SN Computer Science (2025) 6:943
https://doi.org/10.1007/s42979-025-04495-z

ORIGINAL RESEARCH

Evaluating eXtended Reality (XR) and Desktop Modalities for AI Education

Miguel A. Feijoo-Garcia[1] · Yiqun Zhang[1] · Yiyin Gu[2] · Alejandra J. Magana[1] · Bedrich Benes[2] · Voicu Popescu[2]

Received: 16 July 2025 / Accepted: 22 October 2025

Abstract
This paper is an extended version of the best paper from the HUCAPP 2025 conference. The abstract nature of Artificial Intelligence (AI) concepts presents an educational challenge. This paper compares an immersive extended reality (XR) environment to a traditional desktop setup for teaching Neural Networks and Handwritten Digit Recognition. We analyzed differences in student engagement, user experience, and learning outcomes between the two modalities. In this comparative study, 56 participants learned about AI concepts using either an XR headset (Meta Quest 3) or a desktop computer. We collected data on usability, satisfaction, immersion, and likelihood to recommend using the System Usability Scale (SUS), User Satisfaction Questionnaire (USQ), Immersion Presence Questionnaire (IPQ), and Net Promoter Score (NPS) questionnaires. Learning outcomes were assessed via multiple-choice questions administered during the lesson. The XR group reported significantly higher engagement, immersion, satisfaction, and likelihood to recommend the system. However, this increased engagement did not translate to superior learning outcomes; performance on in-lesson questions was comparable between the two groups. XR users also identified challenges, including physical discomfort and unfamiliarity with the technology. The findings suggest that XR can increase student motivation in AI education, but this does not automatically lead to better learning performance. Usability challenges and the novelty of the technology may hinder knowledge absorption. Future work should focus on scaffolding strategies to mitigate these issues and better leverage XR's educational potential by personalizing the learning experience.

Keywords Extended reality · Educational technology · Immersive learning · User experience · Human-Computer interaction

✉ Alejandra J. Magana
admagana@purdue.edu

Miguel A. Feijoo-Garcia
mfeijoog@purdue.edu

Yiqun Zhang
zhan2897@purdue.edu

Yiyin Gu
gu251@purdue.edu

Bedrich Benes
bbenes@purdue.edu

Voicu Popescu
popescu@purdue.edu

[1] School of Applied and Creative Computing, Purdue University, 401 N. Grant St., West Lafayette, IN 47907, USA

[2] Department of Computer Science, Purdue University, 305 N. University St., West Lafayette, IN 47907, USA

Introduction

Artificial Intelligence (AI) can be abstract and difficult to understand for learners. Modern AI generally involves deep neural networks and natural language that have distinct particular complexities, characteristics, limitations, and principles [1]. Consequently, users cannot easily see or comprehend the internal operations of the algorithms, causing a gap between the technology and its users [2], adding to the difficulty in facilitating teaching and learning processes of systemic and critical thinking toward AI concepts from a holistic perspective [3]. Due to the abstract nature of AI, students may not feel prepared to interact with it, leading to a lack of confidence [4, 5]. Considering all these educational challenges when teaching and learning AI, it is necessary to think of new innovative educational methods to facilitate

Published online: 05 November 2025

learners, demystifying the complexity of AI and helping them engage with its concepts in a learner-centered way.

Technology is a driving force for better education [6, 7]. Emerging technologies such as extended reality (XR) have helped create learner-centered environments in education [8, 9]. XR combines computer software with wearable devices to produce interactive settings that blend real and digital components [10, 11]. Interactions can occur through immersive headsets, augmented reality (AR) glasses, or mobile devices [12]. XR includes AR, virtual reality (VR), and mixed reality (MR) [13, 14]. AR integrates real and digital objects, VR offers simulated environments, and MR facilitates interaction between physical and digital elements [15]. The affordability and efficiency of XR make it a practical and versatile tool for education, further demonstrating its potential in education [11, 13]. XR applications span various fields in education, including STEM, architecture, management, medicine, and art [7, 11, 13, 16–18]. As a tool that promotes educational sustainability, XR has gained worldwide interest in research and practice [6].

Using XR in education is beneficial as it can foster soft skills such as teamwork, problem-solving, and technical skills [10, 19]. More benefits include enhanced content sharing and knowledge acquisition [20], personalized learning experiences [21], and better support for students with special needs [22]. However, its implementation has limitations, including the lack of motivation or technical ability among instructors, cybersickness, and the high costs of purchasing virtual learning platforms [11, 23].

As XR technology evolves, educators are encouraged to adopt it to create engaging and effective learning experiences [19]. Immersive technologies like VR are increasingly utilized in computer science education, which involves teaching the fundamentals of abstract concepts in STEM fields [11]. Nevertheless, the teaching process must be incremental, integrating new topics while reinforcing prior knowledge and involving new technologies as instructional resources.

Different questionnaires are related to the effectiveness and user experience of educational tools. System Usability Scale (SUS) [24] is often used to evaluate user experience and to identify the strengths and areas to improve [25]. Similarly, the User Satisfaction Questionnaire (USQ) emphasizes how usable it is and what the user's experience is [26]. The Immersion Presence Questionnaire (IPQ) [27] is used to determine the immersive quality of virtual environments, and the Net Promoter Score (NPS) [28] provides us with evidence of student advocacy and satisfaction. Overall, the results of these assessments help optimize immersive technologies to enhance engagement and improve learning experiences and outcomes in computer science education and related fields [29].

Immersive technologies and simulation-based experiences can be incorporated to increase student engagement and the understanding of complex concepts. This stems further to interactive virtual experience encouraging active participation and practice, resulting in better learning outcomes [11]. That is, evidence-based decision-making is important for improving design and addressing challenges [30], ensuring decisions are grounded in facts rather than assumptions [31]. Therefore, the following question arises: *What is the effectiveness of XR environments in enhancing student engagement and user satisfaction compared to traditional desktop learning setups?* This paper is an extended version of the best paper award from the HUCAPP 2025 conference [32].

Background

Learning approaches in science, technology, engineering, and mathematics (STEM) can be complex due to the abstract nature of its concepts, such as those in AI education. Therefore, among different learning approaches, embodied learning suggests that people learn better and retain knowledge when they use actions, gestures, or when they have the opportunity to interact with the learning materials or the environment to examine abstract concepts that are being studied [33, 34]. In fact, this approach helps students in STEM education by connecting their physical activity with cognitive skills, for instance, by using their hands to interact with 3D objects or moving in space, which can make abstract relationships more straightforward to understand [35, 36]. Such forms of interaction with the environment are associated with theories of embodied cognition.

Theoretical Framework

This study is grounded in *Embodied Cognition*, a theory that explains learning in the form of learners using their bodies to act, move, or gesture, which promote their thinking and learning processes, based on the idea that physical actions help them connect abstract concepts to concrete experiences [35, 36]. Accordingly, the primary assumption of embodied cognition is that knowledge partially relies on neural mechanisms pertaining to perceptual, sensory, and motoric processes [37]. This assumption has important implications for learning abstract concepts in STEM. Specifically, embodied cognition maintains that brain regions in the sensorimotor and nearby association cortex play a prominent role in information processing and retrieval [38]. These same brain regions are also responsive to information within a specific sensory modality responsible for representing the properties of a given object [38]. Thus, concept representations are considered multimodal and rely on networks distributed throughout the cortex. When such representations are reconstructed, they use features (e.g., shape, texture, sound) and modalities (e.g., visual, tactile, phonological). Concepts are organized based on properties, and properties are dependent on specific sensory modalities [39]. That is, abstract concepts are (a) comprehended based on embodied visuospatial representations [40] and (b) understood through analogical reasoning processes involving relational properties emerging from the interaction of two or more objects or agents in a given circumstance [41].

Due to the abstract nature of the concepts in STEM, embodied learning can support learning. For instance, in one study the students learned science topics better when they used gestures and movement in a mixed-reality environment compared to traditional instruction, identifying the amount and type of body movement (i.e., motor activity), how well the gestures match the content (i.e., gestural congruency), and how present and involved students feel in the environment (i.e., immersion) as the three main aspects of embodiment that improve learning, which provides a clear framework for designing and evaluating learning systems [42]. Although the literature like this suggests the benefits of XR for usability and assessment in educational settings, there is a gap in research that examines how different degrees of embodiment (e.g., XR versus desktop-based settings) influence student engagement and, consequently, learning outcomes in AI education [42].

Student engagement is critical, as a significant body of literature has established a robust link between student involvement in educationally purposive activities and positive student learning outcomes [43]. And of relevance for this study, engagement is closely tied to user experience in learning environments because the quality of the user experience helps set the conditions under which students feel motivated, supported, and able to participate deeply. A good learning experience means clear, accessible interfaces, intuitive navigation, timely feedback, interactivity, relevance of content, and scaffolding; all these contribute to learners' perceptions of usefulness, ease, and pleasure. When user experience is positive, students perceive the environment as supportive and meaningful, which boosts their self-efficacy, interest, and attention, and in turn leads to higher engagement. Conversely, poor user experience perceived as difficult navigation, confusing layout, lack of feedback, or relevance, can increase cognitive load, frustration, and disengagement [44]. For example, a study in the context of heritage education in Macau, identified that embodied cognition and immersive experience significantly positively influence learning satisfaction, a user experience metric, and learning effectiveness, confirming that sensory and bodily involvement are part of what makes for a good user/learner experience [45]. Consequently, we argue that positive user experience is a precursor for productive engagement that can foster deeper learning.

Extended Reality for Learning in STEM

Extended reality (XR), which includes other immersive technologies such as virtual reality (VR), augmented reality (AR), and mixed reality (MR), provides to the learners with the exploration and manipulation of objects in a virtual environment, while maintaining the awareness of the real world [22, 36]. In fact, different studies have suggested that XR can increase student motivation, engagement, and presence when learning content in the STEM context, which due to its abstract nature makes it difficult to visualize in traditional classrooms [11, 16]. Hence, this immersive experience supports embodied learning because it allows students to connect their physical actions with the abstract concepts being examined. However, traditional learning approaches, such as those involving desktop-based environments, continue to be widely used in higher education because of their overall accessibility, learners' familiarity, and affordability. The literature suggests that while desktop-based learning provides clear and stable visualizations, it does not offer the same level of immersion and embodied interaction as immersive technologies, such as XR, enable [6, 23].

To understand the differences between environments, it is important to consider how embodied learning connects with the ways students experience them. To address this, the literature has reported different metrics that support assessing user experience from different perspectives. In fact, *system usability* measures whether the students can focus on the content without being distracted by possible technical issues or by the novelty of the technology, while *user satisfaction* measures the quality of the learning experience and how it can influence overall motivation and the willingness to continue learning using the particular learning approach or environment [24, 26]. In addition, *immersion* refers to how much students feel present and can connect sensory and motor activity with the abstract concepts being studied, while the *likelihood of recommendation* measures whether students accept the system overall and find it valuable for their learning [27, 28, 46]. Together, these metrics support the analysis of user experience by focusing on how focused, motivated, and immersed students feel in the learning environment.

Thus, it is important to examine the comparison between different modalities, from traditional to immersive learning, to better understand the support or limitations that student learning may have on either of them. There is a gap in the literature regarding the relationship between the principles of embodied learning and the metrics for assessing user experience, including usability, satisfaction, immersion, and likelihood of recommendation, as well as how these factors together influence overall engagement and learning in STEM contexts. Hence, this manuscript contributes to the literature by comparing XR (i.e., immersive learning) and desktop environments (i.e., desktop-based learning) in the context of learning about neural networks in AI, thereby extending the discussion of how XR technologies can support embodied cognition and how they compare with traditional approaches, such as using desktop-based learning, considering different metrics to evaluate overall user satisfaction and overall short-term learning outcomes.

Implications of Embodied Learning for the Study

Building on our theoretical lens of embodied cognition, this study describes the two learning settings (i.e., XR and desktop) as different levels of embodiment, thereby extending embodied cognition theory to AI education. Particularly, the XR environment provides high embodiment because it allows students to interact with 3D neural network layers using controllers (i.e., hand-held) and panels that combine real and virtual elements, which was chosen to connect the motor actions directly to abstract AI concepts such as Neural Networks. In contrast, the desktop environment offers the same content as the XR environment but with low embodiment, as the students use only a mouse and keyboard and view content on a flat screen. Hence, by framing these two modalities in this way, we are not only testing two tools but also comparing two levels of embodiment. That is, it enables us to assess whether higher embodi- ment leads to greater or lower engagement, stronger presence, higher satisfaction, and improved short-term learning outcomes. Moreover, it also explains why we measure user experience through metrics such as usability, satisfaction, immersion, and recommendation alongside learning outcomes, by considering that each metric corresponds to an element of embodiment. In fact, *usability* shows how easy it is to carry out motor actions, *user satisfaction* shows the learner's feelings about the embodied interaction, *immersion* shows how deeply the learner is involved, and the *likelihood of recommendation* shows whether the willingness of the learners to repeat or share the embodied experience for learning purposes.

Methods

The ACHIEVE System

The *ACHIEVE* system is an eXtended Reality (XR) educational tool for Artificial Intelligence (AI) concepts. It was first developed for the Meta Quest Pro headset but later extended to the Meta Quest 3, which offers full-color video passthrough and a more affordable price [47]. The software for the XR environment was built using Unity 3D (version 2022.3.5f1), combined with Meta's XR All-in-one SDK, and integrated the Barracuda framework [48], which allows the headset to load and run pre-trained neural networks, such as the MNIST digit classifier [49], directly on the device.

The user-defined scenes include several carefully calibrated scenarios showing Neural Networks (NN) and their function in 3D. Particularly, the *input*, *hidden*, and *output* layers appear as interactive panels arranged in a circle on the participant's real desk. With video passthrough, these virtual panels blend with the real world. Participants interact with the system using handheld controllers that work like laser pointers. With them, they can select parts of the NN to view more information (e.g., neuron activation values) and also draw new handwritten digits, seeing how the NN processes the input in real-time. This allows them to observe how the inputs activate the network and how the final classification is made (see Fig. 1).

The system uses Open Neural Network Exchange (ONNX) files to load the structure and weights of the NN, which is compatible with many AI development tools [50].

Since the headset has limited processing power, the system can send the heavy processing to a nearby computer, such as a laptop, using a wireless connection. Training the NN requires a significant amount of computing resources, so it is done separately and offline before using the system. Moreover, different versions of trained models are saved ahead of time, and users can load these versions to explore important AI concepts like untrained networks, underfitting, overfitting, and others.

The 3D visualization provided by the system offers parameterization for customization, allowing users to change the height of the panels, the size of the circle around them, and the angle of the layers, either facing the user or turning sideways. This allows a better view of either the layers or the connections for each learner. Although the interaction is mainly done using a virtual laser pointer, the system also leverages the active haptic feedback from the handheld controllers to guide user interaction or stabilize the selections made by the users. Additionally, since this XR system allows users to see the real world alongside virtual content, they can use a physical keyboard on their desk to input text into the system.

The choice to use XR instead of a fully immersive virtual reality (VR) environment for *ACHIEVE* was made after considering some pedagogical and practical benefits [51]. For example, XR provides natural view control and depth perception, which help users understand the 3D structure of a NN. In addition, XR encourages embodied learning by

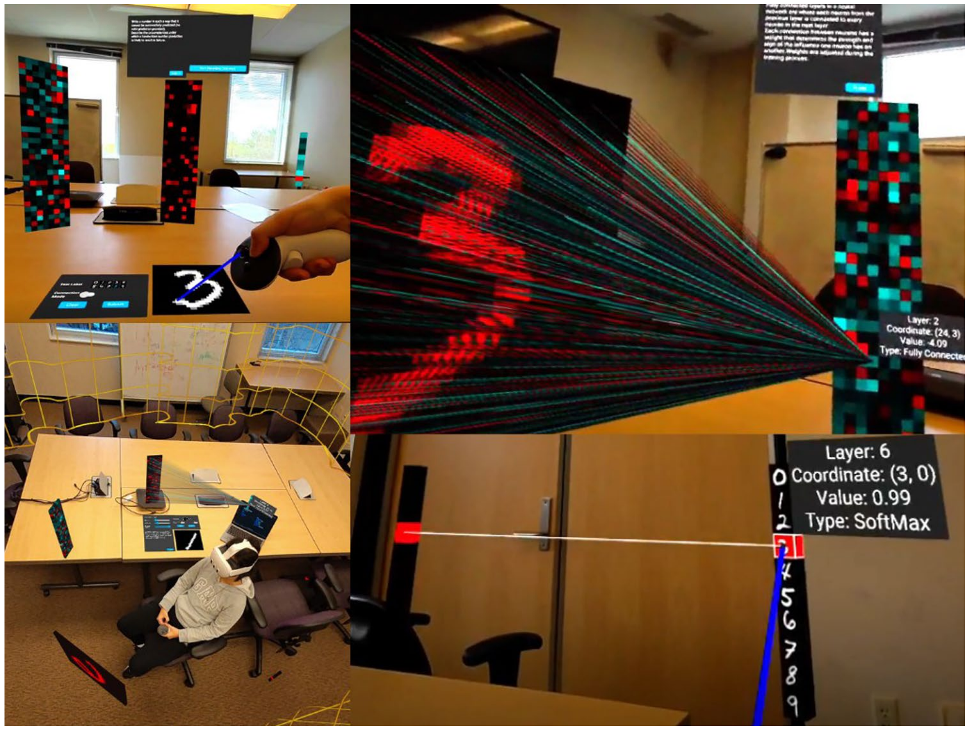

Fig. 1 Features of the ACHIEVE XR System

engaging the users' motor skills as the users interact with the NN. Contrary to fully immersive VR, XR allows users to remain aware of their physical surroundings and interact with instructors or peers, which is important in educational settings. It is important to note that fully immersive VR also provides depth perception and can support seated use. However, XR was chosen in this study because it combines immersion with awareness of the real surroundings, which increases safety and comfort in classroom contexts. In addition, XR reduces the likelihood of cybersickness, which has been reported to be a common limitation of VR in educational settings [11, 23].

Context and Participants

This study is based on the survey responses collected from $N = 56$ participants on how they experienced an educational approach that involved an XR environment or a traditional desktop setup to introduce Neural Networks (NN) and Handwritten Digit Recognition. The study was carried out between the spring and summer of 2024, using a questionnaire to collect demographic information and assess user experiences in VR and XR, including measuring familiarity with technologies such as NN and Handwritten Digit Recognition. In addition, another questionnaire was provided to participants that included different scales to assess their experiences in the XR and desktop environments, which were chosen because of their effectiveness in measuring user experience from different perspectives. These diverse lenses ensure that the overall analysis captures comprehensive user insights.

The age of the participants ranged from 18 to 32 years (M = 22.2, SD = 3.9) and were 67.9% male and 32.1% female. Many reported limited VR headset usage, with 14 (25.0%) having never used one and 17 (30.4%) using it only once. Similarly, 32 participants (57.1%) had never used XR, and 13 (23.2%) had used it once. Most of the participants were Computer Science majors (41, 73.2%), with the majority being undergraduate students (30, 53.6%).

In the XR group (29, 51.8%), most were male (18, 62.1%), undergraduate students (20, 68.9%), and in Computer Science (18, 62.1%). Many had used VR headsets only once (11, 37.9%), and many had never used XR (15, 51.7%). While 17 participants (58.6%) were familiar with Neural Networks (NN), 18 (62.1%) were unfamiliar with Handwritten Digit Recognition. In the Desktop group (27, 48.2%), on the other hand, there was also a male majority (20, 74.1%), with some being undergraduate students (10, 37.1%) and many studying Computer Science (23, 85.2%). Like the XR group, many had limited VR headset experience, with 8 participants (29.6%) using them fewer than five times and another 8 (29.6%) never using them. Familiarity with NN and Handwritten Digit Recognition was relatively balanced, with 13 participants (48.1%) familiar with Handwritten Digit Recognition and 14 (51.9%) with NN (refer to Table 1 reprinted from our previous conference paper [32]).

We conducted an a priori power analysis (two-tailed independent samples t-test, α = 0.05), and the required

Table 1 Participant demographics and characteristics - reprinted from our previous conference paper [32]

Variable	XR ($n = 29$)	Desktop ($n = 27$)	Total ($N = 56$)
Gender			
Male	18 (62.1%)	20 (74.1%)	38 (67.9%)
Female	11 (37.9%)	7 (25.9%)	18 (32.1%)
Other	0 (0%)	0 (0%)	0 (0%)
VR Headset Usage			
Frequently	2 (6.9%)	2 (7.4%)	4 (7.1%)
< 5 times	7 (24.1%)	8 (29.6%)	15 (26.8%)
>5 times	3 (10.3%)	3 (11.1%)	6 (10.7%)
Never	6 (20.7%)	8 (29.6%)	14 (25%)
Once	11 (37.9%)	6 (22.2%)	17 (30.4%)
XR Usage			
Frequently	1 (3.5%)	2 (7.4%)	3 (5.4%)
< 5 times	4 (13.8%)	1 (3.7%)	5 (8.9%)
>5 times	1 (3.5%)	2 (7.4%)	3 (5.4%)
Never	15 (51.7%)	17 (62.9%)	32 (57.1%)
Once	8 (27.6%)	5 (18.5%)	13 (23.2%)
Major			
Computer Science	18 (62.1%)	23 (85.2%)	41 (73.2%)
Data Science	3 (10.3%)	1 (3.7%)	4 (7.1%)
Computer Technology	0 (0%)	1 (3.7%)	1 (1.8%)
Other	8 (27.6%)	2 (7.4%)	10 (17.9%)
Role			
Undergraduate	20 (68.9%)	10 (37.1%)	30 (53.6%)
Graduate	9 (31.1%)	17 (62.9%)	26 (46.4%)
Familiarity with neural networks			
No	12 (41.4%)	13 (48.1%)	25 (44.6%)
Yes	17 (58.6%)	14 (51.9%)	31 (55.4%)
Familiarity with hand-written digit recognition			
No	18 (62.1%)	14 (51.9%)	32 (57.1%)
Yes	11 (37.9%)	13 (48.1%)	24 (42.9%)

Any discrepancies in percentages are due to rounding. Values represent frequencies with percentages in parentheses

sample size to achieve 80% power, which is the conventional standard in educational and HCI research [52, 53], was estimated assuming a medium effect size (Cohen's d = 0.50). This test was performed using the *TTestIndPower* function (from the *statsmodels* library) in Python [54], which calculates the necessary sample size for a two-sample t-test. The analysis suggests that approximately 32 participants per group ($N_{required} \approx 64$ in total) would be needed to reach adequate power [55]. For this study, the sample size of 56 participants (XR group $n = 29$; Desktop group $n = 27$), even if slightly lower than the suggested number of participants, is closely comparable to this conventional requirement and therefore adequate for identifying broader differences (i.e., medium effects) between groups [56, 57].

Data Collection Methods

The data collection methods for this study were as follows:

- **System Usability Scale (SUS)**: Evaluates user satisfaction with the XR experience, covering usage frequency, perceived complexity, ease of use, need for technical support, function integration, consistency, learning curve, and user confidence and comfort. The questionnaire comprises ten elements, rated on a scale [1, 5]. In this range, 1 means strong disagreement and 5 means strong agreement. The items are combined and changed into a total score from 0 to 100, where higher values show better usability. A score above 68 is usually seen as above average usability.
- **Immersion Presence Questionnaire (IPQ)**: Assesses participants' immersion and satisfaction in the XR environment, focusing on visual quality, interaction, realism, disconnection from surroundings, presence, and overall engagement. It includes fourteen elements, rated on a scale [− 3, 3]. In this range, − 3 means *strong disagreement*, 0 is *neutral*, and 3 means *strong agreement*. Thus, higher values show a stronger sense of immersion, presence, and engagement.

- **User Satisfaction Questionnaire (USQ)**: Measures user satisfaction in the XR environment, emphasizing usability, functionality, and overall experience through three elements, rated on a scale [1, 5]. In this range, 1 means very *dissatisfied* and 5 means *very satisfied*. Thus, higher values show greater satisfaction with the XR system.
- **Net Promoter Score (NPS)**: Assesses participants' likelihood or willingness to recommend the experience to others, reflecting the users' overall satisfaction and user advocacy regarding their experience with the system posed. This score is rated on a scale from [0, 10]. In this range, 0 means "*not at all likely* " and 10 means "*extremely likely*." Scores are grouped as: 0–6 = *detractors*, 7–8 = *passives*, and 9–10 = *promoters*. This metric is calculated as the percentage of promoters minus the percentage of detractors. Thus, higher values show stronger positive advocacy for the system.
- **Engagement**: Measures the time participants spent in the activity in both the Desktop and XR conditions. It was measured through the total session duration in seconds. Higher values indicate longer participation, which reflects greater engagement with the learning activity.
- **Learning**: Learning outcomes were measured in both the Desktop and XR conditions, with five pre-lesson multiple-choice questions to assess prior knowledge and six in-lesson multiple-choice questions completed by all participants to evaluate overall understanding during the task. Scores were calculated as the number of correct answers.

It is important to note that all standard instruments (i.e., SUS, IPQ, USQ, and NPS) come from previously validated work, and their original scales were kept to maintain scientific validity and ensure comparability with other studies. Keeping these original scales also supports internal consistency and helps to capture user experience from multiple perspectives. Table 2 presents a summary of the instruments used in this study to assess user experience.

Data Collection Procedures

Data collection sessions lasted up to one hour and involved no more than two participants at a time. Upon arrival, participants completed a consent form. Depending on the experimental condition, they were provided with either VR headsets (i.e., Meta Quest 3) or a laptop (see Fig. 2 reprinted from our previous conference paper [32]). All participants used tablets to respond to demographic, pre-questionnaire, and post-questionnaire, with confidentiality ensured through unique numeric identifiers assigned to each participant.

Table 2 Summary of instruments assessing user experience, learning, and engagement

Metric	No. items	Example	Scale	Source	Meaning
SUS	10	I would like to use this system frequently	1 equals *Strongly Disagree* to 5 equals *Strongly Agree*, converted to 0–100	[24]	Score 68 or higher indicates above-average usability
IPQ	14	Awareness of real-world surroundings in VR	-3 equals *Strongly Disagree*, 0 equals *Neutral*, 3 equals *Strongly Agree*	[27]	Higher scores indicate stronger presence and immersion
USQ	3	Overall satisfaction with the XR experience	1 equals *Very Dissatisfied* to 5 equals *Very Satisfied*	[26]	Higher scores indicate greater user satisfaction
NPS	1	Likelihood to recommend the experience	0 equals *Not at all likely* to 10 equals *Extremely likely*, NPS is equals to *Promoters* minus *Detractors*	[28]	Positive score indicates more promoters than detractors
Learning	11 (Pre-lesson: 5, In-lesson: 6)	Multiple-choice questions	Correct/Incorrect	Reported in study	Higher scores indicate greater knowledge or better learning outcomes (both Desktop and XR)
Engagement	1	Session duration	Seconds (continuous)	Reported in study	Longer duration indicates higher engagement (both Desktop and XR)

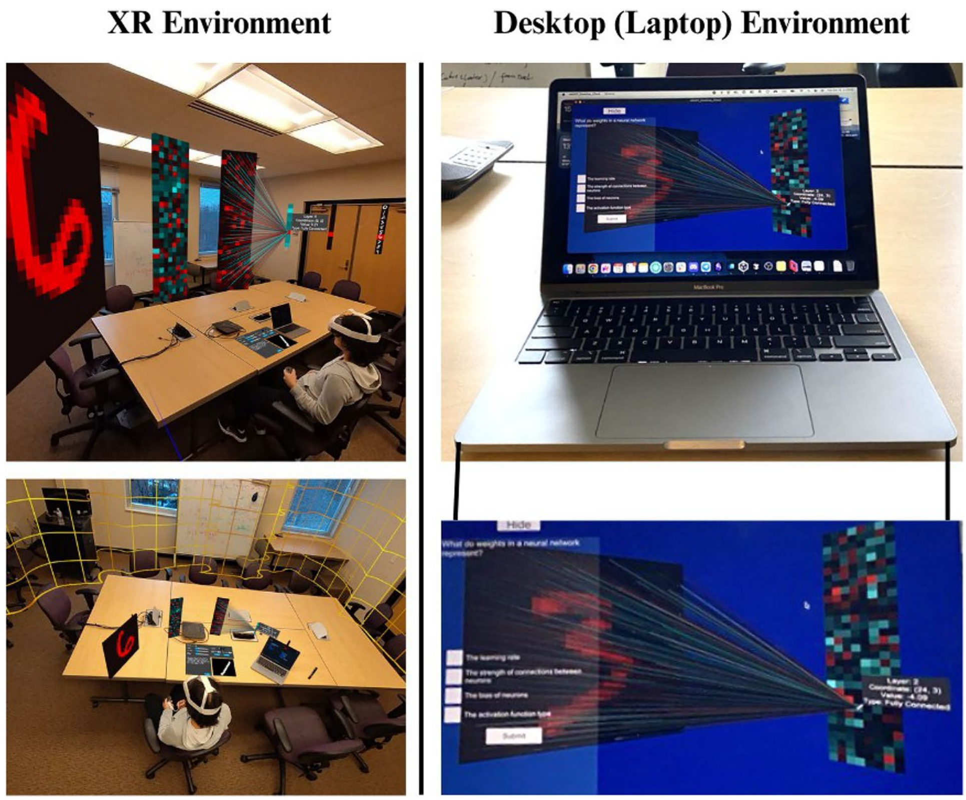

Fig. 2 XR and desktop/laptop environments used for the experimental approach - reprinted from our previous conference paper [32]

The steps in this study are as follows (refer to Fig. 3, reprinted from our previous conference paper [32]): (1) Participants take time to review and sign the consent form [S1]; (2) Participants use the tablets to answer the demographic questionnaire [S2]; (3) Participants use the provided tablets to answer the pre-questionnaire [S3]; (4) Six educational slides were provided to teach participants how to use the headset to complete the study [*VR headset users only*] [S4]; Participants wear the headset to learn the topic. [*VR headset users only*]; (5) Participants use the designated laptop to learn the topic [*Desktop users only*]; Participants respond during the intervention to 6 multiple choice questions [*VR headset and Desktop users*] [S5]; (6) Participants use tablets to answer the post-questionnaire [S6].

The research team guided the participants through each procedure, with the preparation phase lasting 15 min, the learning phase lasting roughly 15 min, and the final step lasting 5 to 10 min. The team was available to attend and answer all inquiries throughout the experimental session. The start times for the learning phase were staggered by 10 min for the two participants using VR headsets in the same session to minimize distractions during verbal response questions. The participants then completed a post-questionnaire with 26 items from the four metrics mentioned—the XR group responded to all the items of the four metrics. The desktop group responded to 14 questions, with three out of the four metrics, excluding the IPQ. Figure 3, reprinted from our previous conference paper [32], depicts the outline of the lesson design for this comparative study.

Data Analysis Methods

The data collected from the questionnaires (i.e., SUS, USQ, IPQ, and NPS) and the multiple-choice tests were analyzed using descriptive statistics. After scoring the students' responses from all metrics as shown in Table 2, we proceeded to apply statistical analyses, descriptive and inferential, to describe our findings in terms of central tendency and spread, as well as for comparing metrics between groups, correspondingly. For each measure, we calculated the mean, standard deviation, skewness, and kurtosis. These values described the general performance of the participants, the variation in their responses, and the distribution of scores. Moreover, the metrics for engagement and learning were also considered. While the participants' engagement was analyzed with the session time, which showed how long participants interacted with the system and how consistent their participation was, learning was analyzed with the pre-lesson knowledge questions and the in-lesson multiple-choice questions, providing insights about the participants' understanding before and during the lesson.

Furthermore, to compare the XR and Desktop groups, inferential statistical tests were applied. For this, independent

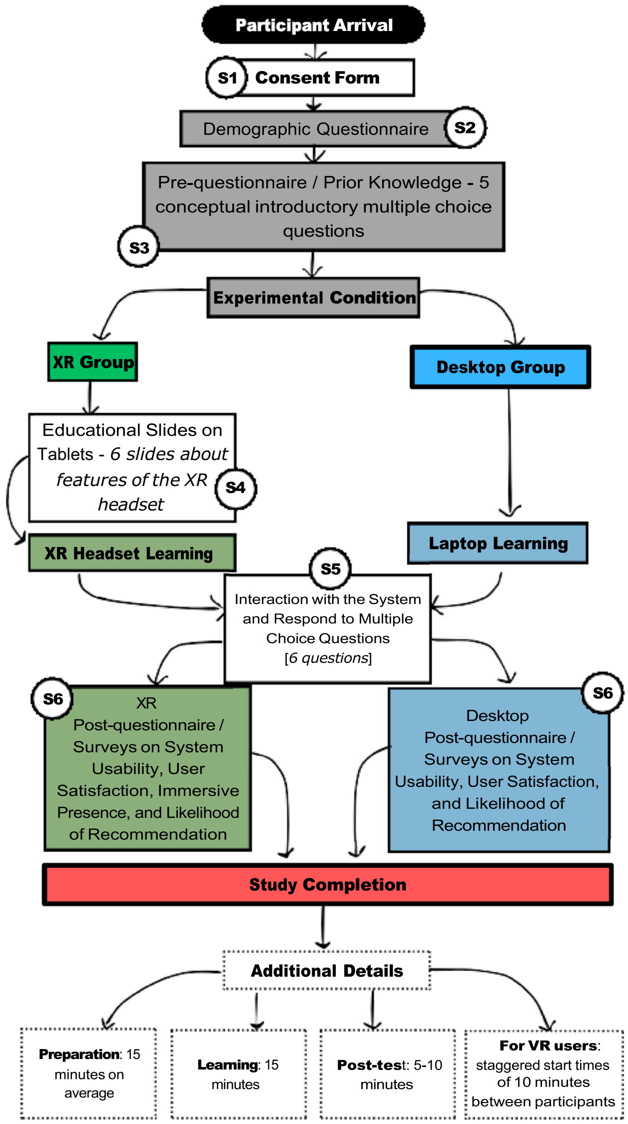

Fig. 3 Outline of the lesson design for the comparative study between XR and desktop environments - reprinted from our previous conference paper [32]

samples *t*-tests were conducted to examine group differences in prior knowledge, in-lesson performance, session duration, and user experience scores. When the assumption of equal variances was not met the Welch's *t*-test was used. In addition, effect sizes (i.e., Cohen's d) were reported to indicate the magnitude of the differences between groups.

Ethical Considerations

This study has been approved by Purdue University's Institutional Review Board under IRB-2024-57. A four-page consent form was designed to inform participants about the purpose, duration, confidentiality, benefits, risks, and other pertinent information of the study. If they chose to participate, participants first signed and dated the consent form. Participants could also withdraw at any time.

Results

Learning Gains

In this study, a lesson was designed on the foundational concepts of Neural Networks and Handwritten Digit Recognition. This study compared the performance of two groups of participants exposed to the same lesson in two settings (i.e., XR and Desktop environments), using the MNIST dataset trained on a fully connected network [58].

This section presents the lesson design and outlines the two distinct types of assessments employed: (1) an initial evaluation of participants' prior knowledge conducted before the lesson, and (2) subsequent assessments of their learning outcomes integrated during the lesson itself.

Pre-Lesson Assessment of Prior Knowledge

Participants began by answering five out of 14 multiple-choice questions assessing their prior knowledge of Neural Networks. These questions were administered before participants engaged with any lesson content, serving as a baseline measure of their existing understanding of neural networks. These questions covered introductory concepts such as the definition of a Neural Network, the roles and functions of nodes and layers, and data processing sequences. Participants also demonstrated an understanding of concepts/practices like supervised and unsupervised learning, the impact of network architecture on energy consumption, and the significance of hardware components like CPUs, GPUs, and TPUs. For example, participants faced questions like *What is a layer in a neural network?* and *How are GPUs, CPUs, and TPUs important for neural networks or CNNs?* All questions were closed-ended, either multiple choice or True / False, to gauge the foundational knowledge of neural network theory. Further analysis of the 14 multiple-choice questions showed a generally high level of understanding by the participants, with an average of 80.7% correct. Most of the questions were correctly answered, but some had rates below 60%, indicating knowledge gaps—topics that need more explanation to help understanding.

Moreover, an independent samples *t*-test compared mean differences between the XR and Desktop groups regarding "correctness" and "time spent responding to questions" (in seconds). Welch's test was employed due to the lack of equal variances, as indicated by Levene's test for

Table 3 Independent samples *t*-test - reprinted from our previous conference paper [32]

	t	df	*p*	Mean Difference	SE Difference	Cohen's d	SE Cohen's d
Correctness	−0.6	54	0.7	0.1	0.3	0.2	0.3
Duration	−4.4	34.7	0.0	−35.3	10.8	−1.2	0.3

For all tests, the alternative hypothesis specifies that group *Desktop* is less than group *XR*

Welch's *t*-test

Table 4 Multiple-choice questions Scores - reprinted from our previous conference paper [32]

Statistic	XR ($n = 29$)	Desktop ($n = 27$)
Mean	4.2	4.4
Standard Error (SE) Skewness	0.2	0.1
Std. Deviation	1.2	0.8
Skewness	−1.2	0
Standard Error (SE) Skewness	0.4	0.4
Kurtosis	1.2	−0.4
Standard Error (SE) Kurtosis	1.9	1.9

"correctness" (refer to Table 3 reprinted from our previous conference paper [32]).

Regarding correctness, both groups had similar mean scores (3.9 for XR and 4.1 for Desktop) with no significant difference ($p = .7$), indicating comparable prior knowledge of Neural Networks. Nevertheless, the period taken by the XR group to complete the pre-questionnaire was significantly longer than the Desktop group (33.4 s versus 68.7 s; $p < .001$). This extended duration might result from XR participants' lack of familiarity with Neural Networks, requiring significantly more cognitive effort.

In-Lesson Assessment of Learning Outcomes

On the other hand, during the intervention, the lesson asked six multiple-choice questions assessing learning outcomes. These questions were strategically embedded throughout the lesson to assess participants' comprehension and retention of the material as it was being presented, thus directly measuring their learning gains. These questions focused on foundational concepts related to Neural Networks and Hand-written Digit Recognition, for example, about *How many input neurons are there in a network handling MNIST images*, and *What do weights in the neural network represent*. Some of the topics addressed in other questions were *What ReLU does when given a negative input* and *What ReLU is used to introduce* the goal of the softmax function?, and *Which of the predefined test cases (0, 1, 3, and 8) has the least confidence in prediction?*. These questions assessed student understanding and interest in the material presented in the lesson.

Table 4, reprinted from our previous conference paper [32], depicts the scores for the multiple-choice questions during the intervention. The XR group scored a mean of 4.2 (Standard Error (SE) Skewness = 0.2, *SD* = 1.2), indicating

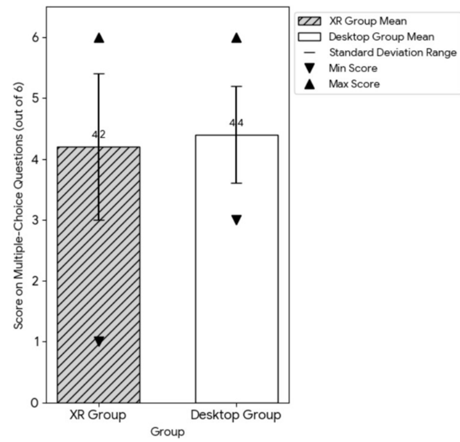

Fig. 4 Comparison of in-lesson multiple-choice scores between XR and Desktop groups

high score variability, with skewness of -1.2, suggesting more participants scored above the mean. The scores varied from 1.0 to 6.0. In contrast, the Desktop group got a mean score of 4.4 (Standard Error (SE) Skewness = 0.1, SD = 0.8), indicating less variability and more consistent performance. The skewness of zero indicates a near-normal distribution, which means the scores are uniformly distributed around the mean. In addition, the kurtosis of −0.4 indicates fewer extreme values, which means fewer high or low scores, resulting in more consistent performance. Their scores ranged from 3.0 to 6.0. Although the Desktop group showed higher average performance and less variability, the XR group's lower mean score and more significant variability may stem from their unfamiliarity with XR technology. As participants become more accustomed to XR, their performance may improve, potentially narrowing the gap with the Desktop group.

Figure 4 illustrates the scores on the multiple-choice questions for both the XR and Desktop groups, reinforcing the statistical data presented in Table 4 reprinted from our previous conference paper [32].

This figure shows the mean scores, with the Desktop group having a mean of 4.4, slightly outperforming the XR group with a mean of 4.2. The error bars, representing the standard deviation range, highlight the greater variability in scores within the XR group with a standard deviation of 1.2, compared to the Desktop group with a standard deviation of

0.8. This is further emphasized by the minimum and maximum score markers, which show a wider range of performance in the XR group (1.0 to 6.0) compared to the Desktop group (3.0 to 6.0). While both groups show high scores, Fig. 4 underscores that the desktop group had a more consistent performance and the XR group had a wider distribution of results, which may be due to the "novelty effect" of the XR technology for many participants.

User Engagement Assessment (Session Time)

Furthermore, user engagement was assessed in both the XR and desktop contexts, with the session time metric considered an important factor in analyzing user engagement and persistence. Extended periods demonstrate a deeper engagement with the topic matter. Table 5, reprinted from our previous conference paper [32], depicts the results on Session Time. Using the XR headset, the XR group had a mean session time of 723.7 s (Standard Error (SE) Skewness = 33.4, *SD* = 179.8), indicating considerable variability. The skewness of −0.2 suggests a near-normal distribution, meaning that session times are mostly balanced around the average. Most of the participants had session times close to the mean, reflecting a balanced experience. Furthermore, a kurtosis of −0.9 suggests that the distribution is slightly flatter, with fewer extreme values and more consistent session times. That is, this kurtosis suggests that there was a more uniform experience among the XR group than among the other group. The session lasted between 359.0 and 1009.0 s.

However, the mean session duration for desktop users was 578.5 s (Standard Error (SE) Skewness = 40.1), showing greater variability due to a standard deviation of 212.2 s, suggesting that the duration of the sessions varied significantly within the desktop group, with certain participants spending more or less time compared to others. This indicates that most of the participants finished their sessions on time, while a small number took significantly longer, resulting in a skewed distribution to the right. The skewness of 0.8 indicates a right-skewed distribution, meaning most session times were shorter, with a few much longer outliers.

Table 5 Session time (in seconds) - reprinted from our previous conference paper [32]

Statistic	XR (n = 29)	Desktop (n = 27)
Mean	723.7	578.5
Standard Error (SE) Skewness	33.4	40.1
Std. Deviation	179.8	212.2
Skewness	−0.2	0.8
Standard Error (SE) Skewness	0.4	0.4
Kurtosis	−0.9	1.5
Standard Error (SE) Kurtosis	1.9	1.9

Most desktop users had shorter session times, but a few participants took much longer than the average. Furthermore, a kurtosis of 1.5 indicates more noticeable tails, implying more extreme session times, both shorter and longer, resulting in increased variability during the intervention. The session times ranged from 239.0 to 1237.0 s.

User Experience

The results show that participants using the XR setting reported higher levels of usability, satisfaction, and likelihood to recommend the system than those using the Desktop version. The XR group (n = 29) experienced consistently positive interactions, while Desktop users (n = 27) faced more usability challenges and lower satisfaction levels. Detailed findings for each scale assessing user experiences in both environments focused on Neural Networks and Handwritten Digit Recognition are provided.

The results of the System Usability Scale (SUS) [24] indicate that users found the XR system more engaging and easier to navigate. For instance, the statement "*I think that I would like to use this system frequently* " (SUS1) received a mean score of 3.3 in the XR group versus 3.2 in the Desktop group, showing a stronger inclination to use the XR system. Although both groups rated the statement "*I thought the system was easy to use*" (SUS3) similarly, the Desktop group exhibited greater variability, reflecting inconsistent experiences.

Moreover, as indicated in Table 6 reprinted from our previous conference paper [32], the XR group achieved an average SUS score of 60.1, which aligned closely with the standard norm of 68. This shows that the perceived usability was acceptable, yet it suggests that there is still room for enhancement. Nonetheless, the average SUS score for the Desktop group of 57.9 indicated increased challenges. This also indicates that while both systems require enhancements, the XR system is showing superior performance in terms of overall usability compared to the Desktop system.

The User Satisfaction Questionnaire (USQ) results indicate high overall satisfaction, particularly among XR participants, who reported greater satisfaction than Desktop users. The XR group achieved a mean score of 4.1 for *"satisfaction with the XR experience"* (USQ1), with a standard deviation of 0.8. In contrast, the Desktop group reported a mean satisfaction score of 3.6 and a higher standard deviation of 0.8, indicating more variability. For satisfaction with visual quality (USQ3), the XR group scored higher at 3.9, compared to the Desktop group's score at 3.6. This suggests a better perceived visual quality and a more consistent positive perception in the XR environment, indicating an effective user expectation management, compared to the lower

Table 6 System Usability Scale—reprinted from our previous conference paper [32]

	SUS1	SUS2	SUS3	SUS4	SUS5	SUS6	SUS7	SUS8	SUS9	SUS10
Descriptive Statistics: XR (n=29)										
Mean	3.3	2.4	3.6	2.5	3.8	2.0	2.2	2.9	2.8	2.2
Standard Error (SE) Mean	.1	.1	.1	.2	.1	.1	.1	.2	.2	.2
Standard Deviation	1.0	1.0	1.0	1.3	.9	.9	1.0	1.1	1.2	1.3
Skewness	.1	.4	−.7	.6	−.5	.9	.4	.2	−.4	.8
Standard Error (SE) Skewness	.3	.3	.3	.3	.3	.3	.3	.3	.3	.3
Kurtosis	−.7	-1.0	.1	−.7	−.4	.2	−.8	−.6	-1.1	−.6
Standard Error (SE) Kurtosis	.6	.6	.6	.6	.6	.6	.6	.6	.6	.6
Descriptive Statistics: Desktop (n=27)										
Mean	3.2	2.2	3.6	2.3	3.7	2.2	2.6	3.3	2.3	2.3
Standard Error (SE) Mean	.2	.2	.2	.2	.2	.2	.2	.2	.3	.3
Standard Deviation	1.0	1.0	1.0	1.1	.9	1.1	1.0	1.0	1.3	1.3
Skewness	.3	.5	−.4	.6	−.5	.7	−.2	−.3	.4	.5
Standard Error (SE) Skewness	.5	.5	.5	.5	.5	.5	.5	.5	.5	.5
Kurtosis	−.8	−.7	−.8	−.4	−.3	−.6	-1.3	−.4	-1.3	-1.2
Standard Error (SE) Kurtosis	.9	.9	.9	.9	.9	.9	.9	.9	.9	.9

Table 7 User satisfaction Questionnaire - reprinted from our previous conference paper [32]

	USQ1	USQ2	USQ3
Descriptive Statistics: XR (*n* = 29)			
Mean	4.1	4.1	3.9
Standard Error (SE) Skewness	0.1	0.2	0.2
Standard Deviation	0.8	1.1	0.9
Skewness	−0.1	−1.0	−0.3
Standard Error (SE) Skewness	0.4	0.4	0.4
Kurtosis	−1.2	0.6	−0.7
Standard Error (SE) Kurtosis	0.9	0.9	0.9
Descriptive Statistics: Desktop (*n* = 27)			
Mean	3.6	3.6	3.6
Standard Error (SE) Skewness	0.2	0.2	0.2
Standard Deviation	0.8	1.2	1.1
Skewness	−1.2	−0.3	−0.7
Standard Error (SE) Skewness	0.5	0.5	0.5
Kurtosis	0.6	−0.7	−0.0
Standard Error (SE) Kurtosis	0.9	0.9	0.9

scores of the Desktop group, indicating some areas needing improvement (refer to Table 7 reprinted from our previous conference paper [32]).

In general, the participants who used the XR system were very satisfied, with an overall average rating of 4.1 out of 5. Despite this, desktop users had an average rating of 3.6, suggesting that there is room for improvement in visual appeal and user-friendliness, although users in this group were generally satisfied.

On the other hand, findings from the Immersive Presence Questionnaire (IPQ) [27] show that participants in the XR group experienced a higher sense of presence than desktop group participants. For instance, the item *How aware were you of the real-world surroundings while navigating in the virtual world?*, had a mean of 1.8, standard deviation of 0.9, indicating high immersion and reduced awareness of the real world. In contrast, the item *"How real did the virtual world seem to you?"* had a mean score of 0.6 and a standard deviation of 1.6, indicating varied perceptions of realism. This variation may stem from individual sensitivities to immersive experiences. The moderate immersion measured by the overall IPQ score of 3.48 on 56 participants demonstrates the strong experience of presence when users are engaged in the virtual world. However, these results also indicate room to improve the consistency in engagement and realism for all users.

Furthermore, the average likelihood of 56 people recommending their experience was 7.2, indicating that their opinions differed. XR users displayed more excitement, with an average NPS score of 7.9, showing their increased likelihood of recommending the experience. In contrast, the Desktop users scored 6.5, indicating the need for improvement to match the recommendability of the XR version.

Another independent sample *t*-test was conducted to investigate the effects of the XR environment compared to the Desktop version on instructional design. This study looked at the System Usability Scale (SUS), User Satisfaction Questionnaire (USQ), and Net Promoter Score (NPS) in both the XR and Desktop groups. The goal was to demonstrate that participants in the XR environment would report higher scores in perceived presence, system usability, and overall satisfaction than those using the Desktop version (refer to Table 8 reprinted from our previous conference paper [32]). This data analysis indicates that users of the XR system reported significantly higher user satisfaction (measured by the User Satisfaction Questionnaire, USQ) and a greater likelihood of recommending the system (indicated by the Net Promoter Score, NPS) compared to Desktop users. This suggests that XR participants were more

Table 8 Independent samples *t*-test - reprinted from our previous conference paper [32]

	t	df	p	Mean Difference	SE Difference	Cohen's d	SE Cohen's d
System Usability Scale							
SUS1	−.7	54	.2	−.2	.3	−.2	.3
SUS2	−.9	54	.2	−.3	.3	−.3	.3
SUS3	.4	53.8	.7	.1	.3	.1	.3
SUS4	−1.2	53.7	.1	−.4	.3	−.3	.3
SUS5	−.1	54	.5	−.0	.3	−.0	.3
SUS6	1.0	48.8	.8	.3	.3	.3	.3
SUS7	2.9	43.5	1.0	.7	.3	.8	.3
SUS8	3.0	53.6	1.0	.8	.3	.8	.3
SUS9	−3.3	45.1	.0	−1.0	.3	−.9	.3
SUS10	.3	52.5	.6	.1	.3	.1	.3
User Satisfaction Questionnaire							
USQ1	−2.1	53.2	.0	−.4	.2	−.6	.3
USQ2	−1.4	53	.1	−.4	.3	−.4	.3
USQ3	−1.2	51.8	.1	−.3	.3	−.3	.3
Net Promoter Score							
NPS	−2.4	47.7	.0	−1.4	.6	−.7	.3

For all tests, the alternative hypothesis specifies that group *Desktop* is less than group *XR*

Welch's *t*-test

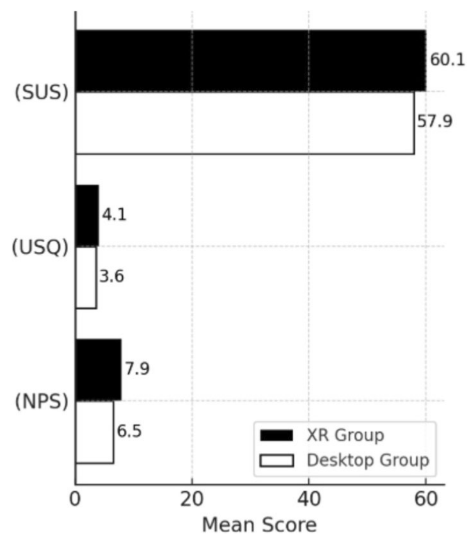

Fig. 5 Comparison between the metrics assessing user experience for the XR and Desktop groups

satisfied and inclined to recommend their experience with the system.

Figure 5 illustrates how the XR group consistently reported higher mean scores across the different metrics. The figure shows how the XR group achieved an average System Usability Scale (SUS) score of 60.1, suggesting a better perceived usability compared to the Desktop group's average of 57.9. In a similar way, regarding the user satisfaction reported by the User Satisfaction Questionnaire (USQ), the XR group scored higher than the Desktop group with a score of 4.1 over 3.6. Moreover, for the Net Promoter Score (NPS), the XR group reached a mean of 7.9, clearly surpassing the Desktop group's of 6.5, indicating a greater likelihood of recommending the system.

Figure 5 highlights that participants using the XR system experienced greater usability, satisfaction, and a stronger intention to recommend the system than those using the traditional desktop setup. However, the two groups did not have significant differences in system usability (as assessed by the System Usability Scale, SUS), indicating similar usability ratings. For this independent samples *t*-test, Welch's test approach was used due to unequal variances indicated by Levene's test for items SUS7 and SUS9, particularly regarding ease of learning and confidence. Although normality tests showed significant deviations ($p < .05$), the Central Limit Theorem supports the idea that the distribution of sample means approximates normality as sample sizes grow larger [59].

The system usability scale (SUS) analysis did not show significant differences between the XR and Desktop groups for most items. However, XR users felt more confident using the system. This was clear in their responses about user confidence (SUS9), where XR users had higher scores ($t = -3.3$, $p < .01$) and a strong effect size (Cohen's $d = -0.9$).

In the User Satisfaction Questionnaire (USQ), apparent differences were observed in the first item (USQ1), with a *p*-value below 0.05. However, effect sizes of −0.6 and −0.4, for USQ1 and USQ2 show that XR users had a better overall experience and rated the visual quality higher than Desktop users. There was also a slight difference in how easy users found the system to use and visual quality (USQ2 and USQ3), with *p*-values of 0.1, suggesting that XR users found it easier to use and were more satisfied regarding the visual quality of the environment.

Finally, for the Net Promoter Score (NPS), a significant difference ($p < .01$) with a medium effect size (Cohen's $d = -0.7$) was found, meaning that XR users were more likely to recommend the system. Overall, XR users reported higher satisfaction, confidence, and a greater willingness or likelihood to recommend their experience with the system.

Discussion

Involving immersion technologies in education, such as XR, to introduce complex concepts like AI-related topics (e.g., neural networks or handwritten digit recognition) may require a mix of sensory engagement, interactivity, and relevance. Research suggests that high-quality visuals, realistic simulations, and interactive features help learners actively explore complex topics (e.g., AI concepts), making them easier to understand and remember [34]. However, it is important to avoid overwhelming students with too much information at once, so technology should support learning without causing distractions [33]. Thus, gradually introducing XR technologies and providing guidance can help students focus on their learning. Guided by assumptions of embodied cognition, this study examined the educational value of students performing actions or movements during learning and how those forms of engagement may have resulted in students' understanding of abstract concepts more effectively. In this study, while the XR environment offered a higher level of embodiment through direct interaction with 3D neural network layers via the hand controllers, the desktop environment limited embodied interaction to mouse and keyboard [35, 36].

The findings suggest that XR environments may boost engagement, promoting deeper cognitive processing via meaningful interactions with the content, as visualizations are particularly useful in comprehending complex ideas, such as Neural Networks, as they aid in understanding [11]. In general, users found the XR system to be easier to navigate and allowing them to focus on learning rather than struggling with technology. This is supported by the System Usability Scale (SUS) scores, where the participants in the XR group had an average score of 60.1, compared to 57.9 for the participants in the desktop group, suggesting that the XR users had a more positive experience and increased participation in their learning tasks. It is important to point out that both the XR and Desktop groups received the same learning content about Neural Networks and Handwritten Digit Recognition, but the XR environment included interactive 3D visualizations and immersive features, while the Desktop version only showed standard 2D graphics.

Moreover, desktop interfaces often lack the immersive elements found in XR, leading to a broader range of user experiences that may hinder the educational progress of some students. However, based on the findings presented in this paper, the interactive quality of XR enables students to interact with 3D models, improving comprehension and memory of complex ideas and grasping deeper complex concepts, giving this technology a notable edge over conventional approaches. However, our findings indicate that increased embodiment did not automatically produce higher learning scores, which aligns with previous research that has suggested that immersive settings may increase cognitive load when insufficient support is provided [60].

Nevertheless, despite XR participants reporting more engagement, as evidenced by longer session lengths and higher user satisfaction, this did not transfer into better learning outcomes, as both the XR and Desktop groups provided similar correct responses. The variability in XR user performance suggests that not all benefited equally from the immersive experience, with some struggling with the technology. Factors, such as the novelty of the XR technology, can distract participants from educational content, and the learning curve associated with XR could hinder material absorption [61]. Then, to overcome these challenges, students may need specific training (i.e., scaffolding) and support to use technology and stay focused on learning objectives properly. Accordingly, providing scaffolding can reduce this cognitive load, as structured guidance on how to use gestures and controls can help students transform their physical actions in XR into a meaningful understanding of the abstract content being studied [60].

These performance gaps in learning within the XR group became evident during our intervention. Although this group reported higher engagement, as suggested by our findings, they also showed a mean score of 4.2 on the multiple-choice questions, and their results (scores) were more variable and less consistent. In fact, the higher standard deviation of the XR group (1.2) compared to the one reported by the Desktop group (0.8) indicates that the individual scores within the XR group were more spread out from their mean, suggesting greater variability and less consistency in performance among the participants. The statistical analysis further supports this claim, as it also showed that the XR group reported a negative skewness in their test scores, suggesting that while several students had an overall high performance, others scored significantly lower, which suggests that there is an inconsistency in how learners responded to the XR experience, due to the resulting wider distribution of the results. In addition, the kurtosis value further suggests that some students either thrived or struggled considerably more than expected (refer to Table 4 reprinted from our previous conference paper [32]). Therefore, such inconsistencies may indicate that the learning benefits of XR are not evenly distributed and may depend on different factors, such as the prior experience or the cognitive readiness of the learners [60]. For instance, prior familiarity influences the effectiveness of embodiment, as without experience, interactive features may demand additional cognitive effort and limit the intended learning outcomes.

The high initial cognitive load caused by the novelty and the learning curve of XR technology helps hypothesize on why the XR group's higher engagement did not always result in better learning outcomes [60, 62, 63]. Although the immersive experience captured the students' attention and created an overall positive learning environment, as suggested by our findings, many had to spend a higher cognitive effort learning how to use the new interface in the immersive XR environment [64]. Consequently, this suggests that it may have likely reduced their ability to focus on understanding the complex AI concepts introduced in the intervention. As a result, the performance within the XR group varied, with some students who adapted quickly and were able to benefit from the immersive features, while others may have struggled with the XR technology, which may have affected how much they learned, despite being actively engaged in the immersive learning approach.

This learning gap could be related to the "novelty effect" that the XR technologies could have, for instance, where the excitement or curiosity about the tools captures users' attention but may unintentionally shift their focus away from the instructional content [65]. Many students were using XR for the first time, and this new experience may have distracted them from the actual learning objectives. In fact, as reported in Table 1, reprinted from our previous conference paper [32], 51.7% of XR users had never used XR before, and 37.9% had used VR only once. Hence, as suggested by our demographics, most students in our study were unfamiliar with this kind of technology and required significant time and effort to understand how to use the system, compared to the desktop-based environment for the same content, following the same instructional design.

Moreover, to improve the educational value of XR, it is important to support both students and instructors to make the most of these technologies [66]. This involves providing a proper scaffolding within the XR environment to facilitate effective learning, which could include, for instance, interactive tutorials for gesture controls within the system, or customizable features that allow students to adjust the difficulty of tasks as they become more comfortable and familiar with the XR system [67]. These approaches aim to have students focus more on understanding AI concepts rather than struggling with the technology itself. At the same time, instructors need training that goes beyond learning how to use XR tools solely. They should also learn how to observe and support student thinking in 3D spaces, guide collaborative problem-solving in virtual environments, and offer real-time feedback that uses XR's features. Therefore, with the right scaffolding for both learners and educators, XR technologies can be used more effectively to create deeper and more meaningful learning experiences [66]. Despite these challenges, technologies like XR have the potential to greatly increase student engagement. They help students learn and retain complex concepts more easily by making them more accessible, improving motivation, and sparking interest [19]. Integrating XR into education could create a more dynamic and effective learning environment [11]. However, as indicated before, further research is needed, using larger sample sizes and more diverse STEM topics, to compare XR with traditional learning methods and their impact on learning outcomes. Additionally, it is important to explore how immersive technologies like XR can support different learning styles, particularly for students who may need extra time or specific support to adapt to this technology. Future studies should focus on how XR can be used to personalize learning experiences, tailoring them to individual needs. Therefore, when used effectively, immersive technologies, such as XR, have great potential to enhance learning experiences and improve educational outcomes.

Conclusions, Limitations, and Future Work

The results of this study show the differences in learning between XR environments and traditional desktop settings, highlighting their distinct advantages and disadvantages in teaching complex AI concepts. The students' experiences and outcomes differed, while both technological approaches intended to introduce complex topics by employing interactive tasks, such as Neural Networks and Handwritten Digit Recognition. XR technology has made learning more engaging and easier to use, but some challenges need to be considered.

In this study, the control group used a laptop, which may have influenced the results. A larger screen, such as a desktop monitor, provides a wider field of view and can enhance immersion and engagement, similar to some benefits of an XR headset but without the physical discomfort or novelty issues. In this study, a laptop was chosen for practical reasons, as laptops are a common setup for computer-based learning. However, the smaller and less immersive laptop screen may have limited the desktop experience, making the comparison with the XR group less direct. To address this limitation, future research should include a standard or extended desktop monitor to examine how screen size influences engagement and learning outcomes.

The outcomes of this research suggest that the variability in performance among XR users may arise from their unfamiliarity with the technology [68]. Although performance may increase with more exposure, this leads to concerns about the initial learning process and differences in adaptability among students, particularly those who are not as familiar with technology [65]. Thus, training and scaffolding are needed to support the adaptability of new immersive technologies in education.

Desktop environments are generally considered more intuitive, as they do not need special equipment like XR environments, making it possible for students without XR access to participate [11]. There is also a higher price for the XR equipment, which makes it difficult to acquire. Fortunately, the performance/price ratio of XR headsets continues to increase. For example, our XR environment can now be deployed on a $300 Meta Quest 3 S.

Nevertheless, there are several limitations of the current version of *ACHIEVE*. For instance, although users can customize many visualization settings, these adjustments must be made manually at this time, and not in an automated way. Moreover, while the system can currently load different pre-trained models to show different learning states of the NN, it currently uses static weights and does not support live training or dynamic weight updates, which would require a more advanced back-end system.

The authors acknowledge that this research only investigated the quantitative perspective. Therefore, further research will be conducted to analyze the effectiveness of immersive technologies, such as XR, in a deeper way using different lenses. For this, multiple methods will be used to identify and reflect on the interplay between overall performance and rationales in measuring students' learning outcomes. That is, further research aims to provide a more complete understanding of how these immersive technologies impact learning outcomes, reflecting on the variability in students' experiences and on how they can be used more effectively in education.

The multiple-choice questions are useful for standardized testing, as reported preliminarily in this manuscript. However, they may not fully measure learning, the deeper understanding, and the skills that students develop in an immersive setting [9]. For instance, XR can enable students to interact with and visualize abstract data flows in a neural network, where learning extends beyond selecting the correct answer in closed-ended questions [9, 69]. Therefore, further research may include open-ended questions that ask students to explain their thinking and how they understand other AI concepts, such as AI bias and generalization. Overall, these responses will help provide a better understanding of how students are learning. In addition, the tasks that students complete in the XR environment will also test their ability to apply what they have learned, for instance, by tracing data through network layers or adjusting an AI model (i.e., a neural network) in real-time. Hence, these approaches may provide a better understanding of what students have learned, examining skills such as critical thinking and hands-on problem-solving, in addition to engagement, which are skills supported by immersive technologies.

Furthermore, the results describing how the in-lesson learning outcome scores were comparable between the XR and Desktop groups suggest that the effect size was likely small. Therefore, to reliably detect effects of a small magnitude with 80% power, a significantly larger sample size is required [55]. However, since this was an exploratory study, the sample size $N = 56$ was sufficient to identify medium-sized differences. Nevertheless, further research should include larger and more diverse samples to improve statistical power to detect smaller effects and to support and complement all claims made in this exploratory study.

Building on the findings from this study, further research could explore whether the benefits of learning in XR environments last over time. Although the students may have demonstrated understanding immediately after the XR lesson, it is still unclear whether this knowledge remains enduring. Hence, conducting a longitudinal study could provide evidence of the long-term effects of XR on learning AI topics [70]. Moreover, further research could explore expanding the XR content beyond other abstract AI concepts, beyond Neural Networks and Handwritten Digit Recognition, to better understand how this technology can support a broader range of challenging AI topics [71]. Further studies could also examine how XR lessons work when implemented in real classroom settings, such as with student groups or full-class settings, which could provide a clearer picture and better understanding of how XR performs under typical learning conditions, including time limitations and diverse student needs. In addition, future work could involve enhancing the XR system with features that enable users to control their view and interaction with AI models (e.g., adjusting network layers or testing different datasets), which may also increase its educational value. These features could also allow for customization based on the topic or the learner's preferences.

Furthermore, to elaborate on the comparison between XR and desktop-based instruction more clearly, future research could use additional instructional designs, such as grouping students with similar backgrounds or having students experience both learning methods—both XR and desktop environments—with different content, as these approaches may help identify and better understand when XR is most helpful and how it can complement traditional learning.

Acknowledgements The authors thank all participants who voluntarily and actively contributed to the study with their participation.

This material is based upon work supported by the National Science Foundation under award numbers 2412928, 2417510, 2212200, 2219842, 2309564, and 2318657.

Author Contributions Conceptualization, M.A.F.-G, Y.Z., Y.G., B.B., A.J.M., VP; methodology, M.A.F.-G, Y.Z., Y.G., B.B., A.J.M., VP; validation, Y.Z. and M.A.F.-G.; formal analysis, M.A.F.-G.; investigation, A.J.M., B.B. and V.P.; resources, A.J.M., B.B. and V.P.; data curation, M.A.F.-G., Y.Z., and Y.G; writing—original draft preparation, M.A.F.-G., Y.Z., B.B., A.J.M., and V.P.; writing—review and editing, M.A.F.-G., Y.Z., Y.G., B.B., A.J.M., and V.P.; supervision, A.J.M.; project administration, V.P.; funding acquisition, A.J.M., B.B., and V.P. All authors have read and agreed to the published version of the manuscript.

Funding This work was supported in part by the National Science Foundation under awards 2412928, 2417510, 2212200, 2219842, 2309564, and 2318657. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Data Availability The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

Materials Availability Not applicable.

Code Availability Not applicable.

Declarations

Conflict of Interest The authors declare no conflicts of interest. The authors have no relevant financial or non-financial interests to disclose.

Ethical Approval and Consent to Participate This study has been approved by Purdue University's Institutional Review Board under IRB-2024-57. A four-page consent form was designed to inform participants about the purpose, duration, confidentiality, benefits, risks, and other pertinent information of the study. If they chose to participate, participants first signed and dated the consent form. Participants could also withdraw at any time.

Consent for Publication The authors declare that this work is original and not considered to be published in any other publication media.

References

1. Pham ST, Sampson PM. The development of artificial intelligence in education: A review in context. J Comput Assist Learn. 2022;38(5):1408–21.
2. Kim S-W. Change in attitude toward artificial intelligence through experiential learning in artificial intelligence education. Int J Adv Sci Eng Inform Technol 13(5), pp. 1953–1959 (2023).
3. Feij´oo-Garcıa MA, Ram´ırez-Ar´evalo HH, Feij´oo-Garc´ıa PG. Collaborative strategy for software engineering courses at a south american university. In: Proceedings of the 13th International Conference on Computer Supported Education - Volume 1: CSEDU, pp. 266–273 (2021). https://doi.org/10.5220/0010460602660273
4. Chen L, Chen P, Lin Z. Artificial intelligence in education: A review. Ieee Access. 2020;8:75264–78.
5. Pedr´o F. Applications of artificial intelligence to higher education: possibilities, evidence, and challenges. IUL Res. 2020;1(1):61–76.
6. Guo X, Guo Y, Liu Y. The development of extended reality in education: inspiration from the research literature. Sustainability. 2021;13(24):13776. https://doi.org/10.3390/su132413776.
7. Zwolin´ski G, Kamin´ska D, Laska-Le´sniewicz A, Haamer RE, Vairinhos M, Raposo R, Urem F, Reisinho P. Extended reality in education and training: case studies in management education. Electronics. 2022;11(3):336. https://doi.org/10.3390/electronics11030366.
8. Rangel-de L´azaro G, Duart JM. You can handle, you can teach it: systematic review on the use of extended reality and artificial intelligence technologies for online higher education. Sustainability. 2023;15(4):3507. https://doi.org/10.3390/su15043507.
9. Kuleto V, Ili´c P, Stanescu M, Rankovi´c M, Popovi´c Sˇevi´c M, Pˇaun N, Teodorescu D. Extended reality in higher education, a responsible innovation approach for generation Y and generation Z. Sustainability. 2021;13(21):11814. https://doi.org/10.3390/su132111814.
10. Gu Y, Feijoo-Garcia MA, Zhang Y, Magana AJ, Benes B, Popescu V. An XR environment for AI education: Design and first implementation. In: 2024 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW), pp. 157–162 (2024). https://doi.org/10.1109/VRW62533.2024.00032. IEEE.
11. Zhang Y, Feijoo-Garcia MA, Gu Y, Popescu V, Benes B, Magana AJ. Virtual and augmented reality in science, technology, engineering, and mathematics (STEM) education: an umbrella review. Information. 2024;15(9):515. https://doi.org/10.3390/info15090515
12. Kosko KW, Ferdig RE, Roche L. Conceptualizing a shared definition and future directions for extended reality (XR) in teacher education. J Technol Teacher Educ. 2021;29(3):257–77.
13. Kharvari F, Kaiser LE. Impact of extended reality on architectural education and the design process. Automation in Construction 141, 104393 (2022) https: https://doi.org/10.1016/j.autcon.2022.104393
14. Jagatheesaperumal SK, Ahmad K, Al-Fuqaha A, Qadir J. Advancing education through extended reality and internet of everything enabled meta-verses: Applications, challenges, and open issues. IEEE Trans Learn Technol. 2024;17:1120–39. https://doi.org/10.1109/TLT.2024.3358859
15. Ferreira JMM, Qureshi ZI. Use of XR technologies to bridge the gap between higher education and continuing education. In: 2020 IEEE Global Engineering Education Conference, pp. 913–918 (2020). https://doi.org/10.1109/EDUCON45650.2020.9125346
16. Chiang Y-C, Liu S-C. The effects of extended reality technologies in STEM education on students' learning response and performance. J Baltic Sci Educ. 2023;22(4):568–78.

17. L´opez-Ojeda W, Hurley RA. Extended-reality technologies: an overview of emerging applications in medical education and clinical care. J Neuropsychiatry Clin Neurosciences. 2021;33(3):4–177. https://doi.org/10.1176/appi.neuropsych.21030067
18. Koukopoulos D, Dafiotis P, Sylaiou S, Koukoulis K, Fidas C. Xr technolo gies for self-regulated student exhibitions in art education: Survey and first design considerations. In: 2022 International Conference on Interactive Media, Smart Systems and Emerging Technologies, pp. 1–8 (2022). https://doi.org/10.1109/IMET54801.2022.9929450
19. Guilbaud P, Guilbaud TC, Jennings D. Extended reality, pedagogy, and career readiness: A review of literature. In: Chen JYC, Fragomeni G, editors. Virtual, augmented and mixed reality. Cham: Springer; 2021. pp. 595–613.
20. Idrees A, Morton M, Dabrowski G. Advancing extended reality teaching and learning opportunities across the disciplines in higher education. In: 2022 8th International Conference of the Immersive Learning Research Network, pp. 1–8 (2022). https://doi.org/10.23919/iLRN55037.2022.9815897
21. Fern´andez-Batanero JM, Montenegro-Rueda M, Fern´andez-Cerero J, L´opez- Meneses E. Extended reality as an educational resource in the primary school classroom: an interview of drawbacks and opportunities. Computers. 2024;13(2):50. https://doi.org/10.3390/computers13020050.
22. Meccawy M. Creating an immersive XR learning experience: A roadmap for educators. Electronics. 2022;11(21):3547. https://doi.org/10.3390/electronics11213547.
23. Obeidallah R, Ahmad AA, Qutishat D. Challenges of extended reality technology in higher education: A review. Int J Emerg Technol Learn 2023;18(14):39–50
24. Grier RA, Bangor A, Kortum P, Peres SC. The system usability scale: Beyond standard usability testing. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 57, pp. 187–191 (2013). SAGE Publications Sage CA: Los Angeles, CA.
25. Romeike R. The role of computer science education for understanding and shaping the digital society. In: Sustainable ICT, Education and Learning: IFIP WG 3.4 International Conference, SUZA 2019, Zanzibar, Tanzania, April 25–27, 2019, Revised Selected Papers 1, pp. 167–176 (2019). Springer.
26. Gil-G´omez J-A, Manzano-Hern´andez P, Albiol-P´erez S, Aula-Valero C, Gil- G´omez H, Lozano-Quilis J-A. Useq: a short questionnaire for satisfaction evaluation of virtual rehabilitation systems. Sensors. 2017;17(7):1589.
27. Schwind V, Knierim P, Haas N, Henze N. Using presence questionnaires in virtual reality. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, pp. 1–12 (2019).
28. Baquero A. Net promoter score (NPS) and customer satisfaction: relationship and efficient management. Sustainability. 2022;14(4):2011.
29. Kara A, Mintu-Wimsatt A, Spillan JE. An application of the net promoter score in higher education. J Market Higher Edu 2021;34(2):478–501.
30. Brown NJ, Furtak EM, Timms M, Nagashima SO, Wilson M. The evidence-based reasoning framework: assessing scientific reasoning. Educational Assess. 2010;15(3–4):123–41.
31. Feijoo-Garcia MA, Holstrom MS, Magana AJ, Newell BA. Simulation-based learning and argumentation to promote informed design decision-making processes within a first-year engineering technology course. Sustainability. 2024;16(7):2633. https://doi.org/10.3390/su16072633
32. Feijoo-Garcia MA, Zhang Y, Gu Y, Magana AJ, Benes B, Popescu V. Exploring extended reality (XR) in teaching AI: A comparative study of XR and desktop environments. In: Proceedings of the 20th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - HUCAPP, pp. 472–482 (2025). https://doi.org/10.5220/0013141000003912
33. Skulmowski A. Learning by doing or doing without learning? The potentials and challenges of activity-based learning. Educational Psychol Rev. 2024;36(1):28.
34. Marougkas A, Troussas C, Krouska A, Sgouropoulou C. Virtual reality in education: a review of learning theories, approaches and methodologies for the last decade. Electronics. 2023;12(13):2832.
35. Stolz SA. Embodied learning. Educational Philos Theory. 2015;47(5):474–87.
36. Nathan MJ. Foundations of embodied learning: A paradigm for education. New York: Routledge; 2021.
37. Barsalou LW. Grounded cognition. Annu Rev Psychol. 2008;59(1):617–45.
38. Hayes JC, Kraemer DJ. Grounded Understanding of abstract concepts: the case of stem learning. Cogn Research: Principles Implications. 2017;2(1):7.
39. Martin A. Grapes—grounding representations in action, perception, and emotion systems: how object properties and categories are represented in the human brain. Psychon Bull Rev. 2016;23(4):979–90.
40. Mason RA, Just MA. Neural representations of physics concepts. Psychological Sci. 2016;27(6):904–13.
41. Gentner D. Structure-mapping: A theoretical framework for analogy. Cogn Sci. 1983;7(2):155–70.
42. Johnson-Glenberg MC, Birchfield DA, Tolentino L, Koziupa T. Collaborative embodied learning in mixed reality motion-capture environments: two science studies. J Educ Psychol. 2014;106(1):86.
43. Trowler V. Student engagement literature review. High Educ Acad. 2010;11(1):1–15.
44. Wong JT, Chen E, Au-Yeung N, Lerner BS, Richland LE. Fostering engaging online learning experiences: investigating situational interest and mind-wandering as mediators through learning experience design. Educ Inform Technol. 2024;29(14):18789–815.
45. Zhao J, Lee M, Li X. Embodied cognition and built heritage education: A case study of macau's historical architecture. Buildings. 2025;15(18):3350.
46. Laarni J, Ravaja N, Saari T, B¨ocking S, Hartmann T, Schramm H. Ways to measure spatial presence: Review and future directions. Immersed in media: Telepresence theory, measurement & technology, 139–185 (2015).
47. Aros M, Tyger CL, Chaparro BS. Unraveling the meta quest 3: An out-of-box experience of the future of mixed reality headsets. In: International Conference on Human-Computer Interaction, pp. 3–8 (2024). Springer.
48. Rivera D. Visualizing machine learning in 3D. In: Proceedings of the 28th ACM Symposium on Virtual Reality Software and Technology, pp. 1–2 (2022).
49. Shrivastava A, Jaggi I, Gupta S, Gupta D. Handwritten digit recognition using machine learning: A review. In: 2019 2nd International Conference on Power Energy, Environment and Intelligent Control (PEEIC), pp. 322–326 (2019). IEEE.
50. Ananthi S, Lokesh B, Chandran S, Sakthi S. Framework for platform independent machine learning (ML) model execution. In: 2024 2nd International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT), pp. 728–732 (2024). IEEE.
51. Alhakamy A. Extended reality (XR) toward Building immersive solutions: the key to unlocking industry 4.0. ACM-CSUR. 2024;56(9):1–38.
52. Brysbaert M. How many participants do we have to include in properly powered experiments? A tutorial of power analysis with reference tables. J Cognition. 2019;2(1):16.

53. Wright DB. Research methods for education with technology: Four concerns, examples, and recommendations. In: Frontiers in Education, vol. 4, p. 147 (2019). Frontiers Media SA.
54. Haslwanter T. An introduction to statistics with python. With applications in the life sciences. Switzerland: Springer International Publishing; 2016.
55. Cohen J. Statistical power analysis for the behavioral sciences. New York: Routledge; 2013.
56. Wagenmakers E-J, Wetzels R, Borsboom D, Maas HL, Kievit RA. An agenda for purely confirmatory research. Perspect Psychol Sci. 2012;7(6):632–8.
57. Haile ZT. Power analysis and exploratory research. J Hum Lactation. 2023;39(4):579–83.
58. Deng L. The MNIST database of handwritten digit images for machine learning research [best of the web]. IEEE Signal Process Mag. 2012;29(6):141–2.
59. Lakens D. Sample size justification. Collabra: Psychol. 2022;8(1):33267.
60. Makransky G, Petersen GB. The cognitive affective model of immersive learn- Ing (CAMIL): A theoretical research-based model of learning in immersive virtual reality. Educational Psychol Rev. 2021;33(3):937–58.
61. Alnagrat A, Ismail RC, Idrus SZS, Alfaqi RMA. A review of extended reality (XR) technologies in the future of human education: current trend and future opportunity. J Hum Centered Technol. 2022;1(2):81–96.
62. Wang J, Ma Q, Wei X. The application of extended reality technology in architectural design education: a review. Buildings. 2023;13(12):2931.
63. Hmoud M, Swaity H, Karram O, Shibli H, Swaity S, Daher W. High school students' engagement in biology in the context of XR technology. IEEE Access. 2023;11:137053–66.
64. Upadhyay B, Madathil C, Hegde K, Anderson S, Wooldridge D, Presley E, Perez D, Reid L. B.: Barriers toward the implementation of extended reality (XR) technologies to support education and training in workforce development programs. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 68, pp. 265–269 (2024). SAGE Publications Sage CA: Los Angeles, CA.
65. Parong J, Mayer RE. Cognitive and affective processes for learning science in immersive virtual reality. J Comput Assist Learn. 2021;37(1):226–41.
66. Dong W, Zhou M, Zhou M, Jiang B, Lu J. An overview of applications and trends in the use of extended reality for teaching effectiveness: an umbrella review based on 20 meta-analysis studies. Electron Libr. 2023;41(5):557–77.
67. Liu R, Wang L, Lei J, Wang Q, Ren Y. Effects of an immersive virtual reality-based classroom on students' learning performance in science lessons. Br J Edu Technol. 2020;51(6):2034–49.
68. Bautista L, Maradei F, Pedraza G. Strategies to reduce visual attention changes while learning and training in extended reality environments. International J Interact Des Manuf (IJIDeM). 2023;17(1):17–43.
69. Lin Y, Wang S, Lan Y. The study of virtual reality adaptive learning method based on learning style model. Comput Appl Eng Educ. 2022;30(2):396–414.
70. Hopwood CJ, Bleidorn W, Wright AG. Connecting theory to methods in longitudinal research. Perspect Psychol Sci. 2022;17(3):884–94.
71. Reiners D, Davahli MR, Karwowski W, Cruz-Neira C. The combination of artificial intelligence and extended reality: A systematic review. Front Virtual Real. 2021;2:721933.