Applications of Artificial Intelligence in Neurosurgical Education: A Scoping Review

Hector Julio Piñera-Castro[1*] https://orcid.org/0000-0002-2491-489X
Christian Borges-García[2] https://orcid.org/0000-0002-1430-8706

[1]Department of Neurosurgery, Hospital Clínico Quirúrgico Hermanos Ameijeiras. Havana, Cuba.
[2]Facultad de Ciencias Médicas Dr. Ernesto Guevara de la Serna, Universidad de Ciencias Médicas de Pinar del Río. Pinar del Río, Cuba.

*Corresponding author: hectorpinera18100@gmail.com

ABSTRACT

Background: Artificial intelligence (AI) has revolutionized medical education due to its capacity to optimize instruction, assess competencies, and tailor learning. Given the interdisciplinary nature and transformative potential of AI, its integration into neurosurgical education merits thorough examination.

Objective: To evaluate systematically the applications of AI in neurosurgical education.

Methods: A scoping review was conducted in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses Extension for Scoping Reviews (PRISMA-ScR) guidelines. Original observational and experimental studies published in peer-reviewed journals that investigated the use of AI in neurosurgical education were included. The search was carried out in Scopus up to May 2024, and studies not written in English or Spanish were excluded. The selection process and data extraction were performed independently by two reviewers. A narrative synthesis approach was employed.

Results: Twenty-three studies were selected, identifying four main areas. In board examination performance, advanced models like ChatGPT-4 outperformed previous

versions and residents in answering neurosurgical exam questions, achieving higher accuracy in complex theoretical and image-based inquiries. In simulation training and tutoring, neural networks analyzed surgical simulation data, classifying participants by expertise levels and providing individualized, metric-based feedback that improved technical skills and identified distinct learning patterns among trainees. In performance evaluation, machine learning techniques assessed surgical skills during simulations with high accuracy, identifying key behavioral metrics associated with expertise. Other innovative applications included AI for neuroanatomical segmentation in imaging studies, analysis of surgical instrument utilization patterns through computer vision, and generation of academic content. These studies highlighted AI's multifaceted roles and ethical considerations, such as overreliance risks and the need for robust databases and realistic simulations.

Conclusions: AI has enhanced neurosurgical education by improving knowledge assessment, simulation training, and performance evaluation. Despite its transformative potential, ethical and technical challenges persist. The continuous development of AI models and their responsible implementation are recommended to further optimize neurosurgical training and ensure safe integration into educational curricula.

Keywords: Artificial Intelligence; Neurosurgery; Medical Education; Neurosurgical Procedures; Neurosurgical Education.

INTRODUCTION

Artificial intelligence (AI) is an interdisciplinary domain dedicated to enabling computers to execute tasks traditionally associated with human cognition, such as perception, reasoning, problem-solving, and planning. Central to AI are machine learning and data analytics, which empower machines to emulate human cognitive processes and address complex, ambiguous problems with intentionality, intelligence, and adaptability. AI finds applications across various sectors, notably within the medical field.[1–3]

AI has a significant impact on medical education by transforming educational paradigms and enhancing learning experiences. AI technology ensures the safety and consistency of medical experiments, fosters diagnostic reasoning, and heightens humanistic care awareness among students. It amalgamates resources, aids instruction, simulates clinical environments, monitors teaching quality, facilitates self-evaluation, and enables personalized learning. Additionally, AI equips healthcare professionals with superior skills and knowledge, leading to improved practical competencies and better patient outcomes. The provision of immediate and interactive feedback reinforces knowledge, and overall, AI elevates the quality of medical education, inspiring innovative applications within the field.[4–8]

Given the interdisciplinary nature and transformative potential of AI, its integration into neurosurgical education merits thorough examination. The complexity and precision required in neurosurgical training amplify the potential benefits of AI, which includes enhancing educational paradigms, ensuring the safety and consistency of training methods, and providing adaptive learning experiences.

Therefore, it is pertinent to ask: What are the current applications of AI in neurosurgical education? This review systematically evaluates these applications, consolidating existing knowledge and highlighting best practices. By doing so, it seeks to inspire innovative applications and inform future research, ultimately aiming to heighten the proficiency and preparedness of neurosurgeons through AI-enhanced education.

METHODS

This scoping review adhered to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA),[9] complemented with the PRISMA for Scoping Reviews (PRISMA-ScR)[10] and the PRISMA for Abstracts (PRISMA-A) extensions.[11]

Eligibility criteria

The inclusion criteria comprised observational or experimental original studies that investigated the application of AI in the context of neurosurgical education. These studies were required to be published in peer-reviewed journals and could encompass various methodological approaches, including but not limited to randomized controlled trials, cohort studies, case-control studies, and cross-sectional studies. The exclusion criterion was limited to studies not written in English or Spanish.

Information Sources and Search Strategy

We conducted a literature search in Scopus, last accessed on May 25th, 2024. No date or language restrictions were applied initially. In addition, we reviewed reference lists of relevant articles to identify any further eligible studies.

Search Strategy

The full search strategy involved using a combination of Medical Subject Headings (MeSH) terms related to "neurosurgery" and "artificial intelligence", within the article title, abstract, and keywords: *(Neurosurg* OR "Neurologic* Surgery") AND ("Artificial Intelligence" OR "AI" OR "Computer Reasoning" OR "Machine Intelligence" OR "Computational Intelligence" OR "Computer Vision System*" OR "Computer Knowledge Acquisition" OR "Computer Knowledge Representation*")*. We decided not to include terms related to education due to the extensive range of possibilities. Instead, we preferred to select the articles pertinent to the topic manually. The filter applied limited the results to original research articles.

Selection Process

Two independent reviewers screened the titles and abstracts to identify studies that might be eligible. The full texts of studies deemed potentially relevant were retrieved and independently assessed for eligibility by the same reviewers. Any discrepancies in judgment were resolved through discussion or by consulting a third reviewer.

Data Collection Process

Data extraction was conducted independently by two reviewers using a standardized form. Any discrepancies were resolved through consensus. Attempts were made to contact the authors of the studies for clarifications or additional data as necessary. No automation tools were employed in the data collection process.

Data items

The following data were extracted: publication year, author(s), study design, sample size, objective, AI model(s) utilized, main results, contributions, limitations, and conclusions.

Risk of Bias Assessment and Effect Measures

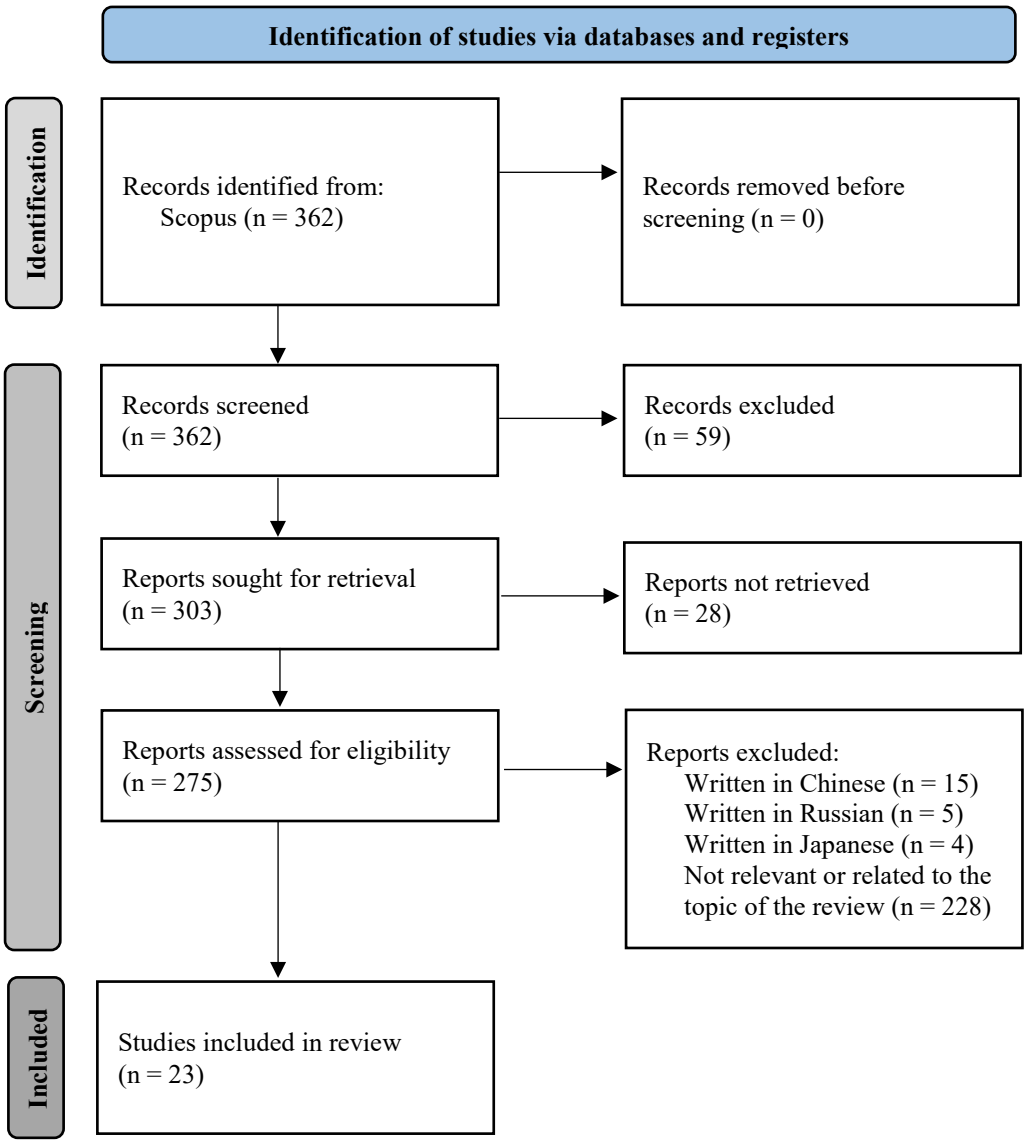
Due to the heterogeneity of the study designs, interventions, and outcomes, effect measures were not utilized and a risk of bias assessment was not performed. The substantial variability in study designs, AI applications, and reported outcomes made it challenging to aggregate the data quantitatively. Additionally, the primary aim of this review was to provide a broad overview of the current landscape and innovative uses of AI in neurosurgical education rather than to quantify the effectiveness of specific interventions.

Data Synthesis

A narrative synthesis approach was employed to collate and summarize the findings across the included studies. This approach allowed for a comprehensive exploration of the various applications of AI in neurosurgical education, highlighting common themes, innovations, and challenges reported in the literature. The narrative synthesis provided a rich contextual understanding of the impact and potential of AI in this field without relying on quantitative effect measures.

RESULTS

The application of the search strategy yielded 362 reports. Following a screening of titles and abstracts, 303 studies were identified as potentially eligible, of which only 275 were successfully retrieved. After conducting an eligibility assessment, 252 reports were excluded for various reasons, resulting in 23 studies being included in the review. The search and study selection process followed the procedure outlined in **Figure 1**.



Source: Elaborated by the authors.
Figure 1 – Flowchart of the search and study selection process.

General overview

This review covers four key areas based on the synthesis of evidence from the selected articles: performance in neurosurgical board examinations and ethical considerations, neurosurgical simulation-based training and tutoring, neurosurgical performance/skills/expertise analysis and assessment, and other applications in neurosurgical education (**Table)**.

Table. Applications and limitations of AI in neurosurgical education.

Key areas	Applications	Limitations
Performance in neurosurgical board examinations and ethical considerations[12–19]	• Answer neurosurgical board examination questions. • Address complex and high-order theoretical questions. • Handle image-based inquiries. • Capability of advanced models like ChatGPT-4 to, in certain cases, surpass the performance of residents, students, and experienced neurosurgeons.	• Inferior performance in practical exams and complex clinical judgment scenarios. • Disparities in accuracy, especially in licensing exams. • Challenges in detecting and processing information from images. • Difficulty in addressing open-ended questions and real-world neurosurgical complexity. • Risk of overreliance on AI, leading to demotivation and loss of fundamental clinical skills.
Neurosurgical simulation-based training and tutoring[20–24]	• Analyze performance metrics during surgical simulations. • Learn the set of metrics for a specific procedure that comprise surgical expertise. • Classify participants by experience level using neural networks and other algorithms. • Provide individualized and quantitative feedback based on metrics obtained in simulations. • Offer real-time audiovisual and metric evaluations. • Assess the impact of various feedback modalities (numeric, visual, visuospatial) on performance improvement. • Develop personalized training plans. • Identify learning patterns.	• Binary classifications are too general and fail to capture nuances in technical competencies. • Difficulties in comprehensively evaluating interdisciplinary skills and team collaboration aspects. • Biases in evaluated metrics, as high values in certain parameters can mask deficiencies in other critical aspects. • Potential negative effects on participants' skills, as AI-enhanced training curriculum were limited to post-exercise comments, potentially lacking important contextual components.
Neurosurgical performance/skills/expertise	• Evaluate competencies and surgical skills during simulated procedures.	• Small datasets hinder model generalization to broader contexts or new scenarios.

analysis and assessment[25–31]	• Use classification algorithms to categorize participant performance. • Analyze specific metrics (e.g., contact time, contact length, inter-instrument distance) to assess technical proficiency. • Identify patterns across different expertise levels. • Identify the surgical task being performed in real-time.	• Simulation scenarios fail to reproduce accurately the biomechanical properties of real tissues, affecting evaluative precision. • Difficulty in holistically integrating all necessary competencies in neurosurgical practice, especially complex behavioral aspects.
Other applications in neurosurgical education[32–34]	• Generate clinical cases and multiple-choice questions for academic purposes. • Draft academic articles. • Implement neuroanatomical segmentation algorithms based on convolutional neural networks for image analysis and structure identification. • Analyze surgical instrument usage patterns using computer vision techniques. • Use measures like Shannon entropy to predict surgical procedure success.	• Generate academic content errors, such as multiple-choice questions with multiple correct answers or inaccurate content. • Limit the accuracy of neuroanatomical segmentation algorithms due to the imprecision of manual reference labels. • Face challenges in analyzing surgical instrument usage patterns, especially in heterogeneous procedures where entropy may vary without reflecting actual success or failure.

Source: Elaborated by the authors.

Performance in Neurosurgical Board Examinations and Ethical Considerations

The review encompassed eight studies[12–19] that evaluated the performance of various AI models in neurosurgical board examinations. These studies compared the results achieved by different AI models with one another, as well as with the performance of medical students and residents preparing for these exams.

In four articles,[12–14,16] a comparative analysis was conducted among various AI models. The findings highlighted the superiority of the ChatGPT-4 model in three of these papers.[12–14]

Ali *et al.*[12] and Guerra *et al.*[14] focused their comparisons on different updates of ChatGPT, revealing that ChatGPT-4 achieved an assertiveness rate of 83.4% on 500 questions[12] and 76.6% on 643 questions,[14] respectively. In contrast, earlier versions like ChatGPT 3.5 demonstrated lower assertiveness, with percentages such as 73.4% on 500

questions[12] and ranging from 53.2% to 60.2% on 643 questions across different models.[14]

These results underscore the significant advancements in the ChatGPT model over time, particularly in its capacity to accurately respond to a broader array of questions. In the research conducted by Ali *et al.*[13], for instance, ChatGPT-4 also exhibited a notably higher performance on image-based questions, with a 68.6% success rate compared to ChatGPT 3.5's 47.1%. Additionally, when incorporating Google Bard model into the comparison, ChatGPT-4 consistently outperformed, with an 82.6% success rate, surpassing both ChatGPT 3.5 and Google Bard, which had success rates of 62.4% and 44.2%, respectively. The high performance of ChatGPT-4 was attributed to its enhanced ability to handle higher-order questions.

Another study[16] indicated that Google Bard outperformed other models, achieving an assertiveness rate of 62% in the written part of the European Board Examination in Neurological Surgery. Although other AIs like ChatGPT and Bing passed the exam, Bard's performance stood out. This paper emphasizes that the highest success rates were observed in theory-based questions (30 out of 86), with ChatGPT at 79%, Bing at 83%, and Bard leading at 86%. These results accentuate the robust text processing capabilities and theoretical answer preparation of these large language models (LLMs).

Seven articles[12,14–19] examined the performance of LLMs in comparison to medical students, residents and neurosurgeons.

On one hand, five of the studies[12,14,16–18] reported that the AI models achieved higher grades than the average grades of the participants examined.

Liu *et al*.[18] devised 50 questions derived from neurosurgical clinical practice guidelines to compare the performance of the GPT-3.5 and GPT-4 models against three groups of neurosurgeons classified by seniority: low, middle, and high. The Chat GPT-4 model demonstrated superior performance, correctly answering 72% of the questions in alignment with neurosurgical practice guidelines. This outperformed the groups with low (40%) and middle (46%) seniority, though it did not surpass the high seniority group, which achieved a 78% success rate. In contrast, the GPT-3.5 model exhibited lower performance at 42%, surpassing only the low seniority neurosurgeons.

Similar behavior is shown by the results presented by Ali *et al.*[12], noting that the ChatGPT models (73.4% for ChatGPT 3.5 and 83.4% for ChatGPT-4) achieved scores comparable to the average user performance (73.7 %), with the average user score being slightly higher than the performance of the ChatGPT 3.5 model and lower than that of the ChatGPT-4 model, emphasizing the superiority of the updated model. However, the assertiveness of the models on the United States Medical Licensing Examination was significantly lower, not exceeding 65%. This disparity might stem from the specialized nature of the Neurosurgical Board Exam questions. Notably, ChatGPT-4 performed best in the fields where students and residents struggled the most (Functional Neurosurgery and Peripheral Nerve Surgery). Another study[16] reiterated that AI models (ChatGPT, Bing, and Bard) showed improved assertiveness in theory-based questions. However, the students enrolled in this study did not exhibit similar improvements, maintaining a 60% success rate.

Sahin *et al.*[17] compared the performance of the ChatGPT model with 260 candidates on 523 questions (600 initial questions minus 77 image-based questions that were excluded) from six exams by the Turkish Society of Neurosurgery Competency Board. The results showed that ChatGPT outperformed the students, with an assertiveness rate of 78.77% compared to the students' average of 62.04%, and even achieving the top grade in three out of the six exams. The study also noted that as the clarity of questions increased, the likelihood of correct answers by the LLM improved—a phenomenon not observed in the students.

Both articles[12,17] emphasize two key points about working with LLMs: their strength in handling purely theoretical elements, which can also be a limitation, and the crucial relationship between the quality and specificity of the prompt and the resulting answer. The value and clarity of the question are proportional to the model's ability to understand the context and provide a better response.

Despite the impressive performance of these models, Guerra *et al.*[14] emphasize that real-world medical practice goes far beyond answering multiple-choice questions. Medical decision-making and patient care involve complex, nuanced responses to open-ended situations. This underlines the necessity for the continuous development and refinement of these technologies, aiming to integrate them effectively into the diverse and intricate tasks of neurosurgical practice.

On the other hand, two studies[15,19] reported that the performance of the LLMs was lower than the average grades of the residents evaluated.

In one study,[15] after having been exposed to the questions for the third time, ChatGPT achieved a correct response rate of 60.2%, which was comparable to the residents' performance (61.5%) but lower than the average Congress of Neurological Surgeons Self-Assessment Neurosurgery user (69.3%). However, ChatGPT still outperformed medical students, who had a success rate of 26.3%.

In the other paper,[19] ChatGPT was ranked ninth out of 11 when compared with 10 residents in a written neurosurgical exam of 50 questions. It is highlighted in the article that to enhance AI capabilities, it is crucial to significantly improve the clarity of the questions.

This phenomenon, observed in the aforementioned articles,[12,17] indicates a relationship between the specificity of the information provided to AI and its ability to generate correct and high-quality responses. This insight can be leveraged to enhance future applications of such models within the field of neurosurgical education.

The ongoing development and integration of LLMs into neurosurgical education indeed bring ethical challenges. Stengel *et al.*[16] highlighted several critical concerns. First, the high performance of these models could lead to overconfidence among professionals, as noted in another study,[12] where ChatGPT omitted a question when tested in the American Board of Neurological Surgery exam, indicating potential blind spots that could have serious clinical and surgical consequences. Additionally, the excessive reliance on AI might demotivate young neurosurgeons at the outset of their learning journey. There is also the risk of cheating during exams, which could distort students' perceived competence and lead to significant challenges when transitioning to real-world clinical and surgical environments. Ensuring these models are used ethically and effectively is crucial for their successful integration into medical education and practice.

Neurosurgical Simulation-Based Training and Tutoring

In fives articles,[20–24] AI models were employed in skills training via surgical simulation technologies.

Two of these articles[20,22] focused on using neural networks to identify performance metrics during simulation exercises (brain tumor resection and anterior cervical

discectomy, respectively) and the subsequent classification of the participants according to their corresponding level of expertise. Classification accuracies of 92%[20] and 83.3%[22] were reported.

These models also provided postoperative feedback, summarizing participants' performances across the metrics. Mirchi *et al.*[20] emphasize the importance of transparent, individualized feedback, noting that experience level classification could be skewed by high metrics compensating for weaker ones. While these quantitative classifications do not encompass the full range of necessary surgical competencies—such as interdisciplinary collaboration—,[22] the integration of these models, combined with objective feedback and instructor input, holds the potential to create a powerful educational tool.[20]

It is important to highlight the significance of multilayer classification for these models, as the binary classification employed by Mirchi *et al.*[20] presents a limitation due to its overly generalist approach (novice or skilled), which may lead to inaccuracies in participant classification based on specific metrics. In contrast, Reich *et al.*[22] adopt a more nuanced classification into three groups (junior residents, senior residents, and post-residents), allowing for a broader spectrum of performance and expertise among participants, thereby enhancing the clarity of the classification.

This ability of artificial neural networks (ANNs) to specifically classify participants' performance enabled Reich *et al.*[22] to describe two distinct learning patterns based on their findings: the continuous pattern, characterized by gradual and linear progress in skill acquisition and refinement over time (e.g., evidenced by a decrease in the application of force on specific anatomical structures proportional to the increase in residents' expertise), and the discontinuous pattern, marked by jumps or nonlinear variations in performance across different experience levels. Instead of demonstrating linear improvement in a particular skill, participants exhibit behaviors characteristic of their experience level. For example, senior residents (more experienced than juniors but less than post-residents) are more likely to make contact with the spinal dura mater when using an active drill compared to the other two groups. This tendency may be attributed to an increase in their confidence, which can manifest as more aggressive movements and a willingness to take greater risks —a phenomenon not observed in post-residents, who, due to their accumulated experience, tend to adopt a more cautious and precise approach.

Conversely, two articles[21,24] investigated how different feedback modalities impact the performance of student groups in surgical simulations. These studies incorporated real-time feedback powered by AI as one of the modalities to which participants were exposed.

On one hand, Fazlollahi *et al.*[21] focused on comparing feedback from remote expert instructors and feedback based on the *Virtual Operative Assistant* (VOA) metrics. As defined by the authors, the VOA is an AI tutoring system "that uses a machine learning algorithm, support vector machine, to classify learner performance and provide goal-oriented, metric-based audiovisual feedback in virtual reality simulations". The research evaluated changes in procedural performance, quantified as an Expertise Score using the validated Intelligent Continuous Expertise Monitoring System (ICEMS; range, −1.00 to 1.00) for each practice resection. Additionally, learning and retention were measured through performance in realistic resections by ICEMS and blinded Objective Structured Assessment of Technical Skills (OSATS; range 1-7). ICEMS assessed 350 practice resections, and ICEMS and OSATS evaluated 70 realistic resections.

In that study,[21] a total of 70 medical students were randomized into three groups. The VOA significantly enhanced practice Expertise Scores by 0.66 points compared to the instructor group and by 0.65 compared to the control group. Realistic Expertise Scores were notably higher for the VOA group compared to both the instructor (mean difference, 0.53) and the control (mean difference, 0.49) groups. The mean global OSATS ratings did not show statistical significance among any of the three groups. These overall OSATS ratings are typically a more holistic assessment that considers multiple aspects of performance, so improvement of certain specific skills from feedback may not be reflected in an overall performance assessment. An important limitation to this comparison was that the expert instruction was conducted remotely and this limited the delivery of product expert feedback to its completion during the pandemic period.

On the other hand, Yilmaz *et al.*[24] assessed four feedback modalities —numerical, visual, visuospatial, and a control group without tailored feedback— in a cohort of 120 medical students using the NeuroVR device with haptic feedback. This device collected data on variables such as applied force, tissue and tumor removal, and bleeding volume.

After completing five repetitions of the simulation exercise and receiving the corresponding personalized feedback, participants in groups 2, 3, and 4 exhibited performance improvements, with scores of 8.40 ± 2.06, 9.10 ± 1.82, and 9.19 ± 1.66 out

of 14, respectively. In contrast, group 1 scored 7.30 ± 1.69, showing no statistically significant improvement over the five repetitions. This finding is further supported by comparing evaluated metrics across the groups, such as the amount of healthy tissue excised or the separation distance of the instrument tip when using two instruments concurrently, with groups 3 and 4 demonstrating better performance from the second to the fifth repetition.[24] These results highlight the critical importance of feedback in developing surgical skills and illustrate how enhancing feedback can significantly advance surgical training.

The integration of AI-based automated systems for personalized feedback enhances learning efficiency and effectiveness, positioning this technology as a highly valuable complementary tool for neurosurgery training. Moreover, as an automated solution, it facilitates large-scale implementations and can be adapted to various expertise levels, thereby easing its integration into advanced training curricula.

Another research[23] evaluated the capacity of an AI model to enhance a training curriculum by comparing two groups of students: one exposed to the AI-enhanced training curriculum and another that did not receive feedback but had expert benchmarks. The first group showed significant improvement in safety-related metrics, such as bimanual control of the operating field (-4.99 maximum instrument divergence) and healthy tissue reduction rate (-7.05×10^5 mm^3). Conversely, negative effects were observed, such as lower speed and acceleration in the dominant hand, and a significant reduction in the rate of tumor removal. These shortcomings may be attributed to the fact that students exposed to the AI-enhanced training curriculum were limited to post-exercise comments, potentially lacking important contextual components, such as instructions to reduce the force or linear acceleration of each attempt, which could result in cumulative behavioral changes. Nevertheless, this presents a potential area for further development of AI and its integration into neurosurgical education, warranting attention in future research projects.

It is possible to infer several clear potentials regarding the utilization of AI models in the training and tutoring of neurosurgical skills. Firstly, the ability to learn and understand the metrics or skills that constitute neurosurgical expertise could enable a classification based on the performance of participants.[20,22] This includes identifying learning patterns that allow educators to create educational tools and strategies tailored to each level of expertise, addressing specific needs previously determined by these AI models.[22] This

approach could serve as an additional evaluative tool in the neurosurgical training process, complementing expert assessments and judgments. It would enable the evaluation of residents in "blind spots" for traditional models, encompassing skills such as ethical decision-making, effective communication with patients and families, and the capacity to collaborate, listen, and contribute within a multidisciplinary team.

Furthermore, an intriguing utility of these models lies in their ability to provide differentiated feedback to medical students and residents, fostering more specialized educational development tailored to their specific needs. This approach not only emphasizes positive behaviors where the resident has achieved expertise but also addresses metrics of negative performance,[20] potentially even surpassing the instructions provided by experts themselves.[21] Additionally, these models could be capable of developing neurosurgical training plans accompanied by the aforementioned feedback.[23] In the future, the integration of AI could be employed to create personalized plans for residents and students, identifying the most critical areas for improvement within their training. This specialized attention would propel them towards reaching their maximum potential within the specialty.

Neurosurgical Performance/Skills/Expertise Analysis and Assessment

Seven studies[25–31] investigated the role of AI models in assessing the performance of neurosurgical students and residents after simulated procedures.

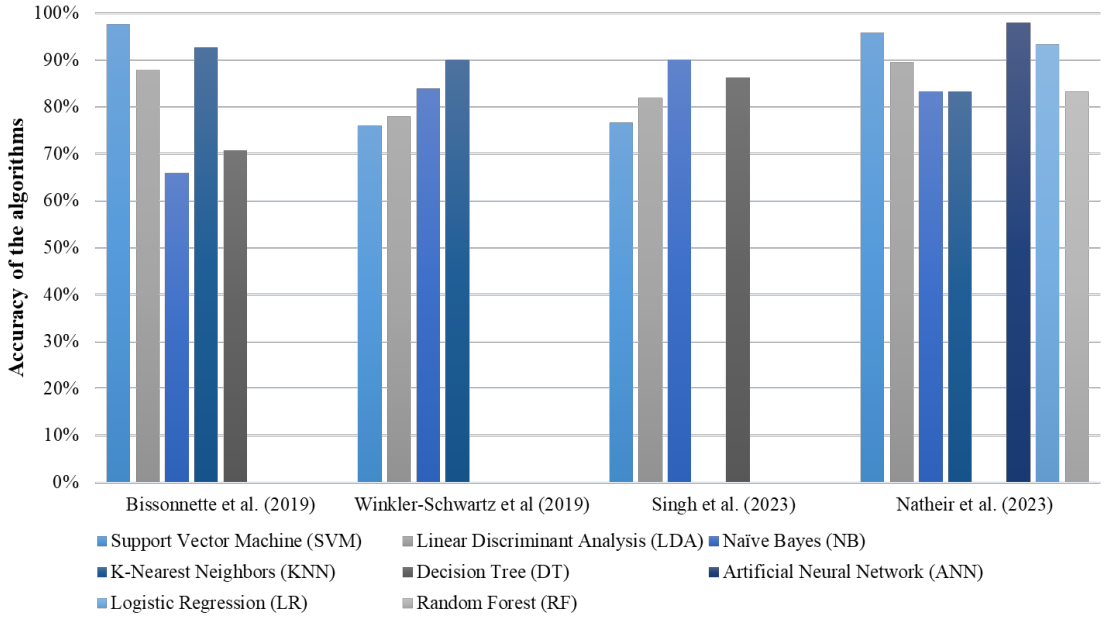
To achieve this, data were collected from participants, classified by expertise level, during various simulation exercises. The aim was to train AI models to identify behavioral trends based on expertise levels. This was accomplished through the visualization of procedures, data from simulation platforms like NeuroVR[27] and Sim-Ortho[28], and real-time movement and force application detection tools such as the SmartForceps System[31] and Electromyographic Force Measurement.[29] The trained models can classify participants based on simulation data, enabling either binary or multiclass classification, depending on the model.

Four of these articles[25,26,29,30] compared the accuracy of various AI models in correctly assessing each participant's expertise level.

Among the most frequently compared algorithms are Support Vector Machine (SVM), Linear Discriminant Analysis (LDA), Naïve Bayes (NB),[25,26,29,30] K-Nearest Neighbors

(KNN),[25,26,30] Decision Tree (DT),[25,29] ANN, Logistic Regression (LR), and Random Forest (RF).[30]

The accuracy values of these models are summarized in **Figure 2**. Notably, the LDA model demonstrated high precision, ranking among the top three models with the highest accuracy in three of the four articles,[25,26,29] with accuracy values of 87.8%, 78%, and 81.9%, respectively. Following this, the SVM,[25,30] NB,[26,29] and KNN[25,26] models appeared twice among the top three models with the highest accuracy.



Source: Elaborated by the authors.
Note: Models absent from the figure were not reported in the respective study.
Figure 2 – Accuracy of AI models used for assessing the participant's expertise level.

These results suggest the substantial potential of integrating these models into surgical assessment systems, offering the opportunity to create a hybrid evaluation model. In such a model, AI would serve as a critical tool for students to assess their development of surgical skills and benchmark their performance against expected trends for their expertise levels. Furthermore, this approach enables the provision of targeted feedback, which can be utilized to enhance skills at each proficiency level.

Given these considerations, it is clear that enhancing these models requires adaptation to more comprehensive datasets, incorporating a greater number of procedures and reference data for participant evaluation.[28] One major limitation is the small performance data sets used in each study, which hinder the application of these trained models beyond their institutional contexts when confronted with new data.[26] Additionally, there is an

emphasis on the importance of using more complex and realistic simulation scenarios, achieved by employing materials with biomechanical properties similar to real tissues[29] and a broader variety of surgical instruments, thereby enabling more accurate evaluative results that reflect the true skill level of each surgical student.[25]

Additionally, four studies[25,27,28,31] utilized AI-provided feedback during participant evaluations to describe specific behavioral differences, linking these differences to the participants' levels of experience.

Alkadri *et al.*[28] demonstrated the utility of using the relative importance of variables identified by an ANN to distinguish performance characteristics assigned to participants' surgical categories. By training an AI model with simulation data from 100 participants, classified via the Sim-Ortho platform, the ANN identified and adjusted the relative importance of various parameters, correlating them with participants' expertise levels. The study found that contact time and contact length were precise indicators for the ANN, with variations in these parameters serving as key evaluation points for surgical skills. Understanding these metrics and their relative importance can help identify critical skills needing improvement among students and residents, thereby facilitating the development of personalized training programs to enhance these specific skills.

Similarly, three articles[25,27,31] describe behavior trends in different participant groups at various stages of simulated surgery. In the educational context, data such as the smaller inter-tool tip distance observed in expert surgeons compared to novice surgeons (46.78 vs. 75.92 pixels), which reflects greater skill and control,[27] and the reduction in force applied to the dura mater during simulated hemilaminectomies,[25] can significantly aid in the development of both junior and senior surgeons. This data, provided as ergonomic feedback by ANN models, can be used as a reference during their training.

Baghdadi *et al.*[31] also presents the use of a Time-Series U-Net model for analyzing force profiles applied at various points during surgeries (50 neurosurgical tumor resection procedures) using the SmartForceps System, allowing for real-time force measurement. This model demonstrated 98% accuracy in identifying force profiles. It distinguished force application patterns associated with novices and experts, detected errors from excessive or insufficient force application, and offered feedback to help prevent undesirable outcomes in real surgical performance.

Furthermore, the model was trained to possess the capability to accurately identify the surgical task being performed in real-time. This includes distinguishing between procedures aimed at halting bleeding (coagulation) and those that are not (non-coagulation), such as manipulation, dissection, pulling, or retracting. With an accuracy rate of 82%, this technology showcases a promising future for the development of specialized robotic assistants.[31] These assistants could receive and process information about ongoing events in the operating room in real-time, thus enabling the establishment of care protocols for surgeons based on the gathered data.

Other Applications in Neurosurgical Education

In addition to the applications already discussed, three articles[32–34] highlighted other uses of AI, showcasing its great potential in neurosurgical education.

Sevgi *et al.*[32] evaluated various applications of ChatGPT, focusing on its reliability in generating clinical cases and drafting academic articles to identify potential limitations and strengths within the surgical academic field. ChatGPT produced three multiple-choice questions—two of which were correct and complete, while one contained two correct answers, which is considered an error in the context of examinations. Additionally, it generated two clinical cases using provided keywords: one involving a patient with meningitis complicated by an epidural hematoma following a lumbar puncture and another concerning a patient with leg pain treated with endoscopic lumbar discectomy. Both the format and content of the generated questions were appropriate, and the clinical cases were coherent and clinically sound.

Therefore, with updates to these models and the incorporation of databases to further support AI, this technology could become a highly valuable tool for developing question banks for neurosurgical residents and creating examinations during their training programs. It is essential to consider the joint participation of expert specialists to ensure the quality of the outcomes.

In another article, Witten *et al.*[33] aimed to develop an AI-based neuroanatomy image segmentation algorithm. They used a dataset of 879 neuroanatomical images from *The Neurosurgical Atlas*, which were manually segmented into categories like arteries, veins, nerves, gross brain tissue, and bone. The images were cleaned to eliminate errors, incorrect rotations, and anomalous labels.

A convolutional neural network model based on a Pyramidal Scene-Parsing Network was trained on these images, divided into 679 for training, 100 for validation, and 100 for testing. Data augmentation techniques such as rotation, random cropping, and blurring were employed to enhance generalization.[33]

The model achieved an overall test accuracy of 91.8%, a training accuracy of 97.1%, and a validation accuracy of 94.5%. Effectiveness varied across tissue types, with an intersection-over-union (IoU) of 0.887 for arteries and 0.674 for brain tissue. The IoU metric, ranging from 0 to 1, evaluates image segmentation model performance, with 1 indicating perfect pixel classification. Variations in scores highlighted challenges in distinguishing brain tissue from nervous tissue due to imprecise manual labels, suggesting the need for a more exhaustive database.[33]

These segmentation models could significantly enhance neurosurgical training by integrating with augmented reality or virtual simulation technologies. They would provide immediate information on anatomical structures being visualized or manipulated and assess students' ability to identify these structures before actual procedures.

Balu *et al.*[34] focused on training the YOLOv4 model to identify surgical instruments in video frames from endonasal simulation procedures treating internal carotid artery injuries, sourced from the Simulated Outcomes After Carotid Artery Laceration (SOCAL) database. The model demonstrated an accuracy of 0.594, which increased to 0.743 with metallic instruments due to greater contrast.

Based on this detection, Shannon entropy (ShEn)—a measure of the variability in combinations of surgical instruments— was calculated. The capacity of a logistic regression model based on ShEn scores to predict surgical procedure success according to observed instrumentation entropy was evaluated. The model achieved 78.3% accuracy with manually annotated instruments and 73.9% with automated detections from YOLOv4, demonstrating the potential of combining both models.[34]

The study showed that procedures with higher entropy were more likely to be successful (with successful procedures having a ShEn of 0.452 compared to 0.370 for unsuccessful ones) and identified instrumentation patterns that differentiated likely successful procedures from unsuccessful ones. Failed attempts lacked evolution in surgical processes.[34]

It is important to note that in heterogeneous surgical procedures, entropy levels might be altered and not directly associated with success or failure. However, analyzing instrument use patterns may contribute to establishing an educational framework, thereby developing more structured and efficient skills.

DISCUSSION

The review identified key areas where AI enhances neurosurgical education based on the synthesis of evidence from 23 selected articles.

In the realm of neurosurgical board examinations, AI models, particularly ChatGPT-4, demonstrated high accuracy in answering exam questions, surpassing earlier versions and other models such as Microsoft Bing and Google Bard. These results align with the availability of more comprehensive databases, optimized machine learning processes, and updates that enable models to visualize images, thereby equipping AI with additional tools to address these questions. Studies highlighted the superior performance of AI models compared to medical students and residents in theoretical assessments, with performance proportional to the specificity of the question and the amount of data included. This underscores the importance of contextualization, clarity, and structure when utilizing machine learning models.

However, this superiority over students and residents diverges from the real needs of clinical-surgical practice, which requires a skill set beyond mere text comprehension and data analysis. Ethical concerns also arise regarding the use of these models, emphasizing the need to avoid dependence on their capabilities due to observed blind spots and potential errors that could have consequences in medical practice. This underscores the necessity of employing these technologies only when patient safety is ensured and avoiding their irresponsible application in medical training.

In neurosurgical simulation-based training and tutoring, AI models have been employed to classify participants' expertise levels during simulated procedures and provide personalized feedback according to observed behaviors. This represents a significant opportunity to integrate these models into neurosurgical education programs, as studies have demonstrated a relationship between AI-driven feedback and significant improvement in surgical skills. The reviewed studies underscored the importance of multilayer classification for specific categorization based on expertise levels. They also identified different learning patterns among participants, allowing for the recognition of specific needs at each training stage and the development of educational tools to address identified challenges. This application was further reflected in the ability of these models to create neurosurgical training plans and facilitate the comparison of various feedback

modalities, such as AI-driven real-time feedback, remote expert instruction, and numerical, visual, and visuospatial feedback, demonstrating AI's potential in this field.

The assessment of neurosurgical performance, skills, or expertise through AI models was another focal point. These models exhibited high accuracy in evaluating participants based on simulation data, allowing for the identification of behaviors typically associated with experts and novices, thereby guiding practical training. The studies stressed the necessity of more comprehensive datasets and realistic simulation scenarios to improve accuracy and applicability.

Additionally, other AI applications in neurosurgical education were explored, showcasing the technology's vast potential. For instance, AI models like ChatGPT were assessed for their capability to generate clinical cases and draft academic articles, indicating potential for developing question banks and examinations for residents, as well as contributing to research processes, provided they do not interfere with accuracy, transparency, and methodology. AI-based neuroanatomy image segmentation algorithms demonstrated high accuracy and could enhance training by integrating with augmented reality or virtual simulation technologies. The YOLOv4 model enabled the identification of surgical instruments with high precision, allowing the detection of usage patterns and predicting surgical success based on instrument variability.

The relevance of these findings extends to neurosurgical students and residents, medical educators and institutions, policymakers, and AI developers and researchers. AI offers tailored feedback, skill development, and adaptive learning experiences for students and residents. For educators and institutions, AI provides supplementary evaluative tools and personalized training plans. Policymakers must address ethical concerns to ensure responsible use, while developers and researchers should refine AI models, enhance datasets, and collaborate with medical experts to further AI's applications in neurosurgical education.

Limitations

This scoping review has limitations. The broad inclusion criteria led to significant heterogeneity in study designs and outcomes, making it difficult to draw definitive conclusions. The narrative synthesis approach, while providing a rich contextual understanding, limited quantitative aggregation and comparability of results. Additionally, the review may have missed relevant studies due to the constraints of the

search strategy, language restrictions, and the unavailability of full texts. The absence of a formal risk of bias assessment means the findings should be interpreted with caution. Finally, given the rapid advancements in AI technology, the review's findings may quickly become outdated, necessitating continuous updates to stay current with the evolving field.

Conclusions

The results of this scoping review highlight the significant potential of AI in neurosurgical education. AI models have demonstrated significant benefits in enhancing board examination performance, providing valuable feedback during simulation-based training, accurately assessing surgical skills, and offering innovative educational tools. However, the findings underscore the importance of addressing ethical considerations, utilizing comprehensive datasets, and ensuring realistic simulation scenarios.

Moving forward, the continuous development and refinement of AI models, coupled with collaboration between educators, AI developers, and policymakers, will be essential to maximize the effective and responsible integration of AI in neurosurgical training.

REFERENCES

1. Zhang L, Pan Y, Wu X, Skibniewski MJ. Introduction to Artificial Intelligence. Lect Notes Civ Eng. 2021;163:1-15.
2. Darda P, Pendse MK. The impact of artificial intelligence (AI) transformation on the financial sector from the trading to security operations. In: Khang A (ed). Shaping Cutting-Edge Technologies and Applications for Digital Banking and Financial Services. London: Routledge; 2025. p. 322-39.
3. Tiwari S. The Rise of Intelligent Machines: An Introduction to Artificial Intelligence. In: Artificial Intelligence and Machine Learning in Drug Design and Development. New York: Scrivener Publishing; 2024. p. 1-22.
4. Wu Z, Liu Z, Zhang Z. Research on the current situation and future governance of artificial intelligence in the field of medical education. Chin Med Ethics. 2024;37(9):1093-100.
5. Zhang J, Li C, Lyu W. The application prospects and challenges of artificial intelligence in medical education, medical research, and clinical practice. Chin J Gen Pract. 2024;22(7).
6. Nagi F, Salih R, Alzubaidi M, Shah H, Alam T, Shah Z, et al. Applications of Artificial Intelligence (AI) in Medical Education: A Scoping Review. Stud Health Technol Inform. 2023;305:648-51.
7. Sohail N, Puyana C, Zimmerman L, Tsoukas MM. Artificial intelligence in dermatology: Bridging the gap in patient care and education. Clin Dermatol. 2024;42(5):434-6.
8. Zhao H, Li G, Feng W. Research on application of artificial intel-ligence in medical education. In: Proc- Int Conf Eng Simul IntellControl ESAIC. Hunan, China; 2018 Nov 9. pp. 340–2
9. Page MJ, Moher D, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. BMJ. 2021;372:n160.
10. Tricco AC, Lillie E, Zarin W, O'Brien KK, Colquhoun H, Levac D, et al. PRISMA Extension for Scoping Reviews (PRISMA-ScR): Checklist and Explanation. Ann Intern Med. 2018;169(7):467-73.
11. Beller EM, Glasziou PP, Altman DG, Hopewell S, Bastian H, Chalmers I, et al. PRISMA for Abstracts: Reporting Systematic Reviews in Journal and Conference Abstracts. PLOS Med. 2013;10(4):e1001419.
12. Ali R, Tang OY, Connolly ID, Zadnik Sullivan PL, Shin JH, Fridley JS, et al. Performance of ChatGPT and GPT-4 on Neurosurgery Written Board Examinations. Neurosurgery. 2023;93(6):1353-65.
13. Ali R, Tang OY, Connolly ID, Fridley JS, Shin JH, Zadnik Sullivan PL, et al. Performance of ChatGPT, GPT-4, and Google Bard on a Neurosurgery Oral Boards Preparation Question Bank. Neurosurgery. 2023;93(5):1090-8.
14. Guerra GA, Hofmann H, Sobhani S, Hofmann G, Gomez D, Soroudi D, et al. GPT-4 Artificial Intelligence Model Outperforms ChatGPT, Medical Students, and

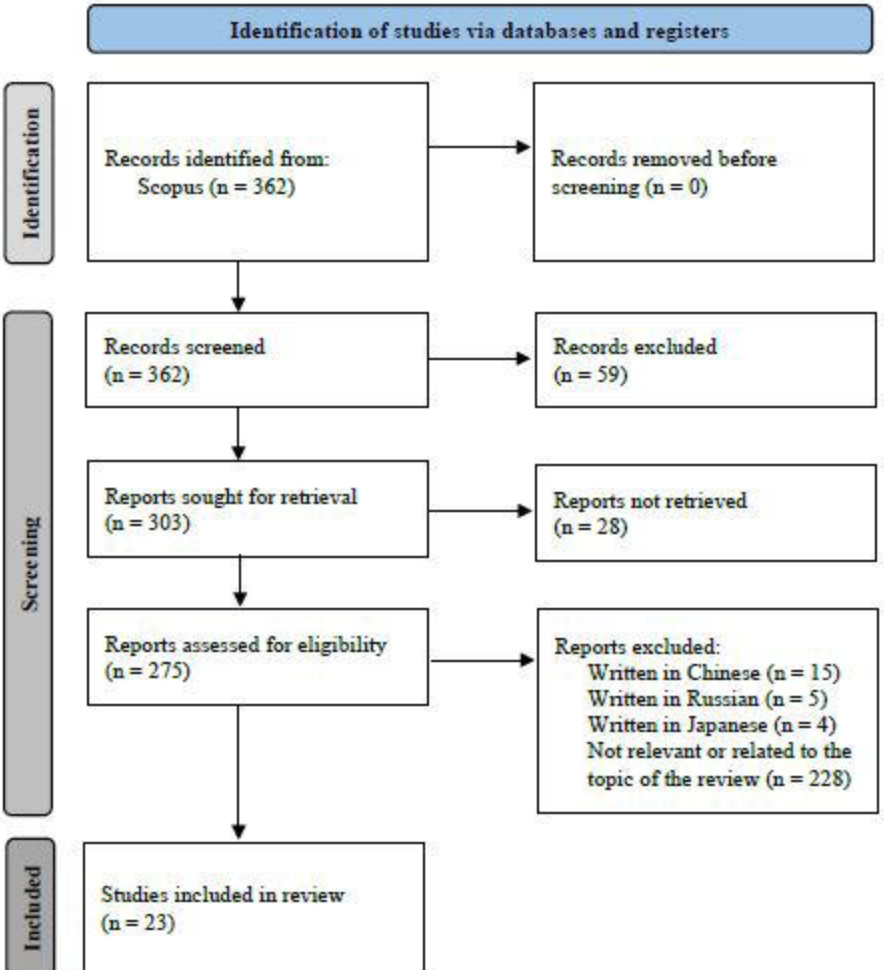
Neurosurgery Residents on Neurosurgery Written Board-Like Questions. World Neurosurg. 2023;179:e160-5.
15. Hopkins BS, Nguyen VN, Dallas J, Texakalidis P, Yang M, Renn A, et al. ChatGPT versus the neurosurgical written boards: a comparative analysis of artificial intelligence/machine learning performance on neurosurgical board–style questions. J Neurosurg. 2023;139(3):904-11.
16. Stengel FC, Stienen MN, Ivanov M, Gandía-González ML, Raffa G, Ganau M, et al. Can AI pass the written European Board Examination in Neurological Surgery? - Ethical and practical issues. Brain Spine. 2024;13:4:1027654.
17. Sahin MC, Sozer A, Kuzucu P, Turkmen T, Sahin MB, Sozer E, et al. Beyond human in neurosurgical exams: ChatGPT's success in the Turkish neurosurgical society proficiency board exams. Comput Biol Med. 2024;169:107807.
18. Liu J, Zheng J, Cai X, Wu D, Yin C. A descriptive study based on the comparison of ChatGPT and evidence-based neurosurgeons. iScience. 2023;26(9):107590.
19. Bartoli A, May AT, Al-Awadhi A, Schaller K. Probing artificial intelligence in neurosurgical training: ChatGPT takes a neurosurgical residents written exam. Brain Spine. 2024;29:4:102715.
20. 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.
21. 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.
22. 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.
23. Fazlollahi AM, Yilmaz R, Winkler-Schwartz A, Mirchi N, Ledwos N, Bakhaidar M, et al. AI in Surgical Curriculum Design and Unintended Outcomes for Technical Competencies in Simulation Training. JAMA Netw Open. 2023;6(9):e2334658.
24. Yilmaz R, Fazlollahi AM, Winkler-Schwartz A, Wang A, Makhani HH, Alsayegh A, et al. Effect of Feedback Modality on Simulated Surgical Skills Learning Using Automated Educational Systems– A Four-Arm Randomized Control Trial. J Surg Educ. 2024;81(2):275-87.
25. Bissonnette V, Mirchi N, Ledwos N, Alsidieri G, Winkler-Schwartz A, Del Maestro RF. Artificial Intelligence Distinguishes Surgical Training Levels in a Virtual Reality Spinal Task. J Bone Jt Surg Am. 2019; 101(23):e127.
26. 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.

27. Davids J, Makariou SG, Ashrafian H, Darzi A, Marcus HJ, Giannarou S. Automated Vision-Based Microsurgical Skill Analysis in Neurosurgery Using Deep Learning: Development and Preclinical Validation. World Neurosurg. 2021;149:e669-86.
28. Alkadri S, Ledwos N, Mirchi N, Reich A, Yilmaz R, Driscoll M, et al. Utilizing a multilayer perceptron artificial neural network to assess a virtual reality surgical procedure. Comput Biol Med. 2021;136:104770.
29. Singh R, Godiyal AK, Chavakula P, Suri A. Craniotomy Simulator with Force Myography and Machine Learning-Based Skills Assessment. Bioengineering. 2023;10(4):465.
30. Natheir S, Christie S, Yilmaz R, Winkler-Schwartz A, Bajunaid K, Sabbagh AJ, et al. Utilizing artificial intelligence and electroencephalography to assess expertise on a simulated neurosurgical task. Comput Biol Med. 2023;152:106286.
31. Baghdadi A, Lama S, Singh R, Sutherland GR. Tool-tissue force segmentation and pattern recognition for evaluating neurosurgical performance. Sci Rep. 2023;13(1):9591.
32. Sevgi UT, Erol G, Doğruel Y, Sönmez OF, Tubbs RS, Güngor A. The role of an open artificial intelligence platform in modern neurosurgical education: a preliminary study. Neurosurg Rev. 2023;46(1):86.
33. Witten AJ, Patel N, Cohen-Gadol A. Image Segmentation of Operative Neuroanatomy Into Tissue Categories Using a Machine Learning Construct and Its Role in Neurosurgical Training. Oper Neurosurg. 2022;23(4):279-86.
34. Balu A, Pangal DJ, Kugener G, Donoho DA. Pilot Analysis of Surgeon Instrument Utilization Signatures Based on Shannon Entropy and Deep Learning for Surgeon Performance Assessment in a Cadaveric Carotid Artery Injury Control Simulation. Oper Neurosurg. 2023;25(6):E330-7.

FUNDING
None.

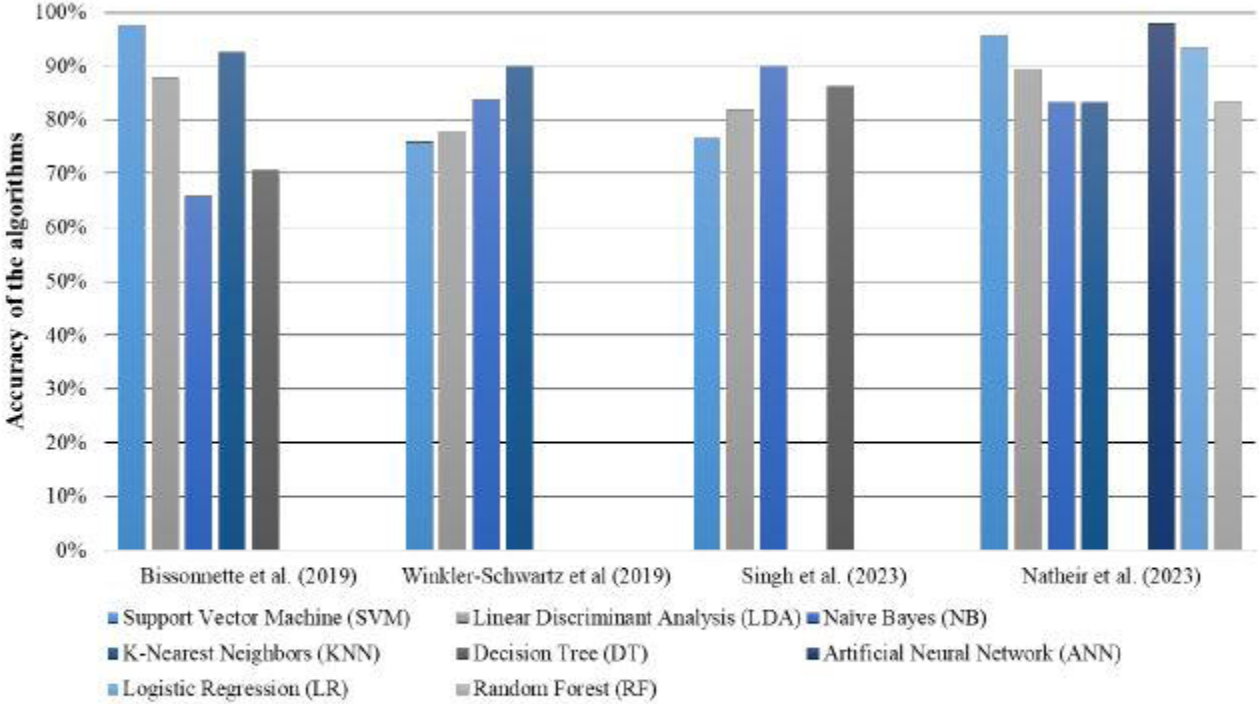
CONFLICT OF INTERESTS
None.

AUTHOR CONTRIBUTIONS
Conceptualization: *Hector Julio Piñera-Castro.*
Data curation: *Hector Julio Piñera-Castro, Christian Borges-García.*
Formal analysis: *Hector Julio Piñera-Castro, Christian Borges-García.*
Investigation: *Hector Julio Piñera-Castro, Christian Borges-García.*
Methodology: *Hector Julio Piñera-Castro.*
Resources: *Hector Julio Piñera-Castro.*
Supervision: *Hector Julio Piñera-Castro.*
Visualization: *Hector Julio Piñera-Castro, Christian Borges-García.*
Writing – original draft: *Hector Julio Piñera-Castro, Christian Borges-García.*
Writing – review & editing: *Hector Julio Piñera-Castro, Christian Borges-García.*



Source: Elaborated by the authors.

Figure 1 – Flowchart of the search and study selection process.



Source: Elaborated by the authors.

Note: Models absent from the figure were not reported in the respective study.

Figure 2 – Accuracy of AI models used for assessing the participant's expertise level.