Education and Information Technologies (2024) 29:4611–4633
https://doi.org/10.1007/s10639-023-12009-8

AI in Medical Education: Global situation, effects and challenges

Wei Zhang[1] · Mingxuan Cai[1] · Hong Joo Lee[2] · Richard Evans[3] · Chengyan Zhu[4] · Chenghan Ming[5]

Received: 11 March 2023 / Accepted: 26 June 2023 / Published online: 10 July 2023

Abstract

Purpose Artificial Intelligence (AI) is transforming healthcare and shows considerable promise for the delivery of medical education. This systematic review provides a comprehensive analysis of the global situation, effects, and challenges associated with applying AI at the different stages of medical education.

Methods This review followed the PRISMA guidelines, and retrieved studies published on Web of Science, PubMed, Scopus, and IEEE Xplore, from 1990 to 2022. After duplicates were removed (n = 1407) from the 6371 identified records, the full text of 179 records were screened. In total, 42 records were eligible.

Results It revealed three teaching stages where AI can be applied in medical education (n = 39), including teaching implementation (n = 24), teaching evaluation (n = 10), and teaching feedback (n = 5). Many studies explored the effectiveness of AI adoption with questionnaire survey and control experiment. The challenges are performance improvement, effectiveness verification, AI training data sample and AI algorithms.

Conclusions AI provides real-time feedback and accurate evaluation, and can be used to monitor teaching quality. A possible reason why AI has not yet been applied widely to practical teaching may be the disciplinary gap between developers and end-user, it is necessary to strengthen the theoretical guidance of medical education that synchronizes with the rapid development of AI. Medical educators are expected to maintain a balance between AI and teacher-led teaching, and medical students need to think independently and critically. It is also highly demanded for research teams with a wide range of disciplines to ensure the applicability of AI in medical education.

Keywords Medical education · Artificial intelligence · Applications of AIMED · Effectiveness of AIMED · Challenges of AIMED

Extended author information available on the last page of the article

1 Introduction

The start of the new millennium with significant developments in Information and Communication Technologies (ICTs), highlights the growth, maturity, and evolution of Artificial Intelligence in Education (AIED). Developed countries represented by the United States have formulated a series of strategies to seize the opportunity to apply AI in education. From 2016 to 2018, the United States released three important policy reports on national strategies for AI (i.e., Preparing for the Future of Artificial Intelligence, The National Artificial Intelligence Research and Development Strategic Plan, and A National Machine Intelligence Strategy for the United States), indicating a huge potential for AI in education. In November 2021, the United Nations Educational, Scientific and Cultural Organization (UNESCO) released *Reimagining Our Future Together: A New Social Contract for Education*, which introduced the international consensus framework for the deep coupling and mutual enhancement of AI and education. In addition, significant research efforts have been made to demonstrate the benefits of AI in education (L. Chen et al., 2020; Hwang et al., 2020; Ouyang et al., 2022). The 2019 Horizon Report also suggested that AI can personalize learners' experiences, reduce the workload of both students and teachers, and assist in analyzing large and complex data-sets, concluding considerable areas for AI in education.

In 1984, a report by the American Association of Medical Colleges (AAMC) claimed that medical schools should reduce their dependence on lectures as the principal method of teaching, and should provide students with more opportunities for independent learning and problem solving (Muller, 1984). The field of medicine has been experiencing significant growth in new knowledge, and the medical education is lifelong in nature. As an effective tool and knowledge platform in medical education, AI will undoubtedly have a profound impact, reducing deficiencies in education.

One of the earliest studies on AI in medical education (AIMED) by Klar and Bayer (1990) suggested the novel idea of integrating expert knowledge into the computer-aided teaching of medicine. As a result, numerous scholars have committed to AIMED, for example, computer-aided diagnosis (Cheng et al., 2020; Fang et al., 2022; Qian et al., 2022), intelligent tutoring systems (Hu et al., 2019; Nakawala et al., 2018; Voss et al., 2000), and chatbots (Li et al., 2021), to prove the application effects of AI in optimizing the learning, teaching, and management of medical education. Some studies have also presented retrospective summaries of AIMED using different approaches, such as by analyzing its sub-disciplinary areas (Hosny et al., 2018; Kirubarajan et al., 2022; Lazarus et al., 2022), its application and challenges at the various stages of medical education (Gorospe-Sarasúa, Munoz-Olmedo, et al., 2022; J. Lee et al., 2021), and the application and challenges of AIMED (K. S. Chan & Zary, 2019). However, although significant interest in AIMED, few studies have examined the effects of AI application in medical education, with most extant research being limited to a single stage in education or a specific discipline. Therefore, this systematic review focuses on the application scenarios and effects of AI at the different stages of

medical education. This study aims to: (1) review the variation of AIMED uses, (2) evaluate the effectiveness of AIMED, and (3) summarize the challenges in applying AI in medical education.

2 Methods

2.1 Data sources and search strategy

This systematic review was conducted in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines. Web of Science, PubMed, Scopus and the IEEE Xplore digital library were adopted to retrieve studies on *artificial intelligence in medical education* published in the English language from 1 January 1990 to 31 December 2022. For the databases, we mainly consider databases related to computer science, medicine and education. Web of Science Core Collection and Scopus involves wide range of studies with computer science, education and health sciences included. While, PubMed mainly included studies in biomedicine and health, and IEEE Xplore features on studies in computer science and electronics. Additionally, the references of relevant reviews published were also checked to include any potentially useful and overlooked studies. Details of the search strategies are provided in Supplementary Appendix 1.

2.2 Inclusion and exclusion criteria

The inclusion criteria are as follows, (1) studies focus on any stage of medical education, including undergraduate training, specialized training and continuing medical education; (2) studies include any technique for applying AI to a certain medical education scenario. While, the exclusion criteria are as follows, (1) studies published as reviews, conference abstracts, bibliographic chapters, news, or letters; (2) studies focused on the use of technology without incorporating AI; (3) studies irrelevant to medical education; (4) studies failed to provide specific AI applications; and (5) qualitative studies of perspectives/attitudes among health professionals.

2.3 Studies screening

Two graduate students in health informatics were trained to independently assess the titles and abstracts of studies based on the inclusion and exclusion criteria. A third reviewer was involved in reaching a consensus when disputes emerged. Among the 6371 studies identified, 1407 duplicated studies were eliminated. 179 studies were selected for full-text screening. Finally, 42 studies were included for the review. Figure 1 shows the details.

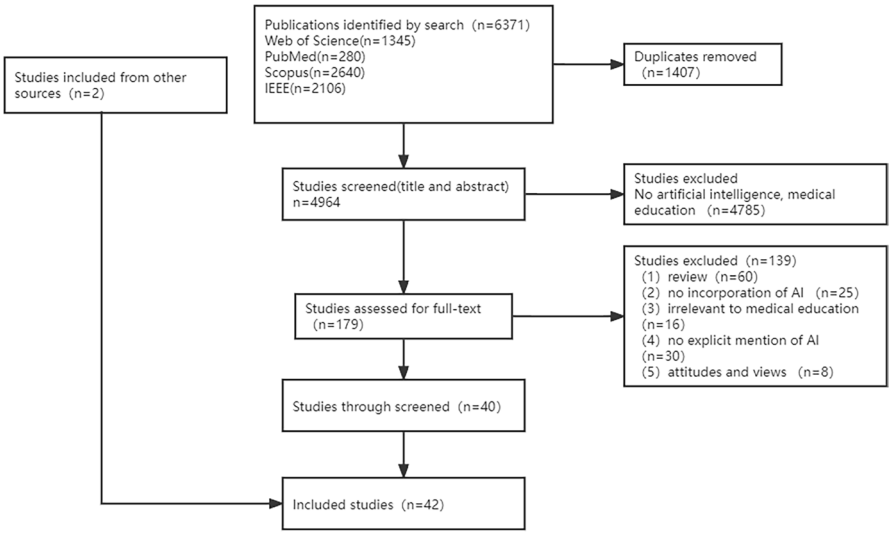


Fig. 1 Flowchart of the screening process

2.4 Data extraction

The data extracted from each study included, year of publication, study location, study classification (i.e., applied study, effectiveness study), study group, stage of medical education, and AI algorithms identified. Further detailed information was also extracted from applied and effectiveness studies. Details of the extracted data are provided in Supplementary Appendix 2 (Table 2, Table 3).

2.5 Data synthesis and analysis

The findings of each eligible study, including its main characteristics and results, are presented in Supplementary Appendix 2 (Table 1, Table 2, Table 3). A narrative synthesis with summarization of the results was also performed.

3 Results

3.1 Study of characteristics

Among the 42 studies, 39 are applied studies, and 19 are effectiveness studies. The yearly number of publication fluctuates as shown in Fig. 2. Since the first study identified in 2000, the number of publication grows slowly until 2010, followed by ups-and-downs. However, it grows in general with a majority of them published in the past seven years.

Regarding of the study location, 11 studies were developed in the USA, followed by eight studies in China and six studies in Canada. See Fig. 3 for details.

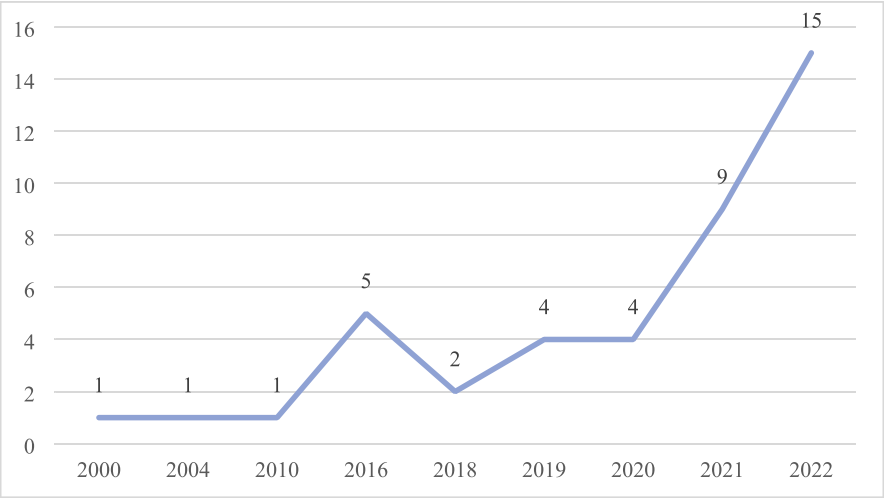
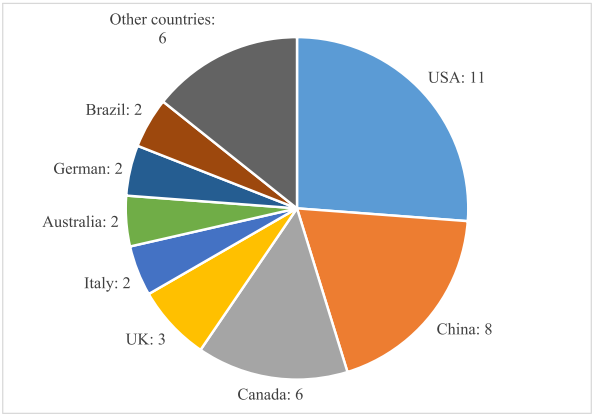


Fig. 2 The distribution of number of studies in year

Fig. 3 The distribution of study locations



Regarding target groups, we divided medical education phases into three stages, undergraduate training, specialized training, and continuing medical education. Over half of the studies (27/42) examined specialized training, and eight and seven studies featured undergraduate training and continuing medical education, respectively. While, six studies did not specify a stage of medical education, collectively referred to as medical education. Figure 4 presents the detail.

3.2 Use of AI in medical education

Through examination of the applied studies (n = 39), three main adoption of AI in medical education emerged in different teaching stages, including teaching implementation (n = 24), teaching evaluation (n = 10), teaching feedback (n = 5). Teaching

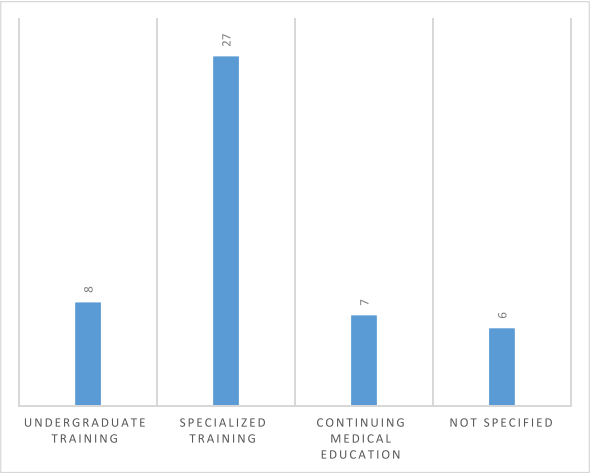


Fig. 4 The distribution of studies in terms of medical education phases

implementation refers to the teaching process of theory and practice education, and teaching evaluation is the evaluation of students' phased learning performance, while, teaching feedback focuses on the reflection and summary related to the quality improvement of medical education.

3.3 Teaching implementation

Teaching implementation is a core component in the cultivation of medical talents, and both theoretical teaching and clinical teaching are critical important in medical education. Among 24 studies featured teaching implementation, 11 studies examined the use of AI in theoretical medical education, while 13 studies explored its use in clinical medical education.

Theoretical teaching aims at the acquisition of basic medical knowledge. Four studies specified AI in anatomy education. Previous research has demonstrated that virtual reality and augmented reality are effective approaches to learn anatomy for its ability of visualized learning in 3D for medical students. However, the user interaction enabled by them are limited, and learners can only rotate or magnify models by mouse click or screen touch (Kurniawan et al., 2018; Lee et al., 2020). AI with virtual reality and augmented reality is essential to improve user interaction and enhanced immersive experience. For example, one study based on AR, using deep convolutional neural network to establish a gesture recognition interface, which could identify eleven gestures such as Pan, Pinch, Fist, etc. (Karambakhsh et al., 2019). Another study created an anatomy learning platform embedded in mobile devices. The AI algorithm is able to identify and track human movements based on points and locations placed by the system, and thus users can learn 3D anatomy through real-time body tracking (Fajrianti et al., 2022). In addition, AI has been also applied to real-time feedback on quiz practice. For example, i-SIDRA e-learning system based on neural network allows for creating, collecting, analyzing responses to multiple-choice questions and

offering real-time feedback to students(Fernández-Alemán et al., 2016). Another study mentioned chatbots capable of answering and asking anatomy questions and providing immediate feedback (Li et al., 2021).

AI technologies, especially machine learning and deep learning, can also assist medical students in the recognition and diagnosis of medical images. For example, deep learning algorithms are used to assist students in identifying hip fracture images(Cheng et al., 2020), recognizing fundus images to diagnose ophthalmic diseases (Fang et al., 2022; Han et al., 2022; Qian et al., 2022), guiding the interpreter to correctly interpret the ECG by analyzing the interpreter's eye movements when observing the ECG (Sqalli et al., 2022). Meanwhile, AI is also applied to other medical disciplines. The SmartPath program adopts decision tree algorithms to assist students in glomerular pathology diagnosis and testing, complementing the remote pathology teaching model (Aldeman et al., 2021). Hu et al., (2019) designed a health statistics intelligent tutoring system (ITS) to automatically generate personalized content and share learning resources.

Clinical teaching is critical of cultivating and maintaining learners' clinical thinking and skills. Among the 13 studies adopting AI as an auxiliary tool in clinical practice, eight studies were designed to provide feedback and instruction to surgical interns. Surgical procedures included laparoscopic minimally invasive surgery, tumor resection, sutures, intravenous catheterization, and thoracentesis. AI algorithms that provide feedback included artificial neural network (ANN), convolutional neural network (CNN) and recurrent neural network (RNN). Seven studies out of eight aimed to enhance psychomotor skills in surgery(Gendia, 2022; Hisey et al., 2022; Islam et al, 2016; Sadeghi Esfahlani et al, 2020; Voss et al., 2000; Yilmaz et al., 2022; Zahiri et al., 2016). For example, Yilmaz et al., (2022) evaluated surgical hand performance at 0.2 s intervals by training a long short-term memory network, aiming to provide continuous assessment, intelligent instruction and risk warning of psychomotor skills. Three out of the seven studies mentioned providing real-time feedback through the analysis of surgical videos(Gendia, 2022; Hisey et al., 2022; Islam et al., 2016), with one study uploaded surgical videos to the cloud, enabling the video analysis based on different parameters and provide time-based step analysis(Gendia, 2022). Another study highlighted cognitive skills training, which designed an intelligent tutoring system for intraoperative situational awareness skills training using thoracentesis as an example (Nakawala et al., 2018).

Five studies discussed the use in clinical diagnostic skills training. One study constituted the CaseB-Pro prototype, which combines neural network and case-based reasoning to help medical students and physicians diagnose patients. The training of the Case-Based Memory Network (CBMN) model is conducted with data from classical prototypes and the associated actual cases. The system can generate appropriate questions based on the main symptoms entered by the user, which allows the user to continue to enter examination results of the case in question until the system finds the best matching case (Peter & Goodridge, 2004). Four studies mentioned the creation of virtual patients learning system to simulate real-life consultation scenarios (de Lima et al., 2016; Furlan et al., 2021; Wang et al., 2022; Yang et al., 2019). In such environments, students can communicate with virtual patients and make

diagnostic decisions in a safe environment. Such systems also provide feedback on the decisions made by students.

3.4 Teaching evaluation

Ten studies concentrated on AI in teaching evaluation. Among them, six studies noted that machine learning algorithms were able to distinguish surgical trainees into two groups, namely proficiency level and novice level(Alonso-Silverio et al., 2018; Bissonnette et al., 2019; Mirchi et al., 2020; Nagaraj et al., 2023; Siyar et al., 2020; Zhao et al., 2021). The common advantages include the extraction of objective and valid indicators and automated skills assessment.

AI can also be used to predict student performance, with two studies showed machine learning algorithms achieved higher accuracy than traditional statistical analysis methods(Baloul et al., 2022; Dharmasaroja & Kingkaew, 2016). Machine learning can also be used for automated scoring of answers. Lam et al., (2022a, 2022b) used Logistic regression, Random forest, XGBoost, and BERT to score short answer question (SAQ) and very short answer question (VSAQ) in online exam, and found that it could achieve high accuracy for question with short sentence length and strong answer certainty. In addition, Y. Yilmaz et al., (2022) explored natural language processing and machine learning to analyze a large workplace-based assessment (WBA) narrative comment data set for identifying trainees at risk.

3.5 Teaching feedback

Five studies discussed the use of AI in teaching feedback, emphasizing quality evaluation of teaching and medical curriculum. Three of the studies assessed the quality of feedback provided by teachers to students, demonstrating that machine learning can reliably identify low-quality, low-utility feedback and flag evaluators who repeatedly produce low-quality feedback (Neves et al., 2021; Ötles et al., 2021; Solano et al., 2021).

Two studies used machine learning algorithms to evaluate medical students' feedback on medical courses(Borakati, 2021; C.-K. Chen, 2010). Assessing medical students' feedback on courses is vital to monitor and improve the quality of medical education. Medical students' course feedback is always collected after the course by questionnaires, but manual systematic evaluation of large amounts of qualitative feedback is time-consuming. These two studies revealed machine learning algorithms are capable of analyzing large and multidimensional free-text data.

3.5.1 Effectiveness studies on AI in medical education

The development and application of AI in medical education have received considerable attention. However, scholars have also begun to explore the effects on medical education to demonstrate that AI-based training is superior to traditional training. Among 19 effectiveness studies reviewed, most of them focused on the stage of

teaching implementation (n = 14), and a small number of studies involved the stage of teaching evaluation (n = 5).

3.6 Teaching implementation

In terms of participants, four studies targeted on surgical residents and physicians, and seven on undergraduate and postgraduate students. Another two studies featured on medical students and healthcare personnel in medical imaging, medical students and supervisors, respectively, while one study was not explicitly specified as medical students.

In terms of sample size, the most extensive study has 200 participants with experimental groups (90 participants) and control groups (110 participants) to explore the effectiveness of the intelligent feedback tool i-SIDRA (Fernández-Alemán et al, 2016). The remaining studies ranged from 10–100 in sample size.

In terms of research method, four were questionnaire survey studies to collect users' subjective experiences and feelings after receiving the intervention (Hedderich et al., 2021; Hisey et al., 2022; Li et al., 2021; Shiang et al., 2022), and six were experiment-control studies to compare the difference before and after the intervention (Cheng et al., 2020; Fazlollahi et al., 2022; Furlan et al., 2021; Islam et al., 2016; Wang et al., 2022; Zahiri et al., 2016). The remaining four were mixed studies with quantitative and qualitative research methods being used (Fang et al., 2022; Fernández-Alemán et al., 2016; Nakawala et al., 2018; Yang et al., 2019).

In terms of study design, two studies applied the AI intervention to the practical teaching of medical education. In one study, 28 healthcare professionals participated in a 12-week AI course on medical imaging to examine participants' perceptions of AI and their self-perception skills (Hedderich et al., 2021). Another study introduced AI-DSS, an AI-based decision support system, into the clinical workflow of radiology(Shiang et al., 2022).

The remaining 12 studies did not apply AI to practical curricula, but rather for initial validation and evaluation. Only two studies investigated users' subjective feelings by questionnaire and qualitative feedback (Hisey et al., 2022; Li et al., 2021). Ten studies obtained more objective measures through experiments(Cheng et al., 2020; Fang et al., 2022; Fazlollahi et al., 2022; Fernández-Alemán et al., 2016; Furlan et al., 2021; Islam et al., 2016; Nakawala et al., 2018; Wang et al., 2022; Yang et al., 2019; Zahiri et al., 2016). Among them, four studies also synthesized users' subjective feelings, which were considered more reliable than users self-reporting(Fang et al., 2022; Fernández-Alemán et al., 2016; Nakawala et al., 2018; Yang et al., 2019). They were all controlled experiments. Specifically, the experimental group received AI intervention and the control group used traditional teaching methods (Cheng et al., 2020; Fang et al., 2022; Fazlollahi et al., 2022; Fernández-Alemán et al., 2016; Islam et al., 2016; Nakawala et al., 2018; Yang et al., 2019; Zahiri et al., 2016), or the participants were used as controls before and after the intervention (Furlan et al., 2021; Wang et al., 2022). For example, Cheng et al., (2020) conducted a Randomized Controlled Trial (RCT) that randomly divided 30 fifth-year undergraduate students into two groups, with

the experimental group receiving additional learning tests from a hip fracture detection system, HipGuide. They found that the experimental group made significant progress. Measures of intervention effectiveness included changes in task completion time, time required to reach a certain level of proficiency, number and score of correct answers, number of mistakes made, learning curve, etc. For example, six studies compared changes in scores or numbers of correct answers (Cheng et al., 2020; Fang et al., 2022; Fernández-Alemán et al., 2016; Furlan et al., 2021; Nakawala et al., 2018; Wang et al., 2022), while three studies compared the time to complete tasks and reach proficiency (Islam et al., 2016; Nakawala et al., 2018; Zahiri et al., 2016).

As for the outcome of the intervention, it can be assessed in terms of teaching effect, user evaluation and system performance. From the perspective of teaching effect, all eight studies proved their effectiveness through experiments with objective and quantitative indicators(Cheng et al., 2020; Fang et al., 2022; Fazlollahi et al., 2022; Fernández-Alemán et al., 2016; Furlan et al., 2021; Islam et al., 2016; Nakawala et al., 2018; Wang et al., 2022). One randomized controlled trial showed that participants trained in AI outperformed the control group even if they were not supported by AI-enhanced images in subsequent learning (Cheng et al., 2020). However, one study concluded that intelligent tutoring system for thoracentesis improved only slightly in aiding learning compared with traditional training (Nakawala et al., 2018).

In terms of user evaluation, participants in six studies showed positive attitudes about the usefulness and usability of the intervention(Fang et al., 2022; Fernández-Alemán et al., 2016; Hisey et al., 2022; Li et al., 2021; Nakawala et al., 2018; Yang et al., 2019). For example, after talking with anatomy chatbot, participants reported that their confidence in anatomy knowledge increased from 2.10 to 3.84 on a scale of one to five, and said that they were more willing to make mistakes and more engaged than when communicating with teachers (Li et al., 2021). At the same time, the participants also put forward suggestions for improvements, such as enhancing system flexibility (Yang et al., 2019) and user-friendliness(Nakawala et al., 2018). Two studies expressed different views. In a study that integrated AI-based decision support system (AI-DSS) into radiology workflow and curriculum(Shiang et al., 2022), participants who used the AI-DSS found that it could complement triage (83.3%) and troubleshooting (66.7%), but there was room for improvement in speed (41.7%), accuracy (33.3%), and diagnosis determination (16.7%). Another study ran a 12-week AI course on medical imaging to investigate participants' perceptions. It revealed that as participants gained a deeper understanding of AI, the answer of whether AI would bring benefits to patients in the foreseeable future changed, with participants becoming less optimistic after completion of the course. The authors argued that the reason may be that their increased knowledge affected their judgments towards the utility of AI (Hedderich et al., 2021).

In terms of system performance, two studies demonstrated the effectiveness of new AI interventions by comparing them with previous systems. One study demonstrated that the portable system PortCAS based on computer vision performed similarly to the traditional laparoscopic simulator FLS (Zahiri et al., 2016). The other study proved that an interactive dialogue training and evaluation system supports

multi-classification and multi-conclusion with higher context variable matching, compared with standard virtual patient system (Yang et al., 2019).

3.7 Teaching evaluation

Five studies were used to validate the role of machine learning in virtual surgical skill assessment (Alonso-Silverio et al., 2018; Bissonnette et al., 2019; Mirchi et al., 2020; Nagaraj et al., 2023; Siyar et al., 2020). AI with machine learning algorithms can use large datasets to analyze the performance of operators. For example, in the study of Bissonnette et al., (2019), two groups of participants with different professional levels were recruited and defined in advance, and all participants were required to perform specified surgical operations on the NeuroVR neurosurgical simulator platform. Then they collected raw operation data, created indicators to train machine learning algorithms to predict whether participants were novice or proficient, and finally evaluated the accuracy of the algorithm through cross-verification. The performance of machine learning model is measured by the accuracy of the algorithm's classification against the basic facts labeled by humans.

In the selection of algorithms, one study used ANN (Alonso-Silverio et al., 2018), and one study used CNN (Nagaraj et al., 2023). Another three studies used support vector machine (SVM) (Bissonnette et al., 2019; Mirchi et al., 2020; Siyar et al., 2020), and two of them also selected k-nearest neighbor, linear discriminant analysis, naïve Bayes, decision trees, parzen window, and fuzzy k-nearest neighbor in order to compare the performance of multiple classifiers (Bissonnette et al., 2019; Siyaret al., 2020). Classification accuracy ranged from 83% to 97.6%. In a study by Bissonnette et al., (2019), it achieved the highest accuracy of 97.6% for support vector machine models. Siyar et al., (2020) proposed a study that supported vector machine performed best when the number of advanced features was increased to 15, with an accuracy value of 90%, outperforming the other three models (k-nearest neighbors, parzen window, fuzyy k-nearest neighbors).

3.7.1 Challenges of applying AI in medical education

The application of AI in medical education has many benefits. However, AI in medical education has several challenges as well. 17 studies discussed the challenges.

Six studies showed that the current prototypes or systems were immature with some technical limitations, which require further improvements in system performance to enhance user-friendliness (Cheng et al., 2020; de Lima et al., 2016; Karambakhsh et al., 2019; Nakawala et al., 2018; Yang et al., 2019; Zahiri et al., 2016). For example, two of these studies highlighted the need to improve interactivity, such as providing real-time automated feedback (Zahiri et al., 2016) and arrow markers (Cheng et al., 2020).

Ten studies concerned the effectiveness of AI applications in medical education. Among them, two studies mentioned that they had not been tested in medical students, so it was difficult to prove its effectiveness in actual medical education(Li et al., 2021; Sadeghi Esfahlani et al., 2020). Six studies noted the need to adopt

different evaluation strategies and the need for sufficient sample sizes in larger areas to prove effectiveness and reproducibility(Bissonnette et al., 2019; de Lima et al., 2016; Islam et al., 2016; Nakawala et al., 2018; Peter & Goodridge., 2004; Zahiri et al., 2016). Two other studies noted difficulties in verifying effects due to a lack of test sets (Furlanet al., 2021) and prior knowledge (Aldemanet al., 2021).

Three studies pointed to issues with samples used to train AI algorithms. One studies referred to the small sample size used to train AI models (Baloul et al., 2022), and one study pointed to quality issues with input data (Nagaraj et al., 2023), which can affect the quality of model training and the accuracy. Another study discussed the issue of data privacy and protection, where real data and sensitive information based on patients can raise legal issues if not handled properly(Gendia, 2022).

Three studies talked about AI algorithms. For example, two studies suggested that the *black box* nature of deep learning limits the interpretability of the results(Baloul et al., 2022; Yilmaz et al., 2022). One study raised the concern for the generalization ability of AI. As Bissonnette et al., (2019) put, testing in new populations is needed to ensure that the algorithm does not overfit.

4 Discussion

This paper presents a comprehensive analysis of the global situation, effects, and challenges associated with applying AI at the different stages of medical education. The main findings of our study are introduced below.

4.1 Global applications of AI in medical education

AI is mostly used in the teaching implementation stage. This may be because theoretical teaching and clinical teaching is an early and key phase in shaping medical talents. In basic medical courses for undergraduate medical students, AI is mostly applied to anatomy. This may attribute to the fact that anatomy is one of the most important foundation medical courses, and a deep understanding of anatomy is fundamental to physicians' safe clinical practice (Estai & Bunt, 2016). The introduction of AI with VR and AR provides students with a visual and interactive learning experience. However, anatomy education is fraught with uncertainty and ambiguity, including individual differences and diversity in human morphology (Willan & Humpherson, 1999), and differences between textbook anatomy and dissection (Stephenset al., 2021). AI is good at the development of anatomical structures and pattern recognition, which reduces *uncertainty* (Lazaruset et al., 2022). Medical students educated in AI may reduce their awareness of the diversity and variability of human morphology, producing the false perception that anatomical knowledge is limited and stable (Lazaruset et al., 2022). Therefore, AI-based teaching tools need to follow the particularities of anatomy. Developers should consider designing *uncertainty* into AI systems. Educators also need to guide students in actual teaching, such as raising conflicts, etc., to prompt students to reflect and question.

Our review also found that deep learning well fits in image recognition, and can be adopted for disease detection, identification and analysis of medical image lesions. It can assist medical students in the interpretation of medical images. By building deep neural network models, training large amounts of data and learning useful data features, deep learning is able to make predictions or classifications. Compared with traditional machine learning algorithms, deep learning can discover feature representations through training without manually designed features. However, before introducing deep learning into medical image processing, possible challenges must be taken into considerations. It is highly dependent on the quality and quantity of training data, and the data of different categories are expected to be balanced. In practice, it is difficult to collect enough abnormal cases (H.-P. Chan et al., 2020). The lack of suitable datasets is one of the biggest barriers to the adoption of deep learning (Razzak et al., 2018). Meanwhile, annotation of large medical image datasets requires much time and effort from medical experts and it alwasys needs multiple expert opinions to solve human errors (Razzak et al., 2018). In addition, there may be legal and ethical issues with the use of real clinical imaging data from patients to develop deep learning systems. If AI tools are introduced into medical education prematurely, without human oversight and an objective analysis of their strengths and limitations, inexperienced medical students may over-rely on the results produced by AI (Gorospe-Sarasúa, Muñoz-Olmedo, et al., 2022a, 2022b). Therefore, before applying AI-assisted tools, medical students should understand how to properly read and interpret medical images. Considering the *black box* nature of deep learning, medical students must also understand the principles of the complex algorithms hidden behind deep learning. In this sense, medical students can understand the varied scenarios of AI-based tools can be applied, and in what scenarios and for what reasons they may produce false results.

In surgical skills training, the important reason to apply AI is to provide feedback. It enables users to receive objective feedback and targeted training in real time, and correct erroneous operations in a timely manner. Studies have shown that learning from mistakes is significant (Foss, 1987). Feedback is a constructive and objective evaluation of performance (Bienstocket al, 2007) and is considered an important approach to improve learner performance (Bing-Youet et al., 2017; Hattie & Timperley, 2007). Feedback comes in two forms, namely, formative feedback and summative feedback. Formative feedback is the improvement of learners' behavior and performance over time (Wood, 2000), and summative feedback is based on comparisons of learners' overall behavioral performance to assign grades and recommend promotions (Bienstocket al., 2007). Although it is generally accepted that high-quality formative feedback is essential for learning, most medical students believe that they receive very little formative feedback, and the feedback they receive is often useless due to late timing and lack of details (Gilet al., 1984). This places a requirement on the use of AI to generate feedback: real-time and targeted. However, the feedback provided by AI is built on the knowledge base and model, which is not always targeted and contextual. For example, AI-based teaching platforms are difficult to understand learners' ongoing emotional and cognitive states. Consequently, it fails to observe medical students' changes and respond accordingly. In this sense, the AI-based teaching platforms cannot establish social connections with medical students,

and stimulate their thoughts and feelings like real medical teachers. Although AI-generated automated feedback enables objective and standardized training, detaching medical teachers from heavy tasks somehow, human connection and emotional communication remain important parts of medical education (Mirchi et al., 2020).

The reason for applying AI in teaching assessment is the demand for objective, repeatable and automated assessment tools. In surgery, visual scoring scales such as objective structured assessment of technical skills (OSATS) tool are considered gold standards for evaluating simulated tasks (Szaszet al., 2015). Although this method has been shown to be effective and reliable (Martin et al., 1997; Niitsu et al., 2013), it relies on the presence of examiners. The number and time of qualified surgical education examiners is limited, and the evaluation results are subjective in some sense(K. Lam et al., 2022a, 2022b; Mirchi et al., 2020). Machine learning enables objective classification of the level of surgical trainees at a scale and speed that humans cannot achieve. This technology allows surgical trainees to receive regular feedback, allowing them to track their progress and improve proficiency, while freeing educators from routine teaching tasks and allowing more time to design educational programs. Although machine learning is promising in the field of surgical skills assessment, it also raises new concerns. Machine learning can only classify surgical trainees into novices and experts through operations on virtual surgical platforms. Some students may *deceive* algorithms through dexterous operation to obtain false ratings and an illusion of proficiency. It is difficult to distinguish whether students are merely proficient in operating the platform or in actual surgical skills. In addition, surgical skill is dependent on both psychomotor and cognitive proficiency. The result provided by the AI-based teaching platform can be used as a reference, and more importantly, it is necessary to evaluate the transferability of professional knowledge to real-life surgical scenarios. AI teaching platforms need to be rigorously validated by expert opinions and multiple research institutions. The AI generated results should also be applied carefully.

Another AI application is during the teaching feedback stage. Our review revealed that AI is rarely used in reviewing curricula like previous studies. Chan and Zary (2019) suggested that this may attribute to the limited digitization in medical education learning management systems. In the context of COVID-19, face-to-face teaching has been restricted because of lockdowns and social distancing measures, which provides opportunities for e-learning, such as Massive Open Online Courses (MOOCs). MOOCs have been shown to provide similar student satisfaction and lower costs when compared to traditional face-to-face instruction (Borakati, 2021). E-learning has demonstrated many benefits during the pandemic, and educators believed a promising for e-learning after the pandemic (Goh, 2021). E-learning generates a large amount date on learning and AI can, therefore, be leveraged to extract meaningful patterns of data and discover typically unidentified knowledge. Educational data mining can facilitate the restructuring of curricula and related research projects in the future.

In addition, this review also found studies applying AI to assess the quality of teacher feedback. Learners value effective feedback, especially when feedback is based on learner performance and tailored to their goals (Hewson & Little, 1998). Wolverton and Bosworth (1985) found that learners perceive the ability to provide

constructive feedback as one of the necessary qualities of a good clinical teacher, second only to clinical competence. Another study found that, high-quality feedback is the strongest predictor of high quality of teaching by medical students (Torreet al., 2003). It is essential for teachers to provide high quality feedback, which is a key factor in determining the quality of medical education. AI can not only assess the learning performance of students, but also can be used to monitor the teaching performance of educators and improve the quality of medical education.

4.1.1 *Effectiveness studies on AI in medical education*

We reviewed effectiveness studies from the perspectives of study population, study method, intervention process, and intervention outcome. There is still a lack of sufficient, high-quality evidence that AI can enable effective learning compared to traditional teaching methods. It is important to note that the effectiveness studies reviewed had small sample sizes, with a maximum sample size of only 200. Small sample sizes may have an impact on confidence in the results. Effectiveness studies require a large sample size for the results to be probabilistic compared with traditional methods of teaching(K. S. Chan & Zary, 2019). Biases in study design is another issue. There may be biases in the self-reporting of participants in the questionnaire, participants may be more interested in new technologies and tend to score higher in self-reports. Clear and quantifiable measures, control for confounding factors such as participant heterogeneity are the basis of controlled experimental studies.

Although most of studies have proved their effectiveness from the results, many studies are only simple verifications and have not been applied to practical teaching on a large scale. We believe that the reason why AI has not yet been applied to teaching on a large scale may be due to the disciplinary gap between the engineers who design and develop AI tools and the end-user population (healthcare personnel and medical educators), especially the mismatch between the rapid development of AI and the slow updating of theoretical guidance on how to integrate AI into medical education. Luan et al., (2020) suggested the need for cognitive and educational psychology theories to guide development teams, understand how AI applications interact with the intrinsic abilities of individual learners to develop the best tools, algorithms, and practices for personalized learning, and to promote the application and effectiveness of AIMED.

4.1.2 Social and ethical implications of AI in medical education

Data-driven AI has the potential to replicate, reinforce, and amplify the inequalities and discrimination that exist in society. The features, metrics, and structure of the training AI model are all selected by the designer, which may embed the designer's bias. The results of AI output may have social discrimination and racial bias that affect diversity and inclusion in medicine. For example, most AI models are designed according to the white race. When confronted with black patients, it can trigger unconscious bias in learners. Responsible data collection, processing, and management are essential components of algorithmic fairness. Developers should

pay attention to the completeness and richness of data sources in design and development engineering, and take necessary measures to correct sample bias.

AI systems need to acquire large amounts of real patient data to learn and train. The protection and privacy of medical data is a concern. The Health Insurance Portability and Accountability Act (HIPAA) specifies patients' legal rights to their personally identifiable information and the obligations of healthcare personnel to protect patient information. Scholars have explored solutions for healthcare data protection. For example, Vayena and Blasimme (2017) proposed three ways to strengthen individual control over data, including clarifying the right to data portability, establishing new models of informed consent, and allowing participatory governance. Another study developed an adaptive privacy protection algorithm based on distributed integration strategy, which can learn the distribution of data accurately while protecting patients' sensitive data (Y. Li et al., 2016). How to balance the protection of patient privacy and the benefits of AI is still a question worthy of further explorations.

Another promising direction is chatbots. ChatGPT, a generative pre-trained transformer (GPT) language model, has attracted much attention. It was first opened to the public in November 2022, which is capable of generating human-like text and engaging in interactive conversations with user. In medical education, it has been proven to be effective, such as providing real-time and personalized feedback, designing and answering medical questions, etc. However, the challenges and limitations must also be considered. The accuracy and reliability of the information provided by ChatGPT is questionable, and if students rely on it as their primary source of information, it will negatively impact their critical thinking and problem-solving skills. In addition, ChatGPT can raise questions of academic dishonesty, for instance, cheating in online exams and assignments, writing academic papers, etc. A recent study conducted an experiment and found that abstracts generated by ChatGPT were able to fool professionals and educators (Hisan & Amri., 2023). The rapid development of language models poses a threat to educational ethics and should be applied with caution in medical education.

4.1.3 The future of medical education for artificial intelligence

Educators need to attend training to clarify the advantages and risks of AI and understand how to strike a balance between AI application and educator-led teaching. It's necessary for educators to understand when human intervention should be prioritized. For example, AI learning platforms cannot provide emotional feedback and understand the emotional state of learners, thus they cannot always provide appropriate feedback. Educators can establish social connections with students, and encourage students' enthusiasm for learning through interpersonal skills. It is important for educators to develop teaching and psychological learning theories for guidance on the application of AI in medical education.

Learners should understand the limitations of AI, develop the ability to think independently, and remain critical of AI. Strengthening medical humanities education is an important prerequisite for the integration of AI into medical education.

Scholars should form multidisciplinary teams to jointly develop AI education systems, and these should include engineers, data scientists, medical educators and students, in order to narrow the disciplinary gap between technology development and medical education, and ensure the applicability and effectiveness of AI in medical education.

4.1.4 Limitations

The number of studies included in this review is limited. We may have missed some important grey literature, such as dissertations and conference proceedings. Future reviews may consider grey literature and pedagogy-related databases. Due to the wide range of fields incorporated, we only provide a general overview to understand the use, effectiveness and challenges of AI in medical education at different stages of teaching. More specific and targeted conclusions must be refined into various subfields for study.

5 Conclusion

This review has comprehensively explored the application, effectiveness and challenges of AI in medical education based on different stages of education. In teaching implementation, the combination of AI and VR is an important trend to provide immersive training environment and real-time feedback. It is widely adopted in anatomy, surgical skills training and clinical thinking training. In teaching evaluation, AI is mainly used to assess the skill level of surgical trainees with binary classifications, namely the novice and the specialist. AI is also conducive to the quality improvement of medical teaching in teaching feedback. Although most of the studies are simple verifications, they have demonstrated their effectiveness by questionnaire surveys and controlled experiments in terms of teaching effect, user evaluation and system performance. The main challenges identified are AI system performance, effectiveness verification, AI training data sample, and AI algorithms. It is necessary to strengthen the theoretical guidance of medical education that synchronizes with the rapid development of AI in the future.

6 Contributions

WZ, MC, HL, and CZ conceptualized the paper, WZ, MC, CZ and CM contributed in the data collection and analysis, WZ and MC wrote the draft, RE, HL, CZ and CM contributed to the Discussion and revision of the paper. All authors have agreed the submission.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10639-023-12009-8.

Funding This paper is supported by National Natural Science Foundation of China (Project No. 72104087, 72004070) and University-Industry Collaborative Education Program supported by Ministry of Education in China (220505084312449).

Data availability The datasets generated and analyzed during the current study are available from the corresponding author on reasonable request.

None.

Declarations

Conflict of interest None.

References

Aldeman, N. L. S., de SáUrtigaAita, K. M., Machado, V. P., da Mata Sousa, L. C. D., Coelho, A. G. B., da Silva, A. S., Silva Mendes, A. P., de Oliveira Neres, F. J., & do Monte, S. J. H. (2021). Smartpathk: A platform for teaching glomerulopathies using machine learning. *BMC Medical Education, 21*(1), 248. https://doi.org/10.1186/s12909-021-02680-1

Alonso-Silverio, G. A., Pérez-Escamirosa, F., Bruno-Sanchez, R., Ortiz-Simon, J. L., Muñoz-Guerrero, R., Minor-Martinez, A., & Alarcón-Paredes, A. (2018). Development of a Laparoscopic Box Trainer Based on Open Source Hardware and Artificial Intelligence for Objective Assessment of Surgical Psychomotor Skills. *Surgical Innovation, 25*(4), 380–388. https://doi.org/10.1177/1553350618777045

Baloul, M. S., Yeh, V.J.-H., Mukhtar, F., Ramachandran, D., Traynor, M. D., Shaikh, N., Rivera, M., & Farley, D. R. (2022). Video Commentary & Machine Learning: Tell Me What You See, I Tell You Who You Are. *Journal of Surgical Education, 79*(6), e263–e272. https://doi.org/10.1016/j.jsurg.2020.09.022

Bienstock, J. L., Katz, N. T., Cox, S. M., Hueppchen, N., Erickson, S., & Puscheck, E. E. (2007). To the point: Medical education reviews—providing feedback. *American Journal of Obstetrics and Gynecology, 196*(6), 508–513. https://doi.org/10.1016/j.ajog.2006.08.021

Bing-You, R., Hayes, V., Varaklis, K., Trowbridge, R., Kemp, H., & McKelvy, D. (2017). *Feedback for Learners in Medical Education: What Is Known? A Scoping Review*. Wolters Kluwer. https://doi.org/10.1097/ACM.0000000000001578

Bissonnette, V., Mirchi, N., Ledwos, N., Alsidieri, G., Winkler-Schwartz, A., Del Maestro, R. F., Yilmaz, R., Siyar, S., Azarnoush, H., Karlik, B., Sawaya, R., Alotaibi, F. E., Bugdadi, A., Bajunaid, K., Ouellet, J., & Berry, G. (2019). Artificial Intelligence Distinguishes Surgical Training Levels in a Virtual Reality Spinal Task. *Journal of Bone and Joint Surgery-American, 101*(23), e127. https://doi.org/10.2106/JBJS.18.01197

Borakati, A. (2021). Evaluation of an international medical E-learning course with natural language processing and machine learning. *BMC Medical Education, 21*(1), 181. https://doi.org/10.1186/s12909-021-02609-8

Chan, H.-P., Samala, R. K., Hadjiiski, L. M., & Zhou, C. (2020). Deep Learning in Medical Image Analysis. In G. Lee & H. Fujita (Ed.), *Deep Learning in Medical Image Analysis: Challenges and Applications* (pp. 3–21). Springer International Publishing. https://doi.org/10.1007/978-3-030-33128-3_1

Chan, K. S., & Zary, N. (2019). Applications and Challenges of Implementing Artificial Intelligence in Medical Education: Integrative Review. *JMIR Medical Education, 5*(1), e13930. https://doi.org/10.2196/13930

Chen, C.-K. (2010). Curriculum Assessment Using Artificial Neural Network and Support Vector Machine Modeling Approaches: A Case Study. IR Applications. Volume 29. In *Association for Institutional Research (NJ1)*. Association for Institutional Research. https://eric.ed.gov/?id=ED524832

Chen, L., Chen, P., & Lin, Z. (2020). Artificial Intelligence in Education: A Review. *Ieee Access, 8*, 75264–75278. https://doi.org/10.1109/ACCESS.2020.2988510

Cheng, C.-T., Chen, C.-C., Fu, C.-Y., Chaou, C.-H., Wu, Y.-T., Hsu, C.-P., Chang, C.-C., Chung, I.-F., Hsieh, C.-H., Hsieh, M.-J., & Liao, C.-H. (2020). Artificial intelligence-based education assists

medical students' interpretation of hip fracture. *Insights into Imaging, 11*(1), 119. https://doi.org/10.1186/s13244-020-00932-0

de Lima, R. M., de Medeiros Santos, A., Mendes Neto, F. M., de Sousa, F., Neto, A., Leão, F. C. P., de Macedo, F. T., & de Paula Canuto, A. M. (2016). A 3D serious game for medical students training in clinical cases. *IEEE International Conference on Serious Games and Applications for Health (SeGAH), 2016*, 1–9. https://doi.org/10.1109/SeGAH.2016.7586255

Dharmasaroja, P., & Kingkaew, N. (2016). Application of artificial neural networks for prediction of learning performances. *2016 12th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*, 745–751. https://doi.org/10.1109/FSKD.2016.7603268

Estai, M., & Bunt, S. (2016). Best teaching practices in anatomy education: A critical review. *Annals of Anatomy - Anatomischer Anzeiger, 208*, 151–157. https://doi.org/10.1016/j.aanat.2016.02.010

Fajrianti, E. D., Sukaridhoto, S., Rasyid, M. U. H. A., Suwito, B. E., Budiarti, R. P. N., Hafidz, I. A. A., Satrio, N. A., & Haz, A. L. (2022). Application of Augmented Intelligence Technology with Human Body Tracking for Human Anatomy Education. *IJIET: International Journal of Information and Education Technology*, *12*(6), Article 6.

Fang, Z., Xu, Z., He, X., & Han, W. (2022).Artificial intelligence-based pathologic myopia identification system in the ophthalmology residency training program. *Frontiers in Cell and Developmental Biology*, *10*. https://doi.org/10.3389/fcell.2022.1053079

Fazlollahi, A. M., Bakhaidar, M., Alsayegh, A., Yilmaz, R., Winkler-Schwartz, A., Mirchi, N., Langleben, I., Ledwos, N., Sabbagh, A. J., Bajunaid, K., Harley, J. M., & Del Maestro, R. F. (2022). Effect of Artificial Intelligence Tutoring vs Expert Instruction on Learning Simulated Surgical Skills Among Medical Students: A Randomized Clinical Trial. *JAMA Network Open, 5*(2), e2149008. https://doi.org/10.1001/jamanetworkopen.2021.49008

Fernández-Alemán, J. L., López-González, L., González-Sequeros, O., Jayne, C., López-Jiménez, J. J., & Toval, A. (2016). The evaluation of i-SIDRA – a tool for intelligent feedback – in a course on the anatomy of the locomotor system. *International Journal of Medical Informatics, 94*, 172–181. https://doi.org/10.1016/j.ijmedinf.2016.07.008

Foss, C. L. (1987). Learning *from errors in ALGEBRALAND*. Institute for Research on Learning.

Furlan, R., Gatti, M., Menè, R., Shiffer, D., Marchiori, C., Levra, A. G., Saturnino, V., Brunetta, E., & Dipaola, F. (2021). A Natural Language Processing-Based Virtual Patient Simulator and Intelligent Tutoring System for the Clinical Diagnostic Process: Simulator Development and Case Study. *JMIR Medical Informatics, 9*(4), e24073. https://doi.org/10.2196/24073

Gendia, A. (2022). Cloud Based AI-Driven Video Analytics (CAVs) in Laparoscopic Surgery: A Step Closer to a Virtual Portfolio. *Cureus*, *14*(9). https://doi.org/10.7759/cureus.29087

Gil, D. H., Heins, M., & Jones, P. B. (1984). Perceptions of medical school faculty members and students on clinical clerkship feedback. *Academic Medicine, 59*(11), 856.

Goh, P. S. (2021). The vision of transformation in medical education after the COVID-19 pandemic. *Korean Journal of Medical Education, 33*(3), 171–174. https://doi.org/10.3946/kjme.2021.197

Gorospe-Sarasúa, L., Munoz-Olmedo, J. M., Sendra-Portero, F., & de Luis-García, R. (2022a). *Challenges of* Radiology *education in the era of artificial intelligence*. 6.

Gorospe-Sarasúa, L., Muñoz-Olmedo, J. M., Sendra-Portero, F., & de Luis-García, R. (2022b). Challenges of Radiology education in the era of artificial intelligence. *Radiología (english Edition), 64*(1), 54–59. https://doi.org/10.1016/j.rxeng.2020.10.012

Han, R., Yu, W., Chen, H., & Chen, Y. (2022). Using artificial intelligence reading label system in diabetic retinopathy grading training of junior ophthalmology residents and medical students. *BMC Medical Education*, *22*(1), Article 1. https://doi.org/10.1186/s12909-022-03272-3

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of educational research, 77*(1), 81–112. https://doi.org/10.3102/003465430298487

Hedderich, D. M., Keicher, M., Wiestler, B., Gruber, M. J., Burwinkel, H., Hinterwimmer, F., Czempiel, T., Spiro, J. E., Pinto dos Santos, D., Heim, D., Zimmer, C., Rückert, D., Kirschke, J. S., & Navab, N. (2021). AI for Doctors—A Course to Educate Medical Professionals in Artificial Intelligence for Medical Imaging. *Healthcare*, *9*(10), Article 10. https://doi.org/10.3390/healthcare9101278

Hewson, M. G., & Little, M. L. (1998). Giving Feedback in Medical Education. *Journal of General Internal Medicine, 13*(2), 111–116. https://doi.org/10.1046/j.1525-1497.1998.00027.x

Hisan, U. K., & Amri, M. M. (2023). ChatGPT and Medical Education: A Double-Edged Sword. *Journal of Pedagogy and Education Science*, *2*(01), Article 01. https://doi.org/10.56741/jpes.v2i01.302

Hisey, R., Camire, D., Erb, J., Howes, D., Fichtinger, G., & Ungi, T. (2022). System for Central Venous Catheterization Training Using Computer Vision-Based Workflow Feedback. *IEEE Transactions on Biomedical Engineering, 69*(5), 1630–1638. https://doi.org/10.1109/TBME.2021.3124422

Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L. H., & Aerts, H. J. W. L. (2018). Artificial intelligence in radiology. *Nature Reviews Cancer*, *18*(8), Article 8. https://doi.org/10.1038/s41568-018-0016-5

Hu, H., Li, J., Lei, X., Qin, P., & Chen, Q. (2019). Design of health statistics intelligent education system based on Internet +. *Journal of Physics: Conference Series, 1168*(6), 062003. https://doi.org/10.1088/1742-6596/1168/6/062003

Hwang, G.-J., Xie, H., Wah, B. W., & Gašević, D. (2020). Vision, challenges, roles and research issues of Artificial Intelligence in Education. *Computers and Education: Artificial Intelligence, 1*, 100001. https://doi.org/10.1016/j.caeai.2020.100001

Islam, G., Kahol, K., Li, B., Smith, M., & Patel, V. L. (2016). Affordable, web-based surgical skill training and evaluation tool. *Journal of Biomedical Informatics, 59*, 102–114. https://doi.org/10.1016/j.jbi.2015.11.002

Karambakhsh, A., Kamel, A., Sheng, B., Li, P., Yang, P., & Feng, D. D. (2019). Deep gesture interaction for augmented anatomy learning. *International Journal of Information Management, 45*, 328–336. https://doi.org/10.1016/j.ijinfomgt.2018.03.004

Kirubarajan, A., Young, D., Khan, S., Crasto, N., Sobel, M., & Sussman, D. (2022). Artificial Intelligence and Surgical Education: A Systematic Scoping Review of Interventions. *Journal of Surgical Education, 79*(2), 500–515. https://doi.org/10.1016/j.jsurg.2021.09.012

Klar, R., & Bayer, U. (1990). Computer-assisted teaching and learning in medicine. *International Journal of Bio-Medical Computing, 26*(1–2), 7–27. https://doi.org/10.1016/0020-7101(90)90016-N

Kurniawan, M. H., Suharjito, Diana, & Witjaksono, G. (2018). *Human Anatomy Learning Systems Using Augmented Reality on Mobile Application*. *135*, 80–88. https://doi.org/10.1016/j.procs.2018.08.152

Lam, A., Lam, L., Blacketer, C., Parnis, R., Franke, K., Wagner, M., Wang, D., Tan, Y., Oakden-Rayner, L., Gallagher, S., Perry, S. W., Licinio, J., Symonds, I., Thomas, J., Duggan, P., & Bacchi, S. (2022a). Professionalism and clinical short answer question marking with machine learning. *Internal Medicine Journal, 52*(7), 1268–1271. https://doi.org/10.1111/imj.15839

Lam, K., Chen, J., Wang, Z., Iqbal, F. M., Darzi, A., Lo, B., Purkayastha, S., & Kinross, J. M. (2022b). Machine learning for technical skill assessment in surgery: A systematic review. *Npj* Digital *Medicine*, 5(1), Article 1. https://doi.org/10.1038/s41746-022-00566-0

Lazarus, M. D., Truong, M., Douglas, P., & Selwyn, N. (2022). Artificial intelligence and clinical anatomical education: Promises and perils. *Anatomical Sciences Education*, *n/a*(n/a). https://doi.org/10.1002/ase.2221

Lee, J., Wu, A. S., Li, D., Kulasegaram, K., & (Mahan). (2021). Artificial Intelligence in Undergraduate Medical Education: A Scoping Review. *Academic Medicine, 96*(11S), S62. https://doi.org/10.1097/ACM.0000000000004291

Lee, L. S., Aluwee, S. A. Z. S., Meng, G. C., Palanisamy, P., & Subramaniam, R. (2020). Interactive Tool Using Augmented Reality (AR) for Learning Knee and Foot Anatomy Based on CT Images 3D Reconstruction. *2020 International Conference on Computational Intelligence (ICCI)*, 281–286. https://doi.org/10.1109/ICCI51257.2020.9247820

Li, Y., Bai, C., & Reddy, C. K. (2016). A distributed ensemble approach for mining healthcare data under privacy constraints. *Information Sciences, 330*, 245–259. https://doi.org/10.1016/j.ins.2015.10.011

Li, Y. S., Lam, C. S. N., & See, C. (2021). Using a Machine Learning Architecture to Create an AI-Powered Chatbot for Anatomy Education. *Medical Science Educator, 31*(6), 1729–1730. https://doi.org/10.1007/s40670-021-01405-9

Luan, H., Geczy, P., Lai, H., Gobert, J., Yang, S. J. H., Ogata, H., Baltes, J., Guerra, R., Li, P., & Tsai, C.-C. (2020). Challenges and Future Directions of Big Data and Artificial Intelligence in Education. *Frontiers in Psychology*, *11*. https://doi.org/10.3389/fpsyg.2020.580820

Martin, J. A., Regehr, G., Reznick, R., Macrae, H., Murnaghan, J., Hutchison, C., & Brown, M. (1997). Objective structured assessment of technical skill (OSATS) for surgical residents. *BJS (british Journal of Surgery), 84*(2), 273–278. https://doi.org/10.1046/j.1365-2168.1997.02502.x

Mirchi, N., Bissonnette, V., Yilmaz, R., Ledwos, N., Winkler-Schwartz, A., & Del Maestro, R. F. (2020). The Virtual Operative Assistant: An explainable artificial intelligence tool for simulation-based training in surgery and medicine. *Plos One, 15*(2), e0229596. https://doi.org/10.1371/journal.pone.0229596

Muller, S. (1984). Physicians for the twenty-first century: Report of the project panel on the general professional education of the physician and college preparation for medicine. *Journal of Medical Education, 59*, 1–208.

Nagaraj, M. B., Namazi, B., Sankaranarayanan, G., & Scott, D. J. (2023). Developing artificial intelligence models for medical student suturing and knot-tying video-based assessment and coaching. *Surgical Endoscopy, 37*(1), 402–411. https://doi.org/10.1007/s00464-022-09509-y

Nakawala, H., Ferrigno, G., & De Momi, E. (2018). Development of an intelligent surgical training system for Thoracentesis. *Artificial Intelligence in Medicine, 84*, 50–63. https://doi.org/10.1016/j.artmed.2017.10.004

Neves, S. E., Chen, M. J., Ku, C. M., Karan, S., DiLorenzo, A. N., Schell, R. M., Lee, D. E., Diachun, C. A. B., Jones, S. B., & Mitchell, J. D. (2021). Using Machine Learning to Evaluate Attending Feedback on Resident Performance. *Anesthesia & Analgesia, 132*(2), 545–555. https://doi.org/10.1213/ANE.0000000000005265

Niitsu, H., Hirabayashi, N., Yoshimitsu, M., Mimura, T., Taomoto, J., Sugiyama, Y., Murakami, S., Saeki, S., Mukaida, H., & Takiyama, W. (2013). Using the Objective Structured Assessment of Technical Skills (OSATS) global rating scale to evaluate the skills of surgical trainees in the operating room. *Surgery Today, 43*(3), 271–275. https://doi.org/10.1007/s00595-012-0313-7

Ötles, E., Kendrick, D. E., Solano, Q. P., Schuller, M., Ahle, S. L., Eskender, M. H., Carnes, E., & George, B. C. (2021). Using Natural Language Processing to Automatically Assess Feedback Quality: Findings From 3 Surgical Residencies. *Academic Medicine, 96*(10), 1457. https://doi.org/10.1097/ACM.0000000000004153

Ouyang, F., Zheng, L., & Jiao, P. (2022). Artificial intelligence in online higher education: A systematic review of empirical research from 2011 to 2020. *Education and Information Technologies*. https://doi.org/10.1007/s10639-022-10925-9

Peter, H., & Goodridge, W. (2004). Integrating Two Artificial Intelligence Theories in a Medical Diagnosis Application. In M. Bramer & V. Devedzic (Ed.), *Artificial Intelligence Applications and Innovations* (pp. 11–23). Springer US. https://doi.org/10.1007/1-4020-8151-0_2

Qian, X., Jingying, H., Xian, S., Yuqing, Z., Lili, W., Baorui, C., Wei, G., Yefeng, Z., Qiang, Z., Chunyan, C., Cheng, B., Kai, M., & Yi, Q. (2022). The effectiveness of artificial intelligence-based automated grading and training system in education of manual detection of diabetic retinopathy. *Frontiers in Public Health, 10*, 1025271. https://doi.org/10.3389/fpubh.2022.1025271

Razzak, M. I., Naz, S., & Zaib, A. (2018). Deep Learning for Medical Image Processing: Overview, Challenges and the Future. In N. Dey, A. S. Ashour, & S. Borra (Ed.), *Classification in BioApps: Automation of Decision Making* (pp. 323–350). Springer International Publishing. https://doi.org/10.1007/978-3-319-65981-7_12

Sadeghi Esfahlani, S., Izsof, V., Minter, S., Kordzadeh, A., Shirvani, H., & Esfahlani, K. S. (2020). Development of an Interactive Virtual Reality for Medical Skills Training Supervised by Artificial Neural Network. In Y. Bi, R. Bhatia, & S. Kapoor (Ed.), *Intelligent Systems and Applications* (pp. 473–482). Springer International Publishing. https://doi.org/10.1007/978-3-030-29513-4_34

Shiang, T., Garwood, E., & Debenedectis, C. M. (2022). Artificial intelligence-based decision support system (AI-DSS) implementation in radiology residency: Introducing residents to AI in the clinical setting. *Clinical Imaging, 92*, 32–37. https://doi.org/10.1016/j.clinimag.2022.09.003

Siyar, S., Azarnoush, H., Rashidi, S., Winkler-Schwartz, A., Bissonnette, V., Ponnudurai, N., & Del Maestro, R. F. (2020). Machine learning distinguishes neurosurgical skill levels in a virtual reality tumor resection task. *Medical and Biological Engineering and Computing, 58*(6), 1357–1367. https://doi.org/10.1007/s11517-020-02155-3

Solano, Q. P., Hayward, L., Chopra, Z., Quanstrom, K., Kendrick, D., Abbott, K. L., Kunzmann, M., Ahle, S., Schuller, M., Ötleş, E., & George, B. C. (2021). Natural Language Processing and Assessment of Resident Feedback Quality. *Journal of Surgical Education, 78*(6), e72–e77. https://doi.org/10.1016/j.jsurg.2021.05.012

Stephens, G. C., Rees, C. E., & Lazarus, M. D. (2021). Exploring the impact of education on preclinical medical students' tolerance of uncertainty: A qualitative longitudinal study. *Advances in Health Sciences Education, 26*(1), 53–77. https://doi.org/10.1007/s10459-020-09971-0

Sqalli, M. T., Al-Thani, D., Elshazly, M. B., & Al-Hijji, M. (2022). A Blueprint for an AI & AR-Based Eye Tracking System to Train Cardiology Professionals Better Interpret Electrocardiograms. In N. Baghaei, J. Vassileva, R. Ali, & K. Oyibo (Ed.), *Persuasive Technology* (pp. 221–229). Springer International Publishing. https://doi.org/10.1007/978-3-030-98438-0_17

Szasz, P., Louridas, M., Harris, K. A., Aggarwal, R., & Grantcharov, T. P. (2015). Assessing Technical Competence in Surgical Trainees: A Systematic Review. *Annals of Surgery, 261*(6), 1046. https://doi.org/10.1097/SLA.0000000000000866

Torre, D. M., Sebastian, J. L., & Simpson, D. E. (2003). Learning Activities and High-Quality Teaching: Perceptions of Third-Year IM Clerkship Students. *Academic Medicine, 78*(8), 812.

Vayena, E., & Blasimme, A. (2017). Biomedical Big Data: New Models of Control Over Access, Use and Governance. *Journal of Bioethical Inquiry, 14*(4), 501–513. https://doi.org/10.1007/s11673-017-9809-6

Voss, G., Bockholt, U., Los Arcos, J. L., Müller, W., Oppelt, P., & Stähler, J. (2000). Lahystotrain. *Studies in Health Technology and Informatics, 70*, 359–364. https://doi.org/10.3233/978-1-60750-914-1-359

Wang, M., Sun, Z., Jia, M., Wang, Y., Wang, H., Zhu, X., Chen, L., & Ji, H. (2022). Intelligent virtual case learning system based on real medical records and natural language processing. *BMC Medical Informatics and Decision Making, 22*(1), 60. https://doi.org/10.1186/s12911-022-01797-7

Willan, P. L. T., & Humpherson, J. R. (1999). Concepts of variation and normality in morphology: Important issues at risk of neglect in modern undergraduate medical courses. *Clinical Anatomy, 12*(3), 186–190. https://doi.org/10.1002/(SICI)1098-2353(1999)12:3%3c186::AID-CA7%3e3.0.CO;2-6

Wolverton, S. E., & Bosworth, M. F. (1985). A survey of resident perceptions of effective teaching behaviors. *Family Medicine, 17*(3), 106–108.

Wood, B. P. (2000). Feedback: A Key Feature of Medical Training. *Radiology, 215*(1), 17–19. https://doi.org/10.1148/radiology.215.1.r00ap5917

Yang, W., Hebert, D., Kim, S., & Kang, B. (2019). MCRDR Knowledge-Based 3D Dialogue Simulation in Clinical Training and Assessment. *Journal of Medical Systems, 43*(7), 200. https://doi.org/10.1007/s10916-019-1262-0

Yilmaz, Y., Nunez, A. J., Ariaeinejad, A., Lee, M., Sherbino, J., & Chan, T. M. (2022). Harnessing Natural Language Processing to Support Decisions Around Workplace-Based Assessment: Machine Learning Study of Competency-Based Medical Education. *JMIR Medical Education, 8*(2), e30537. https://doi.org/10.2196/30537

Yilmaz-Recai, Winkler-Schwartz, A., Mirchi, N., Reich, A., Christie, S., Tran, D. H., Ledwos, N., Fazlollahi, A. M., Santaguida, C., Sabbagh, A. J., Bajunaid, K., & Del Maestro, R. (2022). Continuous monitoring of surgical bimanual expertise using deep neural networks in virtual reality simulation. *Npj Digital Medicine*, *5*(1), Article 1. https://doi.org/10.1038/s41746-022-00596-8

Zahiri, M., Booton, R., Siu, K.-C., & Nelson, C. A. (2016). Design and Evaluation of a Portable Laparoscopic Training System Using Virtual Reality. *Journal of Medical Devices*, 11(1). https://doi.org/10.1115/1.4034881

Zhao, S., Zhang, X., Jin, F., & Hahn, J. (2021). An Auxiliary Tasks Based Framework for Automated Medical Skill Assessment with Limited Data. *2021 43rd Annual International* Conference *of the IEEE Engineering in Medicine Biology Society (EMBC)*, 1613–1617. https://doi.org/10.1109/EMBC46164.2021.9630498

Authors and Affiliations

Wei Zhang[1] · Mingxuan Cai[1] · Hong Joo Lee[2] · Richard Evans[3] · Chengyan Zhu[4] · Chenghan Ming[5]

✉ Chengyan Zhu
zhuchengyan0323@163.com

Wei Zhang
weizhanghust@hust.edu.cn

Mingxuan Cai
caimingxuan0104@163.com

Hong Joo Lee
fastbat@gmail.com

Richard Evans
R.Evans@dal.ca

Chenghan Ming
mingchenghan@163.com

1 School of Medicine and Health Management, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, China

2 Department of Business Administration, The Catholic University of Korea, Seoul, Korea

3 Faulty of Computer Science, Dalhousie University, Halifax, Canada

4 School of Political Science and Public Administration, Wuhan University, Wuhan, China

5 College of Public Administration and Law, Hunan Agricultural University, Changsha, China