated performance rather than time-based criteria.[4,5] The findings of randomized controlled trials (RCTs) further indicate that AI-augmented intelligent tutoring systems can achieve learning outcomes comparable to, and in some cases exceeding, those obtained through traditional expert instruction.[4,8,10,13]

Concurrently with the cultivation of technical proficiencies, there has been a marked surge in the exploration of AI applications in neurosurgical planning and decision support. ML models that incorporate imaging, demographic, and clinical variables have demonstrated greater predictive accuracy than clinician judgment alone, particularly when used as adjunctive tools.[18,21,26] From an educational perspective, exposure to such systems may introduce trainees to data-driven planning paradigms and support reflective comparison between human and algorithmic reasoning.

Despite rapid growth in the field, the extant literature on AI in neurosurgical education remains heterogeneous and underdeveloped. The extant literature focuses predominantly on technical skills training and assessment, while there is a paucity of research on applications that explicitly target clinical reasoning as an independent educational objective.[14,15,34] Furthermore, numerous studies are constrained by several factors, including limited sample sizes, single-center designs, reliance on simulated outcomes, and inadequate external validation of ML models.[6–9,14,17,34] The present systematic review addresses this gap by synthesizing and critically appraising evidence from 35 included studies on AI and ML applications in neurosurgical education.

Materials and Methods

Study Design and Reporting Framework

This systematic review was conducted and reported in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 guidelines. A protocol was developed a priori to define objectives, eligibility criteria, information sources, study selection procedures, and the synthesis approach.[39] Ethical approval was deemed unnecessary in this instance; the scope was confined to analyzing previously published literature.

Information Sources and Search Strategy

In January of 2026, a series of searches was conducted. The PubMed (MEDLINE) and Web of Science (WoS) Core Collection were identified as the primary databases for records reported in the PRISMA flow diagram. Concurrently, Google Scholar and SciSpace were used to enhance sensitivity and validate coverage.

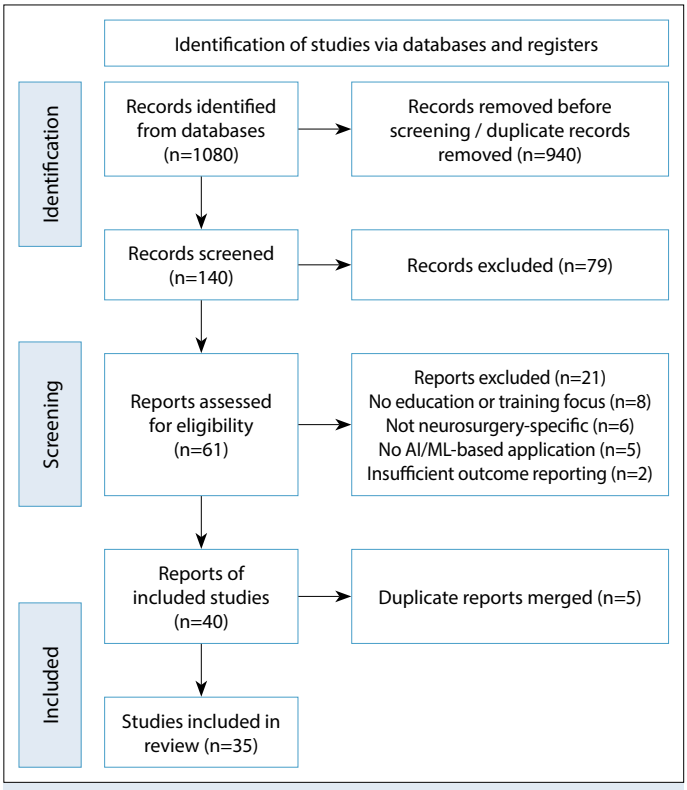


Figure 1. PRISMA 2020 flow diagram for the study selection process. A total of 1080 records were identified through database searches, with no records identified from registers. After removal of 940 duplicate records, 140 records were screened by title and abstract, of which 79 were excluded. Full-text retrieval was sought for 61 reports, all of which were successfully retrieved and assessed for eligibility. Following full-text assessment, 21 reports were excluded with reasons, resulting in 35 studies included in the final qualitative synthesis.

Search strings combined controlled vocabulary (when available) and free-text terms across four concepts: The following areas of study are of particular interest: AI/ML, neurosurgery, education/training, populations, and simulation/VR/augmented reality (AR) or objective technical skill assessment. A variety of strategies were employed; each adapted to the specific indexing and syntax characteristics of the respective platform. The search was confined to the English-language publications from January 2010 to January 2026.

In WoS, the Advanced Search (TS) results were refined to include only article and review document types, excluding proceedings papers, yielding a total of 189 records. A MeSH and Title/Abstract hybrid strategy, restricted to journal articles and reviews, was utilized in PubMed, yielding 173 records. Due to the broad nature of Google Scholar and its lack of reproducibility, the scope of the results was constrained to the period from 2010 to 2026, with a sorting criterion based on relevance. The screening process was further restricted to the initial 200 records from a total of 1,460 hits. A separate title-only (intitle): Google Scholar validation search identified 89 records, thereby confirming the scope of coverage. The use of SciSpace's full-text search and deep review was instrumental in expanding coverage and identifying existing lacunae. The records obtained from Google Scholar and SciSpace were included in the study only if they met the established eligibility criteria and were non-duplicate relative to the records identified through the primary databases. A comprehensive account of the search process is available in the supplementary appendix, specifically Appendix 1.

Eligibility Criteria

Eligible studies were peer-reviewed journal articles or reviews involving medical students, neurosurgery residents/fellows, or practicing neurosurgeons that evaluated AI/ML applications with an explicit educational purpose in neurosurgery. Eligible domains included technical skills training, simulation-based education, VR/AR, and objective technical skill assessment.

Studies were excluded if they addressed clinical/diagnostic AI without an educational component, focused on non-neurosurgical specialties, did not report relevant educational or performance outcomes, or were editorials, conference abstracts, commentaries, or other non-peer-reviewed publications.

Study Selection

All records were imported into a reference management system and deduplicated before screening. Two reviewers independently screened titles/abstracts, followed by full-text assessment of potentially eligible reports. Discrepancies were resolved by consensus. When multiple records referred to the same underlying study (e.g., a preprint and a journal version), the records were merged, and the most complete peer-reviewed version was retained. The selection process and the reasons for full-text exclusion are summarized in Figure 1.

Data Extraction and Synthesis

Two reviewers independently extracted data using a standardized form capturing study design, participant characteristics, educational domain, AI/ML approach, comparator (if applicable), and reported educational or performance outcomes. Owing to heterogeneity in designs, interventions, and outcomes, findings were synthesized narratively.

Risk of Bias (RoB) and Quality Appraisal

RoB and methodological quality were assessed using design-specific tools. These assessments were performed to contextualize and interpret the body of evidence and were not used to exclude studies.

Accordingly, RCTs were assessed using the Cochrane RoB 2 tool, non-randomized and quasi-experimental studies were evaluated using risk of bias in non-randomized studies of interventions, and predictive model studies were assessed using the prediction model risk of bias assessment tool (PROBAST).

Results

Study Selection

Searches across the electronic sources identified 1,080 records. After removal of 940 duplicate records, 140 records remained for title and abstract screening, of which 79 were excluded for failing to meet the predefined eligibility criteria. Sixty-one reports were sought for full-text retrieval; all were successfully retrieved and assessed for eligibility. Following full-text evaluation, 21 reports were excluded with documented reasons. Of the remaining 40 reports, 5 referred to duplicate publications of the same underlying study and were merged, yielding 35 unique studies included in the qualitative synthesis. The study selection process is summarized in Figure 1.

Study Characteristics

The 35 studies included in this review were published between 2017 and 2025, with a clear inflection point after 2020, reflecting accelerating scholarly interest in AI-enabled neurosurgical education. Geographically, the evidence base was dominated by studies conducted in North America (51.4%), followed by Europe (22.9%), with additional contributions from Asia and multi-continental collaborative networks, underscoring the field's global expansion.

Methodologically, the included literature was diverse. RCTs accounted for 5 studies (14.3%), while quasi-experimental designs comprised 7 studies (20.0%). Observational cohort studies accounted for the most significant proportion (n=10, 28.6%), followed by cross-sectional studies (n=6, 17.1%). Evidence synthesis studies, including systematic and scoping reviews, accounted for 4 studies (11.4%), and pilot studies or case series accounted for the remaining 3 (8.6%).

Across primary empirical studies, sample sizes ranged from 14 to 156 participants, encompassing medical students, neurosurgery residents and fellows, and practicing neurosurgeons. Collectively, these characteristics highlight both the methodological heterogeneity and the evolving maturity of the literature on AI-supported educational interventions in neurosurgery.

Educational Focus Areas

Majority of the included studies focused on technical skills training (n=25, 71.4%), indicating the predominant use of AI and ML methods for performance measurement, automated feedback, and procedural proficiency in neurosurgical training. A smaller yet noteworthy proportion of studies focused on surgical planning training (n=10, 28.6%), predominantly using AI-assisted simulation environments and decision support interfaces.

A thorough review of the extant literature indicates that clinical reasoning has not been explored as a standalone educational domain. Conversely, reasoning-related constructions were implicitly incorporated within surgical planning tools or hybrid training frameworks. In these frameworks, cognitive processes were assessed indirectly alongside technical or procedural outcomes, rather than as independent learning endpoints.

AI and ML Technologies

Conventional supervised ML classifiers, including support vector machines, k-nearest neighbors, naive Bayes, and random forests, were predominantly utilized for automated performance assessment (n=25, 71.4%). The implementation of deep learning architecture (n=15, 42.9%) facilitated continuous performance monitoring and video-based analysis. Computer vision systems were employed in 22.9% of the studies. A review of the extant literature reveals that fewer than one-third of the studies reported external validation

Technical Skills Training

AI-based interventions for technical skills training demonstrated the strongest and most consistent evidence of efficacy. Automated assessment systems have shown the ability to accurately differentiate expertise levels across simulated tasks, with reported accuracies ranging from 83% to 100%.[2,6–9,11,16,30] Siyar et al.[2] achieved 91.7% accuracy using Fuzzy K-Nearest Neighbors. Winkler-Schwartz et al.[7] reported a 90% accuracy rate in distinguishing between four levels of expertise. Li et al.[9] achieved an accuracy of 92.41% with an area under the curve of 0.98. Karlik et al.[11] demonstrated a 100% classification accuracy using a hybrid fuzzy clustering neural network (Table 1).

Table 1. Distribution of AI/ML technologies in included studies (n=35)

AI/ML technology	n	%
Machine learning classifiers	25	69.4
Support vector machines	12	33.3
K-nearest neighbors	10	27.8
Naive bayes	8	22.2
Decision trees/Random forests	7	19.4
Deep learning/Neural networks	15	41.7
Convolutional neural networks	8	22.2
LSTM/Recurrent networks	5	13.9
Fully connected networks	6	16.7
Computer vision	8	22.2
Hybrid/Ensemble methods	6	16.7
Large language models	2	5.6

AI: Artificial intelligence; ML: Machine learning; LSTM: Long short-term memory.

A substantial body of research has emerged from RCTs, which have reported performance improvements comparable to those observed with traditional expert instruction through the implementation of AI-augmented intelligent tutoring systems.[4,8,10,13] As demonstrated by Fazlollahi et al.,[4] the virtual operative assistant has been shown to improve expertise scores by 0.66 points (95% confidence interval 0.55–0.77, $p<0.001$). Giglio et al.[8] demonstrated that AI-augmented instruction resulted in higher performance scores (mean difference 0.20, $p=0.02$). In a related study, Yilmaz et al.[13] showed that AI instruction led to a significant improvement in composite scores ($p=0.017$, $p=0.005$). In contrast, the administration of human instruction resulted in a decline in performance ($p=0.004$). As illustrated in Table 2, the study's primary outcomes are summarized.

Surgical Planning and Clinical Reasoning

In a systematic/scoping review of AI applications in surgical planning, Senders et al.[20] reported a median improvement in accuracy of approximately 13% across included studies (as recalculated in the present review). In their systematic review of AI applications in surgical planning, Senders et al.[20] reported accuracy improvements; recalculation of their data in the present review yielded a median improvement of approximately 13%.

From an educational perspective, the most significant benefits were observed when trainees engaged in reflective practice with AI outputs, utilizing them to compare algorithmic predictions with expert reasoning.[3,5,28] However, the findings of these studies were predominantly short-term and simulation-based, with no study assessing clinical reasoning as an independent educational outcome. Consequently, the evidence for sustained improvement in higher-order cognitive skills remains inconclusive. To integrate the identified AI technologies, academic domains, and outcome dimensions across technical skills training, clinical reasoning, and surgical planning, a conceptual framework was developed to summarize AI's roles in neurosurgical education (Fig. 2).

The framework synthesizes evidence from the included studies, illustrating how AI technologies are applied across technical skills training, clinical reasoning, and surgical planning, and how these applications relate to assessment, feedback, and decision-support outcomes in neurosurgical education.

Quality Assessment

Among the studies that employed PROBAST to assess prediction models (n=31), the most prevalent concerns were identified in the analysis domain (18/31, 58.1%) and the outcome domain (23/31, 74.2%), while participant selection was less frequently highlighted (11/31, 35.5%). The limitations of the studies included small sample sizes, single-center designs, reliance on simulated outcomes, and limited external validation. In PROBAST-based prediction models, the most common sources of bias were small training datasets (58.1%), lack of external validation (74.2%), and potential overfitting (35.5%). Despite the limitations, the uniformity of the findings lends credibility to the conclusions drawn.

Discussion

This systematic review synthesizes evidence from 35 studies examining AI/ML applications with an explicit educational purpose in neurosurgery. A thorough review of the extant literature reveals an uneven evidence base across educational domains. The most robust and consistent support is for simulation-based technical skills training and objective performance assessment.[3,7,14,15,17] AI-enabled systems have demonstrated consistent proficiency in discriminating expertise levels and providing precise, replicable feedback across diverse platforms and tasks. This capability substantiates their function as scalable measurement instruments within competency-based training frameworks,[8,14,15,17] aligning with Miller's pyramid of clinical competence, which positions technical performance ("shows how") as a foundational prerequisite for autonomous clinical practice.[40] The observed accuracy

Table 2. Summary of key studies: AI technologies, applications, and outcomes

Study	AI technology	Educational application	Sample	Key outcomes
Siyar et al.[2]	Fuzzy KNN, SVM	VR tumor resection assessment	115	91.7% accuracy, 8.3% EER
Winkler-Schwartz et al.[7]	KNN, Naive Bayes, SVM	Expertise classification in VR	50	90% accuracy with KNN
Yilmaz et al.[6]	Deep neural network, LSTM	Continuous bimanual monitoring	50	4 expertise levels ($p<0.001$), $R^2=27.7\%$
Fazlollahi et al.[4]	Deep learning (ICEMS)	AI tutoring versus expert (RCT)	70	VOA improved 0.66 pts ($p<0.001$)
Giglio et al.[8]	AI tutoring system	AI-augmented instruction (RCT)	87	Mean difference 0.20 ($p=0.02$)
Yilmaz et al.[13]	AI intelligent instruction	Real-time AI versus in-person (RCT)	25	AI improved ($p=0.017$); human decreased
Li et al.[9]	SVM with shapley values	Personalized VR assessment	79	92.41% accuracy, AUC=0.98
Karlik et al.[11]	Fuzzy clustering NN	Expertise classification in VR	NS	100% accuracy with FCNN
Singh et al.[15]	Naive Bayes, SVM, DT	Craniotomy drilling assessment	22	90.0% accuracy
Ledwos et al.[16]	KNN for metric selection	Learning curve analysis	50	5/6 metrics significant ($p<0.05$)
Senders et al.[20]	Various ML models	Surgical planning (SR)	23 studies	13% median accuracy improvement
Sugiyama et al.[12]	Semantic segmentation	Microsurgical video analysis	14	Strong correlation with ratings
Witten et al.[32]	ML segmentation	Neuroanatomical classification	NS	91.8% test accuracy
Reich et al.[5]	Artificial neural networks	Competency-based VR training	NS	Objective tracking of skill progression
Mirchi et al.[29]	Explainable AI	Simulation-based training (VOA)	NS	Transparent and personalized feedback
Pangal et al.[30]	Video-based ML	Automated performance metrics	NS	Automated expertise differentiation
Bocanegra-Becerra[18]	Various ML algorithms	Pre-operative planning (SR)	NS	Integration of imaging/clinical data

NS: Not specified; EER: Equal error rate; SR: Systematic/Scoping review; DT: Decision tree; NN: Neural network; VR: Virtual reality; RCT: Randomized controlled trial; FCNN: Fuzzy clustering neural network; VOA: Virtual operative assistant; AI: Artificial intelligence; SVM: Support vector machines; KNN: K-nearest neighbors; ICEMS: Intelligent continuous expertise monitoring system; AUC: Area under the curve.

of AI systems in expertise classification (83–100%) suggests their potential utility in systematically documenting learner progression through the stages of skill acquisition described in the Dreyfus model—from novice reliance on explicit rules to expert intuitive performance.[41]

A substantial body of research, supported by empirical evidence from randomized and comparative studies, suggests that AI-augmented instruction, particularly when integrated within structured simulation curricula, can enhance technical performance and learning efficiency.[4,5] The merits of AI in education are most clearly delineated when it functions as an augmentative mechanism, a standardizing agent of feedback, and an enabler of deliberate practice as conceptualized by Ericsson.[42] Deliberate practice, characterized by focused repetition with immediate corrective feedback, represents the cornerstone of expert performance development. AI systems that provide granular, real-time performance metrics and individualized coaching operationalize this framework at scale, enabling trainees to engage in high-repetition practice with consistent feedback quality independent of faculty availability. While virtual and MR platforms provide immersive training environments,[38] AI integration adds objective, quantitative assessment capabilities that VR alone cannot deliver, exemplifying how complementary technologies address different dimensions of surgical competence. In such cases, AI serves to complement, rather than supplant, expert supervision. This approach preserves the educational value of mentorship while concomitantly reducing variability in assessment.[3–5,7]

Conversely, applications designed for surgical planning education were less numerous and more heterogeneous. Despite the demonstrated potential advantages of AI-assisted planning tools when used in conjunction with human judgment, educational outcomes were frequently inferred indirectly rather than measured longitudinally.[1,13] It is important to note that clinical reasoning was rarely operationalized as a standalone educational endpoint. Instead, reasoning-related elements were embedded within hybrid technical–cognitive or planning frameworks. This finding underscores the intricacies of measuring higher-order cognition and the limitations of current methodological approaches.[1,13] Within Bloom's taxonomy of educational objectives,[43] most AI applications in this review address lower cognitive domains (knowledge, comprehension, application) rather than higher-order processes such as analysis, synthesis, and evaluation—the very competencies that distinguish expert clinical reasoning. The predominant focus on

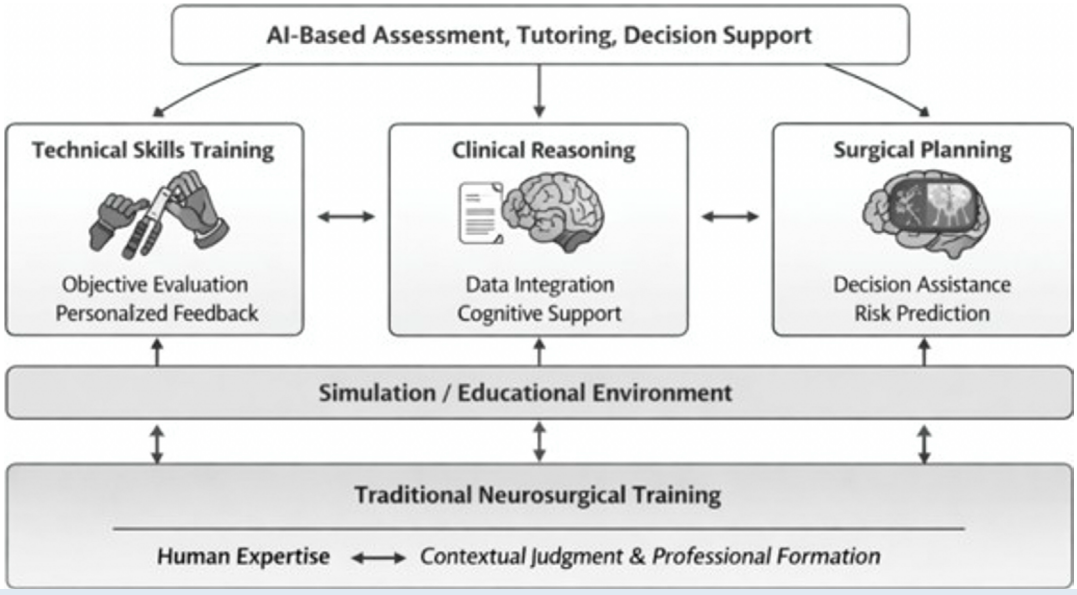


Figure 2. Conceptual framework of artificial intelligence applications in neurosurgical education.

psychomotor skill assessment reflects a natural alignment between simulation-generated kinematic data and ML algorithms, yet leaves unaddressed the critical cognitive skills that enable surgeons to adapt procedural knowledge to novel clinical contexts, anticipate complications, and make sound judgments under uncertainty.

This "cognitive lacuna" in the literature underscores a salient risk of automation bias, in which trainees may become excessively reliant on algorithmic outputs, potentially at the expense of developing their own clinical intuition. As the field of AI matures, it is imperative to transition from using AI as a passive scoring tool to a "cognitive partner." This necessitates the incorporation of explainable AI (XAI) frameworks[29] that do not merely provide a performance score but also deconstruct the underlying cognitive architecture of surgical decision-making. Moreover, while AI systems demonstrate accuracy ranging from 83% to 100%, the absence of external validation in 74.2% of studies[15,31] constitutes a significant impediment to clinical translation. Future research must prioritize multi-institutional validation and the tracking of skill retention and transfer from simulated environments to the operating room. The actual value of AI in neurosurgery will not lie in its capacity to supplant the mentor; instead, it will be found in its ability to deconstruct the "black box" of surgical expertise, thereby providing a transparent, equitable, and evidence-based foundation for the next generation of mastery.[32,35]

In the context of surgical planning education, pedagogical risks merit explicit attention. When trainees interact with high-performing planning or prediction tools, they may develop overconfidence or a false sense of security, defer prematurely to algorithmic recommendations, or reduce reflective comparison with expert reasoning, especially when model uncertainty and failure modes are not transparent. Such effects can inadvertently weaken deliberate practice in clinical judgment and promote "deskilling" in scenario appraisal. Mitigation strategies include supervised use with structured debriefing, calibrated trust training (requiring trainees to justify decisions independent of the model), presentation of confidence/uncertainty estimates, and adoption of explainable AI interfaces that highlight rationale and limitations rather than only outputs.

From a theoretical standpoint, the prevalence of technical skills studies in this field indicates the inherent congruence between simulation-generated data streams and ML methodologies. High-frequency kinematic and performance data are compatible with supervised learning approaches, thereby facilitating model development and validation.[8,14,15] Conversely, reasoning and judgment are context-dependent and less directly observable. This observation underscores the need for greater integration of educational theory, cognitive science, and explainable AI to ensure that AI outputs support learning rather than automation bias.[13]

This asymmetry in AI application, robust support for technical skills, but limited engagement with clinical reasoning, carries important educational implications. Competency in neurosurgery requires not only procedural fluency but also the cultivation of what Schön termed "reflection-in-action":[44] The capacity to think critically during performance, recognize anomalies, and adaptively modify one's approach. Current

AI systems excel at evaluating what trainees do (observable behaviors, tool paths, efficiency metrics) but cannot yet adequately assess how they think or why they make specific decisions. This limitation becomes particularly salient when considering the progression from novice to expert as described by the dreyfus model,[41] where advanced practitioners move beyond rule-following to develop holistic, context-sensitive judgment. Future AI applications must therefore evolve beyond passive performance measurement to actively scaffold metacognitive processes, prompting trainees to articulate their reasoning, compare their decision-making with expert approaches, and develop the reflective habits that characterize surgical expertise.

Several practical considerations must be considered. AI-enabled simulation has the potential to enhance practice density and facilitate individualized learning trajectories, while operating within constraints such as reduced operative exposure and limited faculty time.[3,7] However, the comprehensive implementation of the curriculum is contingent upon substantiating its correlation with clinical performance, skill retention, and program-level impact.[3,7,8,13] Equity considerations underscore the importance of multi-institutional validation and transparent reporting to mitigate algorithmic bias and site-specific overfitting.[3,13,17]

A review of the extant literature reveals several limitations. Firstly, the samples are often small, and studies are predominantly conducted at a single center. Furthermore, there is considerable methodological heterogeneity. These factors prevented conducting a quantitative synthesis. The application of external validation and standardized reporting remains constrained, thereby limiting the generalizability of findings.[3,13,17] At the review level, the available literature that has undergone peer review, as well as the variability across supplementary search platforms, may have influenced the extent of coverage.[3]

Conclusion

A thorough review of the extant literature indicates that AI and ML applications have demonstrated high precision and reproducibility in simulation-based technical skills assessment. These technologies have proven effective as objective measurement instruments, offering standardized evaluation and scalable learning support that transcends the limitations of traditional apprenticeship models. The most significant instructional gains are derived from hybrid training frameworks where AI-augmented systems complement expert supervision by standardizing feedback and extending deliberate practice, rather than replacing the essential role of human mentorship.[28,30]

Conversely, while AI-assisted tools demonstrate potential in surgical planning, their application in cultivating and assessing clinical reasoning remains nascent and inconsistently operationalized. The assessment of learning outcomes in these higher-order cognitive domains is often based on inferences rather than on longitudinal, evidence-based metrics. The development of a cognitive foundation necessitates a transformation in educational paradigms, encompassing the integration of transparent, XAI frameworks. These frameworks should prioritize critical judgment over automation bias, thereby facilitating the advancement of knowledge and skills in a systematic and well-informed manner.[29]

In addition, the incorporation of ML algorithms for automated segmentation of operative neuroanatomy has demonstrated considerable potential to enhance the precision of preoperative planning. For instance, Witten et al.[32] demonstrated that AI-driven image segmentation can effectively delineate complex neuroanatomical structures, thereby providing trainees with a high-fidelity visual roadmap that facilitates better spatial orientation during the transition from simulation to the operating theater. The ultimate objective for neurosurgical training is to ensure that these innovations translate into improvements that are both generalizable and evidence-based.[17]

In contemplating the imminent integration of AI-driven innovations within clinical practice, it is imperative to address the prevailing methodological limitations. Future research must prioritize multicenter validation, tracking skill retention, and objective measurement of technical transfer to the operating room.[15,31] Adherence to consistent reporting standards and a focus on equitable, data-driven advancements are pivotal for the potential of AI to fundamentally redefine neurosurgical mastery and ensure superior patient-safety outcomes in an increasingly complex operative landscape.[32,35]

Ethics Committee Approval: Ethics committee approval was not required for this study because it is a systematic review and does not involve direct human or animal participants, patient intervention, or access to identifiable personal data.

Conflict of Interest: None declared.

Financial Disclosure: The authors declared that this study received no financial support.

Use of AI for Writing Assistance: The authors would like to acknowledge the use of AI-assisted writing tools for language editing, structural refinement, and reference formatting. All intellectual content, including the study rationale, analysis, interpretation, and conclusions, is the original work of the authors, who take full responsibility for the manuscript.

Authorship Contributions: Concept: ÖSŞ; Design: ÖSŞ; Supervision: ÖSŞ, SD; Resource: ÖSŞ; Materials: ÖSŞ; Data collection and/or processing: ÖSŞ, SD; Analysis and/or interpretation: ÖSŞ, SD; Literature review: ÖSŞ; Writing: ÖSŞ, SD; Critical review: ÖSŞ.

Peer-review: Double blind peer-reviewed.

References

1. Wu J, Liang X, Bai XF, Chen Z. SurgBox: agent-driven operating room sandbox with surgery copilot. IEEE Int Conf Big Data 2024:1449-58. [CrossRef]
2. Siyar S, Azarnoush H, Rashidi S, Winkler-Schwartz A, Bissonnette V, Ponnudurai N, et al. Machine learning distinguishes neurosurgical skill levels in a virtual reality tumor resection task. Med Biol Eng Comput 2020;58(6):1357-67. [CrossRef]
3. Yilmaz R, Winkler-Schwartz A, Mirchi N, Reich A, Ledwos N, Del Maestro RF. Development and predictive validation of an intelligent, continuous assessment system for bimanual surgical skills. Can J Neurol Sci 2022;49(6):1-10. [CrossRef]
4. Fazlollahi AM, Bakhaidar M, Alsayegh A, Yilmaz R, Winkler-Schwartz A, Mirchi N, et al. Effect of artificial intelligence tutoring vs expert instruction on learning simulated surgical skills among medical students: a randomized clinical trial. JAMA Netw Open 2022;5(2):e2149008. [CrossRef]
5. Reich A, Mirchi N, Yilmaz R, Ledwos N, Bissonnette V, Tran DH, et al. Artificial neural network approach to competency-based training using a virtual reality neurosurgical simulation. Oper Neurosurg 2022;23(1):31-9. [CrossRef]
6. Yilmaz R, Winkler-Schwartz A, Mirchi N, Reich A, Christie S, Tran DH, et al. Continuous monitoring of surgical bimanual expertise using deep neural networks in virtual reality simulation. NPJ Digit Med 2022;5(1):54. [CrossRef]
7. Winkler-Schwartz A, Yilmaz R, Mirchi N, Bissonnette V, Ledwos N, Siyar S, et al. Machine learning identification of surgical and operative factors associated with surgical expertise in virtual reality simulation. JAMA Netw Open 2019;2(8):e198363. [CrossRef]
8. Giglio B, Albeloushi A, Alhaj A, Alhantoobi M, Saeedi R, Davidovic V, et al. Artificial intelligence-augmented human instruction and surgical simulation performance. JAMA Surg 2025;160(4):1-8. [CrossRef]
9. Li F, Qin Z, Qian K, Liang S, Li C, Tai Y. Personalized assessment and training of neurosurgical skills in virtual reality: an interpretable machine learning approach. Virtual Reality Intell Hardware 2024;6(1):27-43. [CrossRef]
10. Piñera-Castro HJ, Borges-García C. Applications of artificial intelligence in neurosurgical education: a scoping review. Egypt J Neurol Psychiatr Neurosurg 2025;61(1):15. [CrossRef]
11. Karlik B, Yilmaz R, Winkler-Schwartz A, Mirchi N, Bissonnette V, Ledwos N, et al. Assessment of surgical expertise in virtual reality simulation by hybrid deep neural network algorithms. Int J Artif Intell Expert Syst 2021;10(3):47-59.
12. Sugiyama T, Tang M, Sugimori H, Sakamoto M, Fujimura M. Artificial intelligence-integrated video analysis of vessel area changes and instrument motion for microsurgical skill assessment. Sci Rep 2025;15(1):27898. [CrossRef]
13. Yilmaz R, Bakhaidar M, Alsayegh A, Abou Hamdan N, Fazlollahi AM, Tee T, et al. Real-time multifaceted artificial intelligence vs in-person instruction in teaching surgical technical skills: a randomized controlled trial. Sci Rep 2024;14(1):15130. [CrossRef]
14. Titov O, Bykanov A, Pitskhelauri D. Neurosurgical skills analysis by machine learning models: systematic review. Neurosurg Rev 2023;46(1):121. [CrossRef]
15. Singh R, Godiyal AK, Suri A. Craniotomy simulator with force myography and machine learning-based skills assessment. Bioengineering 2023;10(4):465. [CrossRef]
16. Ledwos N, Mirchi N, Yilmaz R, Winkler-Schwartz A, Sawni A, Fazlollahi AM, et al. Assessment of learning curves on a simulated neurosurgical task using metrics selected by artificial intelligence. J Neurosurg 2022;137(4):1160-71. [CrossRef]
17. Davids J, Manivannan S, Darzi A, Giannarou S, Ashrafian H, Marcus HJ. Simulation for skills training in neurosurgery: a systematic review, meta-analysis, and analysis of progressive scholarly acceptance. Neurosurg Rev 2021;44(4):1853-67. [CrossRef]
18. Bocanegra-Becerra JE, Neves Ferreira JS, Simoni G, Hong A, Rios-Garcia W, Eraghi MM, et al. Machine learning algorithms for neurosurgical preoperative planning: a scoping review. World Neurosurg 2025;194:123465. [CrossRef]
19. Danushka N, Wijesinghe D, Jayasinghe R, Attanayake D. AI-powered precision: transforming neurosurgical practice through intelligent technologies. In: Artificial Intelligence. London (UK): IntechOpen; 2025. [CrossRef]
20. Senders JT, Arnaout O, Karhade AV, Dasenbrock HH, Gormley WB, Broekman ML, et al. Natural and artificial intelligence in neurosurgery: a systematic review. Neurosurgery 2018;83(2):181-92. [CrossRef]
21. Dundar TT, Yurtsever I, Pehlivanoglu MK, Yildiz U, Eker A, Demir MA, et al. Machine learning-based surgical planning for neurosurgery: artificial intelligent approaches to the cranium. Front Surg 2022;9:863633. [CrossRef]
22. Khizar A. Artificial intelligence and neurosurgery: a revolution in the field. Pak J Neurol Sci 2024;18(4):244. [CrossRef]
23. Mofatteh M. Neurosurgery and artificial intelligence. AIMS Neurosci 2021;8(4):477-95. [CrossRef]
24. Awuah WA, Adebusoye FT, Wellington J, David L, Salam A, Weng Yee AL, et al. Recent outcomes and challenges of artificial intelligence, machine learning, and deep learning in neurosurgery. World Neurosurg X 2024;23:100301. [CrossRef]
25. Singh C, Gharde P, Verma P, et al. Artificial intelligence in neurosurgery: enhancing diagnosis, treatment and patient outcomes: a narrative review. J Clin Diagn Res 2025;19(9):PE01-PE05. [CrossRef]
26. Yin S, Ming J, Chen H, Sun Y, Jiang C. Integrating deep learning and radiomics for preoperative glioma grading using multi-center MRI data. Sci Rep 2025;15(1):36756. [CrossRef]

27. Maghrabi Y, Jamjoom AB, Algahtani A, Alshareef OH, Jamjoom OM, Alzahrani M. Highly cited artificial intelligence research studies published in neurosurgical journals: a bibliometric analysis. Cureus 2025;17(11):e98191. [CrossRef]
28. Sugiyama T, Sugimori H, Tang M, Fujimura M. Artificial intelligence for patient safety and surgical education in neurosurgery. JMA J 2025;8(1):76-85. [CrossRef]
29. Mirchi N, Bissonnette V, Yilmaz R, Ledwos N, Winkler-Schwartz A, Del Maestro RF. The virtual operative assistant: an explainable artificial intelligence tool for simulation-based training in surgery and medicine. PLoS One 2020;15(2):e0229596. [CrossRef]
30. Pangal DJ, Kugener G, Cardinal T, Lechtholz-Zey E, Collet C, Lasky S, et al. Use of surgical video-based automated performance metrics to predict blood loss and success of simulated vascular injury control in neurosurgery: a pilot study. J Neurosurg 2021;137(3):840-9. [CrossRef]
31. Escobar-Castillejos D, Barrera-Animas AY, Noguez J, Magana AJ, Benes B. Transforming surgical training with AI techniques for training, assessment, and evaluation: scoping review. J Med Internet Res 2025;27:e58966. [CrossRef]
32. Witten AJ, Patel NB, Cohen-Gadol AA. Image segmentation of operative neuroanatomy using machine learning. Oper Neurosurg 2022;23(4):e322-8. [CrossRef]
33. Konakondla S, Fong R, Schirmer CM. Simulation training in neurosurgery: advances in education and practice. Adv Med Educ Pract 2017;8:465-73. [CrossRef]
34. Harley JM, Tawakol T, Azher S, Quaiattini A, Maestro RD. The role of AI in neurosurgical education: an umbrella review. Glob Surg Educ 2024;3(1):83. [CrossRef]
35. Tariciotti L, Palmisciano P, Giordano M, Remoli G, Lacorte E, Bertani G, et al. Artificial intelligence-enhanced intraoperative neurosurgical workflow: current knowledge and future perspectives. J Neurosurg Sci 2022;66:139-50. [CrossRef]
36. Shao X, Yuan Q, Qian D, Ye Z, Chen G, Zhuang K, et al. Virtual reality technology for teaching neurosurgery of skull base tumor. BMC Med Educ 2020;20(1):3. [CrossRef]
37. Jain S, Timofeev I, Kirollos RW, Helmy A. Use of mixed reality in neurosurgery training: a single centre experience. World Neurosurg 2023;176:e68-76. [CrossRef]
38. Silvero Isidre A, Friederichs H, Müther M, Gallus M, Stummer W, Holling M. Mixed reality as a teaching tool for medical students in neurosurgery. Medicina (Kaunas) 2023;59(10):1720. [CrossRef]
39. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 2021;372:n71. [CrossRef]
40. Miller GE. The assessment of clinical skills/competence/performance. Acad Med 1990;65(9 Suppl):S63-7. [CrossRef]
41. Dreyfus SE, Dreyfus HL. A five-stage model of the mental activities involved in directed skill acquisition. Berkeley: University of California, Operations Research Center; 1980. [CrossRef]
42. Ericsson KA. Deliberate practice and acquisition of expert performance: a general overview. Acad Emerg Med 2008;15(11):988-94. [CrossRef]
43. Bloom BS, Engelhart MD, Furst EJ, Hill WH, Krathwohl DR. Taxonomy of educational objectives: the classification of educational goals. Handbook I: cognitive domain. New York: David McKay Company; 1956.
44. Schön DA. The reflective practitioner: how professionals think in action. New York: Basic Books; 1983.

Appendix 1. Database-specific search strategies and retrieval details

This appendix reports the full search strategies, platform-specific adaptations, and retrieval decisions applied in this systematic review, ensuring transparency and reproducibility in accordance with PRISMA 2020 guidelines.

Table A1. Harmonized search strategy across databases

Database / platform	Search strategy (final, high-specificity)	Search date	Platform mode	Filters / restrictions applied	Hits retrieved (n)	Role in review
PubMed (MEDLINE)	MeSH + Title/Abstract hybrid query requiring AI/ML, neurosurgery, education/training, and simulation/VR or objective skill assessment.	January 2026	Advanced Search	English; Publication date 2010–Jan 2026; Journal Article OR Review	173	Primary database for systematic identification
Web of Science Core Collection	TS-based query requiring AI/ML, neurosurgery, education/training, and simulation/VR or skill assessment.	January 2026	Advanced Search (TS)	Article (and Review); Proceedings Papers excluded; English; 2010–2026	189	Primary database for systematic identification
Google Scholar	Keyword-based Boolean query aligned with the four-domain structure (AI/ML + neurosurgery + education + simulation/skills).	January 2026	Relevance-sorted search	Custom year range 2010–2026; first 200 results screened; peer-reviewed journal articles/reviews retained	1460	Supplementary source to enhance sensitivity
Google Scholar (title-only validation search)	intitle:(simulation OR "virtual reality" OR "skill assessment") AND (neurosurgery OR neurosurgical OR "spine surgery") AND ("artificial intelligence" OR "machine learning" OR "deep learning")	January 2026	intitle: operator	English; 2010–2026	89	Validation and specificity check
SciSpace – Full-Text Search	Full-text keyword query using the same four-domain structure (AI/ML, neurosurgery, education, simulation/skills).	January 2026	Full-Text Search	Year ≥2010; English where available; peer-reviewed focus during screening	100	Supplementary source for coverage enhancement
SciSpace – Deep Review	Structured AI-assisted query restricted to peer-reviewed AI/ML educational applications in neurosurgery involving simulation/VR or skill assessment.	January 2026	Deep Review (AI-assisted)	English; 2010–Jan 2026; peer-reviewed focus	529	Curated cross-check and gap detection

PRISMA Compliance and Reporting Notes: Search strategies were prospectively defined and consistently structured across platforms around four core concepts: artificial intelligence/machine learning, neurosurgery, education or training, and simulation-based or objective skill assessment outcomes. Database-specific syntax and filters were applied as required. Google Scholar searches were capped to ensure feasibility and reproducibility, and title-only searches were used solely as validation exercises. Retrieval counts were recorded prior to de-duplication and are summarized in the PRISMA 2020 flow diagram.