

Did I Do Well? Personalized Assessment of Trainees' Performance in Augmented Reality-assisted Neurosurgical Training

Sangjun Eom*
Department of Electrical and
Computer Engineering
Duke University

Tiffany Ma†
Department of Computer
Science
Duke University
Joshua Jackson‡
Department of Neurosurgery
Duke University

Tianyi Hu‡
Department of Electrical and
Computer Engineering
Duke University
Maria Gorlatova¶
Department of Electrical and
Computer Engineering
Duke University

Neha Vutakuri§
Department of Neuroscience
Duke University

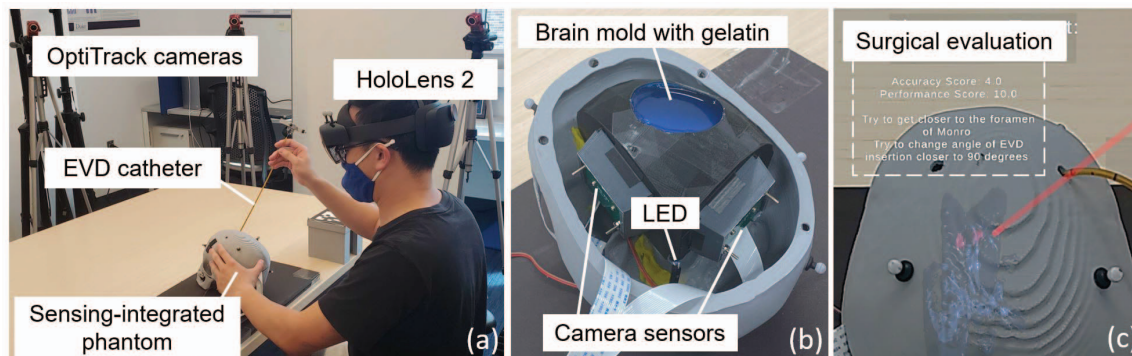


Figure 1: The overall setup of our AR-assisted EVD system (a), the 3D-printed brain mold with two camera sensors and LED placed inside (b), and the surgical assessment of the trainee's EVD performance in scores and assessment feedback (c).

ABSTRACT

The integration of Augmented Reality (AR) with the external ventricular drain (EVD) procedure improves catheter placement accuracy by providing guidance to surgeons through a visualization of the brain ventricular anatomy and other contextual information. However, the junior trainees often require feedback about their EVD performance during the AR-assisted EVD procedure. Therefore, we present an AR-assisted neurosurgical training tool for EVD that provides a personalized assessment of the trainees' EVD performance. We use a sensing-integrated brain phantom model to compute catheter placement accuracy, surgical task recognition to identify trainees' ongoing surgical tasks in real time, and provide accuracy scores, performance scores, and text feedback. Our user study with 16 medical and 12 non-medical students shows that the personalized assessment helped them improve the distance to target by 36.6% and 31.7% accordingly during a more challenging case of EVD procedure. Furthermore, most of the students agreed that the scores and assessment feedback given by our system were helpful in improving their EVD performance.

Index Terms: Human-centered computing—Human computer interaction (HCI)—Interaction paradigms—Mixed / augmented reality; Applied computing—Life and medical sciences—Health care information systems;

*e-mail: sangjun.eom@duke.edu

†e-mail: tiffany.ma@duke.edu

‡e-mail: tianyi.hu@duke.edu

§e-mail: neha.vutakuri@duke.edu

¶e-mail: joshua.jackson@duke.edu

||e-mail: maria.gorlatova@duke.edu

1 INTRODUCTION

The placement of an external ventricular drain (EVD) is a routine neurosurgical procedure conducted to redirect cerebrospinal fluid out of the brain, providing relief from elevated intracranial pressure associated with conditions like hemorrhage or obstructive tumors. The conventional method for EVD placement involves freehand insertion, relying on external anatomic landmarks, often performed by junior neurosurgical trainees. Due to the number of repetitions needed to improve catheter placement accuracy [29], junior trainees are more likely to miss the target point of the foramen of Monro and make multiple passes of the catheter that can increase the risk of iatrogenic hemorrhages, infections, or damage to eloquent structures deep within the brain. Furthermore, distorted anatomies such as those caused by tumors or hemorrhages result in lower rates of accurately placed EVDs [19]. This requires junior trainees to evaluate their operative skills based on procedure-based skill assessments [1] to improve their performance during EVD training.

Recent work in Augmented Reality (AR)-assisted EVD procedures demonstrates that real-time visualization of the brain ventricles [31, 32], optimal location of entry points for catheter [10] and other contextual information about catheter placement [11, 37] in AR holograms during the procedure can help surgeons improve the EVD accuracy. However, junior trainees often need feedback about their surgical performance, such as catheter handling or real-time distance to the target point [6]. While computed tomography (CT) scanning is used to evaluate catheter placement accuracy [23], a manual assessment by senior faculty or residents using a checklist or global rating scale [25] is further required to evaluate the trainees' performance, such as catheter handling, in EVD training. This limits AR-assisted EVD placement training in providing personalized feedback to the trainees on their EVD performance.

To address these challenges, we designed, developed, and evaluated an AR-assisted neurosurgical training tool for EVD placement that integrates automatic segmentation of the brain ventricle, a sensing-integrated phantom, and surgical task recognition to pro-

vide feedback on the trainees' surgical performance. Based on the evaluation of the trainees' EVD placement performance, we provide accuracy and performance scores after the completed procedure with text feedback to instruct the trainees on improving catheter placement accuracy as shown in Fig. 1c. *To the best of our knowledge, our system is the first to provide personalized feedback about trainees' performance in EVD placement training.* We evaluated our system by conducting 224 trials of AR-assisted EVD placements with 16 medical and 12 non-medical students across normal and abnormal ventricular anatomies via an Institutional Review Board (IRB)-approved study. Furthermore, we open-sourced our research artifacts including the implementation of the threshold-based segmentation¹, the sensing-integrated phantom model¹, and the surgical task recognition². Our contributions are as follows:

- We developed an AR-assisted neurosurgical training tool for EVD placement that provides personalized feedback on trainees' EVD performance. We compute catheter placement accuracy by integrating camera sensors into the brain phantom model and identify trainees' ongoing surgical tasks using the phantom and catheter tracking from OptiTrack cameras and hand gesture data from HoloLens 2 in real time.
- We integrated a threshold-based segmentation of brain ventricular anatomy to create 8 different brain ventricular models for AR visualization. We open-sourced our threshold-based segmentation algorithm and sample datasets. Furthermore, we conducted the first evaluation of AR-assisted EVD placement training on automatically-segmented hemorrhage brain ventricles.
- We conducted more than 200 AR-assisted EVD placement trials with 16 medical and 12 non-medical students, providing personalized feedback on their performance. We demonstrated that, given assessment, both medical and non-medical students improved their catheter placement accuracy during more challenging EVD cases with hemorrhage ventricular anatomies by 36.6% and 31.7%, accordingly. Furthermore, our user study showed that most of the participants agreed that scores and feedback were helpful in improving their EVD placement performance.

2 RELATED WORK

AR-assisted EVD. The AR visualization of the intracranial anatomy enhances the surgeon's perception by improving the field of view in neurosurgery [13, 15, 26]. In the EVD placement procedure, the brain ventricle holograms provide guidance to surgeons for more accurate targeting of the foramen of Monro, hence improving catheter placement accuracy by reducing the distance to the target [5]. Prior work in AR-assisted EVD reports that the distance to the target was reduced over 40% [11, 24, 36] compared to trials without AR guidance (*freehand*). The AR guidance enhances the surgeons' perception intraoperatively; however, junior trainees and medical students often require feedback to improve their surgical performance in training [3, 4]. Therefore, we integrate sensing with a brain phantom to provide catheter placement accuracy to trainees in real time, and surgical task recognition to provide feedback on the trainees' catheter handling. *Furthermore, prior work has only been evaluated on standard brain ventricular anatomy* [11, 31, 36]. Hence, we introduce automated brain ventricular segmentation to conduct user studies in more challenging EVD insertion scenarios (e.g., hemorrhage).

Phantom Models with Sensing. Phantom models, widely employed for realistic simulation in surgical training, have traditionally been utilized for post-trial analysis, involving measurements of factors like positioning or the length of skin incisions [27]. However, there is potential for sensorized phantom models to offer real-time

¹<https://github.com/AREVD>

²<https://github.com/Duke-I3T-Lab/Hand-gesture-recognition>

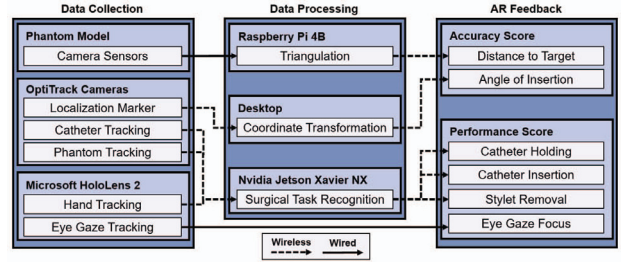


Figure 2: Overall architecture of our AR-assisted neurosurgical training system.

contextual information to trainees during surgical procedures. Previous studies have demonstrated the incorporation of various sensors within phantom models for data collection in surgical applications, such as electrodes within a liver phantom for detecting the position of an inserted needle [35], strain sensors for impact sensing [17], or wire coil sensors for measuring distances [33]. In EVD training, the conventional approach involves a post-analysis that includes scanning the brain phantom. This method proves to be both costly and time-consuming in evaluating catheter placement accuracy, making it impractical for providing immediate feedback to trainees after their trials [5, 34]; thus, trainees are not able to receive feedback on their catheter placement performance right after the trials. In contrast to related work primarily geared for post-analysis purposes [2], we propose a phantom model with embedded camera sensors, allowing for the real-time computation of catheter placement accuracy to enhance AR guidance and provide immediate assessment to trainees.

Surgical Task Evaluation. In surgical training, skill-based evaluation is used for trainees to learn and improve their surgical skills in medical training. This evaluation is commonly performed by senior faculty or residents observing the training session to evaluate the trainees' performance using criteria based on surgical tasks [22]. During the EVD procedure, the surgeon's operative skills such as completion time and handling of the catheter are important metrics for surgical performance evaluation [1]. Recent work demonstrates the use of deep learning-based methods [7, 12] to automate surgical evaluation. Prior work in surgical AR guidance system for neonatal endotracheal intubation (ETI) [40] is the most relevant to our work. However, it assesses ETI performance based on the motions of a laryngoscope and a manikin. It does not take full advantage of the AR headset, leaving the integrated hand gesture tracking untouched. Therefore, we propose the first AR-based training system for EVD that fuses the position and orientation of the phantom model and catheter with trainees' hand gesture data to provide feedback about EVD performance through real-time surgical task recognition.

3 OVERALL ARCHITECTURE

In this section, we describe the hardware setup of our AR-assisted system in Section 3.1, the automatic segmentation module with the evaluation of Dice coefficients across CT scans of various brain ventricular anatomies in Section 3.2, the automatic computation of catheter placement accuracy using a sensing-integrated phantom in Section 3.3, surgical task recognition module in Section 3.4, and the criteria for surgical performance evaluation in providing scores and feedback to trainees in Section 3.5.

3.1 Hardware Setup

Fig. 2 shows the hardware for data collection and data processing included in our setup. To overlay brain ventricle and catheter holograms in AR, we used six Flex 3 OptiTrack cameras with lens specs of 57.5 degrees in the field of view (FoV) to track the phantom model, the localization marker (following a state-of-the-art approach [11]), and the EVD catheter in real time, as shown in Fig. 1a. We used the localization marker to compute the transformation of world coordi-

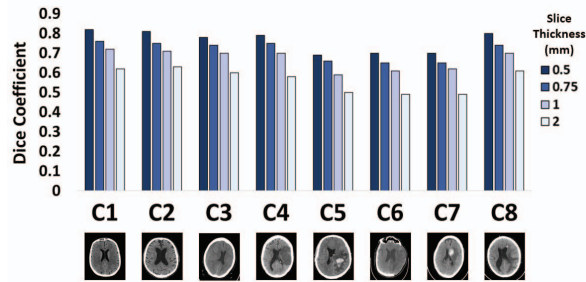


Figure 3: Segmentation results with varying slice thickness across various brain ventricular anatomies.

nates between the HoloLens 2 and the OptiTrack system to ensure high accuracy of image registration. The Transmission Control Protocol (TCP) is used as a data communication protocol between the HoloLens 2 and the desktop server that transmits optical marker location captured by the OptiTrack system, keeping the latency low with 12.32ms averaged over 15 trials. Using the transformed coordinate system, we created an AR app using Unity 2021.2.11f1 to display the brain ventricular anatomy and the guidance (i.e., text holograms for distance to target and angle of insertion).

On top of standard marker-based tracking for AR-assisted EVD systems, we added a sensing-integrated phantom model and a surgical task recognition module. We used a Raspberry Pi 4B to connect with camera sensors embedded inside the phantom model as a low-cost portable solution, sending the catheter placement accuracy results to HoloLens 2 via TCP. Furthermore, we used Nvidia Jetson Xavier NX to run the inference for surgical task recognition. Although we could have run the inference on the desktop server with GPU, due to the modularity and compatibility in collecting and processing the data, we ran the surgical task recognition module on Nvidia Jetson by receiving hand gesture data from HoloLens 2 and phantom and catheter tracking data from the OptiTrack cameras, then sending the inference results to HoloLens 2 via the User Datagram Protocol.

3.2 Automatic Segmentation of Brain Ventricle

Prior work in automatic segmentation algorithms demonstrates the use of deep learning [8, 21] or threshold-based [30] approaches to segment a brain model. This model can be automatically displayed in AR for surgical planning [23] or guidance in EVD procedures [31]. However, only the standard brain ventricular anatomy has been evaluated in state-of-the-art AR-assisted EVD [11, 31, 36]. Hence, we integrated threshold-based automatic segmentation with our AR system to segment non-standard brain ventricular anatomies, creating more challenging scenarios during EVD training. We used threshold-based segmentation over deep learning since we only need to segment the lateral ventricle and the foramen of Monro, as seen in Fig. 1c, which are distinct regions in the brain that can be segmented by a specific value of the threshold. We evaluated our segmentation by comparing dice coefficients across 8 different anatomies from an open-source anonymous patients' database [20].

Our automatic segmentation segments the brain ventricle using CT Digital Imaging and Communications in Medicine (DICOM) images as inputs, then visualizes the segmented ventricular model as a hologram in AR. We first created a 3D matrix and converted the raw voxel values to Hounsfield units through linear transformation for standardization across different scanners in the DICOM. We resampled the slices to standardize pixel and slice spacing to 1 cubic mm. We then set a threshold with a lower value of -10 and an upper value of 15 to capture values corresponding to the ventricle. We performed erosion to remove non-ventricle regions from the model. Because the ventricles shrink after erosion, we then dilated them to their original sizes. Once we obtained the 3D ventricular

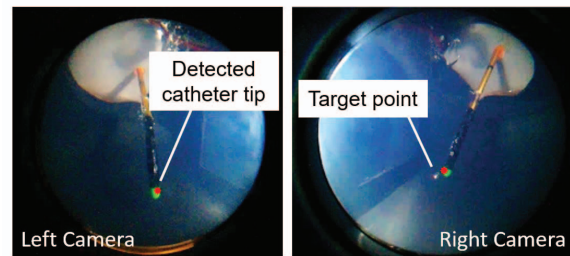


Figure 4: Captured images from the two camera sensors inside the phantom model with the catheter tip labeled as red points and the metal bead as the target location of the foramen of Monro.

model, we created a mesh representation of the model using the marching cubes algorithm. The mesh is constructed by connecting the ventricle vertices to create the ventricle faces. These data are then sent over from the edge server to HoloLens 2 via TCP. For validation, we segmented eight different brain CT scans (C1-C8) to evaluate the accuracy of the automatically-segmented ventricular model for different scenarios of the EVD procedures, as shown in Fig. 3. More information about our algorithm and the evaluation of datasets can be found in our repository¹.

3.3 Sensing-integrated Phantom

Existing phantom models used in AR-assisted EVD placements require CT scanning to evaluate catheter placement accuracy [5, 34]. To automate the EVD evaluation without the need for CT scanning, we designed a sensing-integrated phantom model to calculate the distance to the target in real time. Inspired by prior work [2], we used two stereo camera sensors attached to adjacent sides of our custom brain mold to capture image frames in real-time within the phantom skull, as shown in Fig. 1b. The locations of the catheter tip and foramen of Monro are identified and triangulated to calculate the distance to the target, as shown in Fig. 4. This real-time data is sent to a HoloLens 2 over a wireless network and displayed as textual guidance to allow trainees to understand the accurate depth of the catheter insertion.

Our custom-designed brain mold with camera sensors filled with colorless gelatin to imitate the realistic brain texture is shown in Fig. 1b. We used Arducam wide-angle OV5647 camera sensors that support a frame rate of up to 15fps at a resolution of 2592x1944. The camera sensors are equipped with an M12 lens that captures a wide angle of the image inside the brain mold to ensure the catheter tip is visible on the captured images in all trajectories. A Raspberry Pi 4B is used to obtain captured images from the camera sensors for triangulation and send the results to HoloLens 2 over a wireless network. Additionally, we placed an LED at the corner of the mold between the two camera sensors to enhance the brightness of the captured images. We painted the catheter tip with a green polish to identify its location through image filtering. We filled the inside of the mold using a transparent gelatin solution to provide realistic textural feedback to the trainees.

We evaluated the accuracy of the distance to the target from our triangulation algorithm by comparing the computed distance to the ground truth obtained from CT scans using a Nikon XTH 225 ST, a high-resolution micro X-ray CT scanner. We used a 3D graphical software, Avizo, to render the 3D volume and measure the Euclidean distance between the catheter tip and the foramen of Monro for ground truth. The average Euclidean distance error was 0.386mm (standard deviation, $SD = 0.349$ mm). The latency for HoloLens 2 to receive the distance to the target from Raspberry Pi 4B was 116.8ms, on average over 15 trials, at an image resolution of 320x240. More information about our implementation can be found in our repository¹.

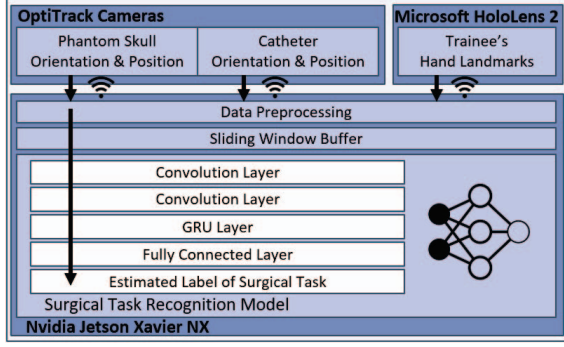


Figure 5: Overall pipeline of the surgical task recognition module.

3.4 Surgical Task Recognition

We formulated surgical task recognition as a multivariate time-series classification problem. Our model utilizes data aggregated from various sensors over time, formatted as a multivariate time series, to predict the category of the surgical task being performed. We used five features: (1) the orientation and (2) position of the phantom model, (3) the orientation and (4) the position of the catheter, and (5) the trainee's hand gestures, which include 25 coordinates encompassing hand joints, fingertips, and the wrist. These features enable the classification of four essential EVD surgical tasks: catheter holding, catheter insertion, stylet removal, and miscellaneous movements. The latter category encompasses actions not included in the first three. Fig. 5 shows the overall pipeline of our module. More details about our model can be found in our repository².

Data Collection and Labeling. We used OptiTrack cameras to track the orientations and positions of both the phantom model and the catheter and MRTK2 on HoloLens 2 to capture the trainee's hand gestures. We implemented a sliding window buffer to store the last two seconds of the five features to infer ongoing surgical tasks. To train our model, we collected the five features from 10 trainees. Each trainee was instructed on the EVD procedures and subsequently performed 10 EVD trials. During these trials, we manually labeled the collected feature data by observing trainees' gestures. To ensure the diversity of the dataset, the trainees were not instructed on specific gestures to perform the EVD. On average, each trainee completed the EVD procedure in 30.04 seconds.

Model Architecture and Data Preprocessing. Prior work shows the effectiveness of a hybrid model architecture, where convolutional layers extract implicit representations among different features, and gated recurrent units (GRU) capture the temporal patterns in time series data [16,28]. Inspired by this, we implemented a hybrid model architecture, as shown in Fig 5. In our data preprocessing phase, to ensure our model is invariant to the phantom model's absolute position and hand landmarks' absolute coordinates, we repositioned the phantom model's location as the origin and centered the hand landmarks relative to the trainee's wrist position. This preprocessed data is then stored in a sliding window buffer to keep the last 2 seconds of features to create time series data. We also adopted time series augmentation, e.g., adding jitters and window warping [18] in the data preprocessing, to generate more data and improve the model generalizability.

Evaluation. We evaluated our model using leave-one-out cross-validation (LOOCV), which takes each trainee's data as the validation set and the rest nine trainees' data as the training set. With LOOCV, we can evaluate the model performance on unseen individual data for all ten trainees in our dataset. The confusion matrix in Fig. 6 shows the model's classification accuracies for different surgical tasks: 65.5% for catheter holding, 70.6% for catheter insertion, 81.4% for stylet removal, and 90.1% for miscellaneous gestures. Furthermore, the performance metrics of our model, as shown in

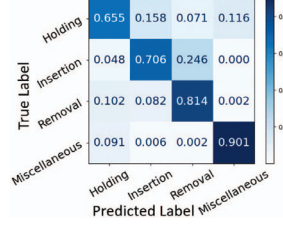


Figure 6: Confusion matrix of surgical task recognition.

Table 1: Evaluation of the surgical task recognition using different evaluation metrics.

Evaluation Metrics	Evaluation Scores
Accuracy	0.760
Precision	0.786
Recall	0.759
F1	0.765

Table 1, including accuracy, precision, recall, and F1 score, are all exceeding 0.75. This demonstrates the model's robust performance in surgical task recognition. On average, the HoloLens 2 took 116.67ms to collect features, send them to the Nvidia Jetson for inference, and then receive the results back, across 10 trials. This latency was adequate for calculating the performance score for the feedback provided to the trainees.

3.5 Surgical Performance Evaluation

The purpose of surgical performance assessment is to provide constructive feedback to trainees and improve specific technical skills during training. More objective assessments of surgical performance using checklists or global rating scales [25] have been used as conventional methods for providing reliable and valid assessments to the trainees. Motivated by this, we automated the assessment of EVD performance for our system to provide feedback to the trainees on their EVD performance after the EVD procedure is completed. The feedback comprises three components: accuracy score, performance score, and text feedback. The accuracy score is an evaluation of the EVD placement accuracy, and the performance score is an evaluation of the trainees' catheter handling and eye gaze focus on AR contextual guidance. In coordination with an expert neurosurgeon, we assigned 5 different scores for each criterion based on the percentiles of AR-assisted EVD performance results from our prior training with medical students [11]. This aligns with the target group of participants in our user study (in Section 4), allowing us to evaluate their EVD accuracy and performance. Lastly, we generated text feedback based on the criteria of low accuracy and performance scores to provide instructions on how to improve the scores.

Accuracy Score. The distance to the target from the sensing-integrated phantom module, d , and the angle of the catheter, θ , are important metrics for the optimal EVD catheter trajectory [9, 39]. Therefore, we formulated the score with these two criteria obtained from the sensing-integrated phantom and the OptiTrack tracking of the EVD catheter. The total accuracy score, which is calculated on a scale of 0-10, assigns equal weight to both the distance to the target (d) and the angle of the catheter insertion (θ), as outlined in Eq. 1.

$$S_a = \sum (0.5\theta + 0.5d) \quad (1)$$

Performance Score. Catheter placement accuracy is not the only criterion that matters in performance evaluation during EVD training. Trainees also get evaluated on other criteria such as the handling of the EVD catheter or knowledge of the procedure [1, 14]. Hence, we developed a performance score that comprises four criteria of the time spent on each surgical task (i.e., c_h for catheter holding, c_i for catheter insertion, and c_r for stylet removal), as shown in Eq. 2, and the percentage of eye gaze focus on AR guidance, g . The total performance score is out of 10 and the weight is evenly distributed between each criterion, as shown in Eq. 3.

$$S(c_h, c_i, c_r) = \sum (0.25c_h + 0.25c_i + 0.25c_r) \quad (2)$$

$$S_p = \sum (S(c_h, c_i, c_r) + 0.25g) \quad (3)$$

Assessment Feedback. During the conventional assessment, trainees receive either verbal or written instructions based on the checklist or scores determined by the senior faculty [25]. Hence, in addition to the accuracy and performance scores, our system provided a textual assessment to the trainees in their AR view about the instructions on how to improve the scores after each trial. For a low score on the distance to the target, the trainee is instructed to reduce the distance between the catheter tip and the target point, and for a low score on the angle of the catheter insertion, the trainee is instructed to adjust the angle of the catheter insertion to be closer to 90 degrees. Furthermore, for the time spent on each surgical task, the trainee is instructed to reduce the time spent on surgical tasks with low scores. For a low score on the percentage of eye gaze focus on AR guidance, the trainee is instructed to utilize the AR guidance more. This assessment was only visualized in AR when each criterion of the score was lower than the maximum score of 10.

4 USER STUDY DESIGN

In our user study, each participant was asked to perform eight AR-assisted EVD trials; the first four EVD trials were performed on standard ventricular anatomy (SVA) and the last four EVD trials were performed on a hemorrhage ventricular anatomy (HVA). HVA had an asymmetrical shape that created challenges in estimating the location of the foramen of Monro and an optimal trajectory. The first 2 trials were conducted on SVA without assessment, followed by another 2 trials with assessment, then 2 trials were conducted on HVA without assessment, followed by another 2 trials with assessment. We recruited 16 medical and 12 non-medical students by emailing medical schools in our metropolitan area. Our user study was approved by our institution's IRB.

Participant Selection. Out of 16 medical students, 9 were year 1 students and 7 were year 3-4 students. Out of 12 non-medical students, 7 were graduate students and 5 were undergraduate students. The age range of all participants was 20-50 years ($MEAN = 25.1$, $SD = 6.04$). Among all the participants, half of them were male and the other half of them were female. None of the participants have performed the EVD procedure; however, 3 of the medical students have participated in an EVD simulation in the past. Among all the participants, 4 of them use an AR headset frequently (more than once a week) and 3 of them use it infrequently (less than once a week). 13 of them had worn an AR headset once or twice, and 8 of them had never worn an AR headset before. None of the participants had any eyesight-related conditions such as strabismus or colorblindness.

AR-assisted EVD Trials. We used two different ventricular anatomies of anonymous patients to vary the level of difficulty in AR-assisted EVD placement. The SVA in Fig. 7a presents a symmetrical brain ventricle with the foramen of Monro in a red hologram, and the HVA in Fig. 7b presents an asymmetrical shape.

The steps for our AR-assisted EVD trials were as follows: participants first watched the instructional video about freehand EVD, recorded by a neurosurgeon with 9 years of clinical experience³. The participants were not instructed on specific gestures for holding the catheter during the procedure. The eye calibration on HoloLens 2 was performed to ensure the rendering of holograms at accurate locations and the collection of accurate eye gaze data of the participants. Upon the initialization of the AR app on HoloLens 2, the participants detected the localization marker to enable AR visualization, allowing participants to start the catheter insertion with constant guidance of displaying the distance to the target and the angle of insertion. During the procedure, the scores and assessment feedback were constantly updated based on the participants' hand movements. Once the participants placed the catheter at the estimated target point, they removed the inner stylet, and final scores and assessment feedback were displayed in front of their AR view.

³The instructional video is provided at <https://youtu.be/wCKOd4m7jK4>

Table 2: Post-experiment survey questions.

	Questions
Q1	The assessment feedback was helpful in identifying areas of improvement for my EVD performance.
Q2	I aimed to address the assessment feedback provided to me in subsequent trials.
Q3	Overall, the assessment feedback was helpful in learning the EVD procedure.
Q4	Overall, the assessment feedback was helpful in improving my EVD performance.
Q5	The assessment feedback was more helpful in the following scenarios.
Q6	The accuracy score was helpful in improving my EVD accuracy.
Q7	The performance score was helpful in handling the EVD catheter.
Q8	Overall, the scores were helpful in learning the EVD procedure.
Q9	Overall, the scores were helpful in improving my EVD performance.
Q10	The scores were more helpful in the following scenarios.
Q11	I didn't feel tired or fatigued at some point during the experiment.
Q12	The hologram visualization was robust without significant lagging.
Q13	The hologram visualization did not obstruct my view.
Q14	If you have any other comments or feedback about your experience, please write below:

Survey Questions. The pre-experiment and post-experiment surveys were given to each participant to fill out before and after the user study. In the pre-experiment survey, we asked the participants about prior experiences in AR and EVD. Table 2 shows the list of post-experiment survey questions in three different categories.

We asked participants a set of questions in three categories: feedback, score, and AR experience. For the feedback category, we asked the participants on a five-point Likert scale if the text feedback was helpful during the AR-assisted EVD trial (Q1-Q4). We also asked the participants to select whether the text feedback was helpful in the SVA, the HVA, both, or neither (Q5). For the score category, we asked the participants on a five-point Likert scale if the accuracy and performance scores helped them learn and improve EVD accuracy during the AR-assisted EVD trials (Q6-Q9). Similarly, we also asked the participants to select whether the scores were helpful in the SVA, the HVA, in both cases, or unhelpful in both cases (Q10). For the AR experience category, we asked the participants on a five-point Likert scale if the system was robust without lagging, drift, and obstruction of view, and if the participants experienced fatigue (Q11-Q13). At the end of the survey, we asked the participants to leave any open-ended feedback about the overall experience (Q14).

5 RESULTS

5.1 EVD Accuracy

We evaluated the levels of accuracy improvements between the first and second trials, as well as between the third and fourth trials on both SVA and HVA, as shown in Table 3. On SVA, the medical students improved their EVD accuracy similarly for both trials with assessment and without assessment. The accuracy improved from 5.24mm ($SD = 3.00mm$) to 3.37mm ($SD = 2.53mm$; improvement rate: 35.6%) with assessment and from 5.53mm ($SD = 4.81mm$) to 3.41mm ($SD = 3.86mm$; improvement rate: 38.4%) without assessment. However, the non-medical students' accuracy improvement was higher for trials without assessment than the trials with assessment. The accuracy improved from 20.9mm ($SD = 12.7mm$) to 10.7mm ($SD = 9.99mm$; improvement rate: 48.8%) without assessment and from 14.3mm ($SD = 11.6mm$) to 10.6mm ($SD = 7.94mm$; improvement rate: 25.4%) with assessment. These differences in improvements between medical and non-medical groups were statistically significant ($p < 0.0001$ for trials with assessment and $p = 0.0003$ for trials without assessment) using a two-tailed

Table 3: Results of AR-assisted EVD trials on SVA and HVA.

Level of Expertise	Medical Students (n=16)				Non-medical Students (n=12)			
Assessment	No		Yes		No		Yes	
Trials on SVA	1	2	3	4	1	2	3	4
Accuracy (mm)	5.24	3.37	5.53	3.41	20.9	10.7	14.3	10.6
Angle (deg)	91.5	90.5	92.1	94.6	92.2	93.1	95.0	93.1
Total Time (s)	129.6	101.2	76.8	76.9	102.1	102.7	96.7	95.4
Trials on HVA	1	2	3	4	1	2	3	4
Accuracy (mm)	5.21	5.05	5.24	3.32	12.3	11.4	10.6	7.25
Angle (deg)	90.9	93.4	92.6	92.5	95.3	91.2	93.3	93.0
Total Time (s)	76.5	90.5	89.0	92.3	72.7	52.9	58.0	54.8

paired Student's t-test with equal variance. We hypothesize that this was due to the learning curve of the EVD procedure. Non-medical students lacked an understanding of the procedure due to the absence of a medical background. This resulted in large accuracy improvements between the two trials on the SVA without assessment, which were the first two trials in the user study. However, there was little difference in the accuracy improvements of medical students (35-38%) for trials performed on the SVA with regard to assessment.

On the contrary, the average distance to the target did not show significant improvements for trials on HVA without assessment for both medical and non-medical students. The accuracy stayed consistent from 5.21mm ($SD = 3.73$ mm) to 5.05mm ($SD = 3.71$ mm; improvement rate: 3.08%) for medical students, and from 12.3mm ($SD = 10.6$ mm) to 10.6mm ($SD = 11.4$ mm; improvement rate: 7.72%) for non-medical students. However, with assessment, the average distance to the target was improved for both medical and non-medical students. The accuracy was improved from 5.24mm ($SD = 3.41$ mm) to 3.32mm ($SD = 3.48$ mm; improvement rate: 36.6%) for medical students, and from 10.62mm ($SD = 9.97$ mm) to 7.25mm ($SD = 6.25$ mm; improvement rate: 31.7%) for non-medical students. These improvements were found to be statistically significant using a two-tailed Student's t-test ($p = 0.0182$) with equal variance. We believe that providing assessment to trainees has the potential to improve catheter placement accuracy in more challenging EVD scenarios such as HVA, which poses difficulties in identifying target locations and determining optimal trajectories. However, the learning curve over the trials remains a limitation of our study due to the lack of randomization in the trial order.

5.2 Total Completion Time

For the trials on SVA, both medical students and non-medical students completed the procedure with assessment in a shorter time (76.8s and 76.9s for medical; 96.7s and 95.4s for non-medical students) than without assessment (129.6s and 101.2s for medical; 102.1s and 102.7s for non-medical students) on average. We believe that the assessment feedback helped students identify areas of improvement, resulting in less time spent determining trajectories and completing the procedure. However, for the trials on HVA, medical students spent a longer time completing the procedure with assessment (89.0s and 92.3s) than without assessment (76.5s and 90.5s), while non-medical students spent a shorter time completing the procedure with assessment (58.0s and 54.8s) than without assessment (72.7s and 52.9s) on average. We hypothesize that due to the higher levels of difficulty in determining trajectories for trials on HVA, medical students spent more time addressing the assessment feedback provided to them and changing their approaches to improve their performance.

5.3 Survey Response

Our post-experiment survey responses are summarized in Fig. 8. We define *positivity rate* as the percentage of participants' responses

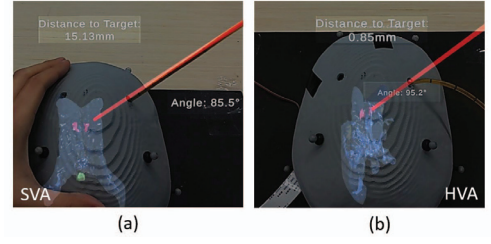


Figure 7: Ventricular holograms used in our study: symmetrical anatomy for the SVA (a) and asymmetrical anatomy for the HVA (b), posing different levels of challenges.

in the “strongly agree” and “agree” categories. The participants' free-text responses are quoted with the participant number, P .

AR feedback. The participants mostly agreed that the text feedback was helpful in identifying areas of improvement (positivity rate: 71.4%) and improving the overall EVD performance (positivity rate: 75.0%). 92.9% of the participants agreed or strongly agreed that they aimed to address the assessment feedback in subsequent trials. However, only 53.6% of the participants agreed or strongly agreed that the assessment feedback was helpful in learning the EVD procedure. Overall, most of the participants agreed that the assessment feedback was equally helpful for both cases of AR-assisted EVD trials on SVA and HVA (74.1%) while a lower percentage of participants agreed for only SVA (18.5%) and both equally unhelpful (3.7%). This aligns with the high percentage of accuracy improvements for AR-assisted trials with assessment on both SVA and HVA seen in Table 3. The participants provided additional feedback that real-time assessment during the surgical procedure would be useful (P11, P12, P17). Two participants felt that the assessment was very helpful (P8, P20). In future work, we will provide assessment feedback during the EVD procedure based on the real-time evaluation of the participants' surgical performance.

Scores. Similarly, the participants agreed that scores were helpful overall in learning the EVD procedure (positivity rate: 64.3%) and improving the EVD performance (positivity rate: 71.4%). However, lower positivity rates were recorded on the helpfulness of each category of the scores: accuracy score in improving the EVD accuracy (53.6%) and performance score in handling the EVD catheter (46.4%). The participants who provided additional feedback felt that more context about the criteria of scores would be helpful (P2, P8, P12). In the future, we plan to provide the breakdown of their scores in each criterion. Overall, most of the participants agreed that the scores were equally helpful for both cases of AR-assisted EVD trials on SVA and HVA (74.1%) while a low percentage of the participants agreed that they were only helpful for SVA (7.4%) and for HVA (7.4%). Only 11.1% of the participants reported that the scores were equally unhelpful for both cases of SVA and HVA, which was higher than the percentage reported for assessment feedback (3.7%). We think this was due to assessment feedback providing detailed instructions on how to improve their performance, while scores only provided them with a numerical evaluation of their performance.

AR Experience. Overall, the participants appreciated that the AR experience provided hologram visualizations without obstructing their view (positivity rate: 85.7%), that there was no significant lagging (positivity rate: 82.1%), and that they did not feel tired or fatigued during the trials (positivity rate: 60.7%). However, a significant number of the participants stated in the additional feedback that they observed “the drift or jump of catheter hologram” (P2, P7) and “misalignment of the catheter when rotating the stylet” (P6). Additionally, some participants felt that “the catheter hologram was difficult to align” (P8, P15) and that the “blocking of optical mark-

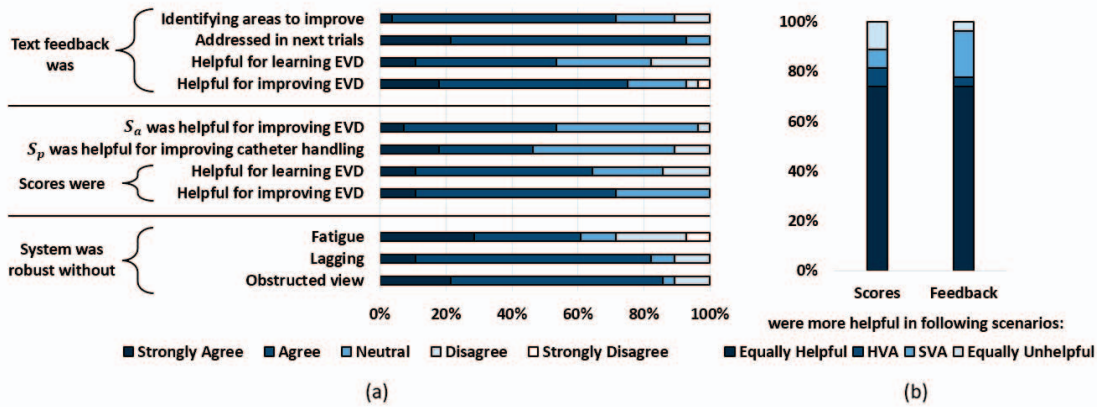


Figure 8: Survey responses on a five-point Likert scale for (a) categories of AR feedback, scores, and AR experiences and (b) comparison on the helpfulness of AR feedback and scores between EVD scenarios on SVA and HVA.

ers might have impacted the results” (P10). While the overall AR experience received high positivity rates from the participants, the occlusion of optical markers and more robust tool tracking remain as future work to improve the robustness of our system.

Additional Feedback. In the open-ended feedback, participants were positive about the AR system as a training tool for practicing EVD and improving their performance. The participants thought that the system was “a cool technology” (P6, P7, P17), “great experience” (P2), “robust enough to allow me to understand the procedure” (P13), and “helpful practice” (P6). The participants also stated that “AR made the procedure richer and engaging” (P3), “the distance and angle guidance assisted in estimating the target point” (P19, P21, P22, P26), and “this was a lot of fun” (P7, P9). The overall positive feedback from all participants demonstrates the potential of our AR-assisted EVD system as a future neurosurgical training tool used for the assessment of medical students’ EVD performance. Our system could potentially complement traditional skill assessments that use a checklist or a global scale rating [1, 25], making the assessment more efficient and interactive, and enabling trainees to practice on various types of challenging brain ventricular anatomies.

6 DISCUSSION AND FUTURE WORK

Our AR system used a stereo camera setup for triangulating 3D coordinates and calculating the distance from the catheter tip to the foramen of Monro. Due to the characteristics of the cameras’ fish-eye lens, the captured images used for both calibration and triangulation were distorted to a higher degree than traditional cameras. This led to higher errors in triangulation around the edges of the images. We plan to improve the current triangulation method by using a larger number of sample images during calibration as well as experimenting with different calibration methods that suit the qualities of fish-eye images. Additionally, our participants reported a mismatch between the distance values reported by the textual guidance and the catheter hologram visualization. We plan to calculate the transformations necessary to obtain the location of the catheter tip in the same plane as the HoloLens 2 coordinate system, which will improve the accuracy of the catheter hologram alignment.

We used the HoloLens 2’s built-in hand gesture tracking to collect the trainees’ hand movements during the AR-assisted EVD procedure. However, due to the limited FoV, HoloLens 2 loses track of the hand gestures when the trainees’ hands are out of the camera view. This often happens during the stylet removal when the trainees grab the top of the stylet to pull it out of the catheter. For future work, we plan to enhance hand gesture tracking by adding other external camera sensors to the AR headset to run hand gesture detection separately or using the OptiTrack system to track the hand gestures

with optical markers [38]. This could potentially be enhanced by future AR headsets that have a built-in camera with better FoV.

One of the limitations of our current AR-based assessment system is that it only provides feedback on trainees’ performance in terms of accuracy and time spent for each surgical phase. However, these metrics do not capture the trainees’ overall behavior and attentional states, which can also affect their surgical performance. For example, trainees may perform poorly due to lack of focus, high stress, or fatigue. Hence, we believe that providing feedback about trainees’ attentional states such as whether the trainee is focused on the task or not can improve our AR-based assessment in surgical training. In future work, we will use eye tracking to analyze the gaze-based attention pattern of trainees to enhance the feedback and alert them in real time.

7 CONCLUSIONS

This paper presents the first AR-based neurosurgical training tool for EVD that provides personalized feedback on surgeons’ performance to guide them to improve catheter placement accuracy during EVD training. We automated the segmentation of brain ventricular anatomy to enable the AR system to display the ventricular hologram and provide catheter placement accuracy using a sensing-integrated brain phantom in real time. Our evaluation with 16 medical and 12 non-medical students demonstrated that the students reduced the distance to target by 36.6% and 31.7% accordingly for trials on HVA, which posed more challenges due to asymmetrical shape. In future work, we will evaluate our system with other brain ventricular anatomies for more challenging EVD scenarios, enhance the calibration for triangulation, and use additional cameras with better FoV to enhance hand gesture tracking during the surgical procedure.

ACKNOWLEDGMENTS

We thank all participants in our user study. We also thank Shervin Rahimpour for his contributions to this project. This work was supported in part by NSF grants CNS-2112562 and CNS-1908051, NSF CAREER Award IIS-2046072, and by a Thomas Lord Educational Innovation Grant.

REFERENCES

- [1] G. Aldave, D. Hansen, V. Briceño, T. G. Luerssen, and A. Jea. Assessing residents’ operative skills for external ventricular drain placement and shunt surgery in pediatric neurosurgery. *Journal of Neurosurgery: Pediatrics*, 19(4):377–383, 2017.
- [2] E. Azimi, Z. Niu, M. Stiber, N. Greene, R. Liu, C. Molina, J. Huang, C.-M. Huang, and P. Kazanzides. An interactive mixed reality platform

- for bedside surgical procedures. In *Proc. MICCAI*, pp. 65–75. Springer, 2020.
- [3] N. Bagher Zadeh Ansari, É. Léger, and M. Kersten-Oertel. VentroAR: An augmented reality platform for ventriculostomy using the microsoft hololens. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, pp. 1–9, 2022.
- [4] M. Benmahdjoub, A. Thabit, M.-L. C. van Veelen, W. J. Niessen, E. B. Wolvius, and T. Van Walsum. Evaluation of AR visualization approaches for catheter insertion into the ventricle cavity. *IEEE Transactions on Visualization and Computer Graphics*, 29(5):2434–2445, 2023.
- [5] M. T. Bounajem, B. Cameron, K. Sorensen, R. Parr, W. Gibby, G. Prashant, J. J. Evans, and M. Karsy. Improved accuracy and lowered learning curve of ventricular targeting using augmented reality—phantom and cadaveric model testing. *Neurosurgery*, pp. 10–1227, 2022.
- [6] L. Chen, T. W. Day, W. Tang, and N. W. John. Recent developments and future challenges in medical mixed reality. In *Proc. IEEE ISMAR*, 2017.
- [7] M. Doughty, K. Singh, and N. R. Ghugre. SurgeonAssist-Net: Towards context-aware head-mounted display-based augmented reality for surgical guidance. In *Proc. MICCAI*. Springer, 2021.
- [8] G. Du, X. Cao, J. Liang, X. Chen, and Y. Zhan. Medical image segmentation based on U-net: A review. *Journal of Imaging Science and Technology*, 2020.
- [9] C. V. Eisenring, F. Burn, M. Baumann, L. H. Stieglitz, R. A. Kockro, J. Beck, A. Raabe, and M. F. Oertel. sEVD—smartphone-navigated placement of external ventricular drains. *Acta Neurochirurgica*, 162:513–521, 2020.
- [10] S. Eom, T. S. Ma, N. Vutakuri, T. Hu, A. P. Haskell-Mendoza, D. A. Sykes, M. Gorlatova, and J. Jackson. Accuracy of routine external ventricular drain placement following a mixed reality-guided twist-drill craniostomy. *Neurosurgical Focus*, 56(1):E11, 2024.
- [11] S. Eom, D. Sykes, S. Rahimpour, and M. Gorlatova. NeuroLens: Augmented reality-based contextual guidance through surgical tool tracking in neurosurgery. In *Proc. IEEE ISMAR*, 2022.
- [12] C. R. Garrow, K.-F. Kowalewski, L. Li, M. Wagner, M. W. Schmidt, S. Engelhardt, D. A. Hashimoto, H. G. Kenngott, S. Bodenstedt, S. Speidel, et al. Machine learning for surgical phase recognition: A systematic review. *Annals of surgery*, 273(4):684–693, 2021.
- [13] C. Gsaxner, J. Li, A. Pepe, Y. Jin, J. Kleesiek, D. Schmalstieg, and J. Egger. The HoloLens in medicine: A systematic review and taxonomy. *Medical Image Analysis*, p. 102757, 2023.
- [14] C. Hadley, S. K. Lam, V. Briceño, T. G. Luerssen, and A. Jea. Use of a formal assessment instrument for evaluation of resident operative skills in pediatric neurosurgery. *Journal of Neurosurgery: Pediatrics*, 16(5):497–504, 2015.
- [15] J. Haemmerli, A. Davidovic, T. R. Meling, L. Chavaz, K. Schaller, and P. Bijlenga. Evaluation of the precision of operative augmented reality compared to standard neuronavigation using a 3D-printed skull. *Neurosurgical Focus*, 50(1):E17, 2021.
- [16] D. Hughes, A. Krauthammer, and N. Correll. Recognizing social touch gestures using recurrent and convolutional neural networks. In *Proc. IEEE ICRA*, 2017.
- [17] J. Hughes, P. Maiolino, T. Nanayakkara, and F. Iida. Sensorized phantom for characterizing large area deformation of soft bodies for medical applications. In *Proc. IEEE RoboSoft*, 2020.
- [18] B. K. Iwana and S. Uchida. An empirical survey of data augmentation for time series classification with neural networks. *PLOS One*, 16(7):e0254841, 2021.
- [19] B. A. P. Jayasekera, A. Al-Mousa, A. Shtaya, and E. Pereira. Freehand external ventricular drain insertion-is there a learning curve? *Surgical Neurology International*, 12, 2021.
- [20] Kaggle. RSNA intracranial hemorrhage detection. <https://www.kaggle.com/c/rsna-intracranial-hemorrhage-detection>, 2023.
- [21] M. Klimont, M. Flieger, J. Rzeszutek, J. Stachera, A. Zakrzewska, and K. Jończyk-Potoczna. Automated ventricular system segmentation in paediatric patients treated for hydrocephalus using deep learning methods. *BioMed Research International*, 2019, 2019.
- [22] R. J. Koehler, S. Amsdell, E. A. Arendt, L. J. Bisson, J. P. Braman, A. Butler, A. J. Cosgarea, C. D. Harner, W. E. Garrett, T. Olson, et al. The arthroscopic surgical skill evaluation tool (ASSET). *The American Journal of Sports Medicine*, 41(6):1229–1237, 2013.
- [23] C. Kunz, M. Hlaváč, M. Schneider, A. Pala, P. Henrich, B. Jickeli, H. Wörn, B. Hein, R. Wirtz, and F. Mathis-Ullrich. Autonomous planning and intraoperative augmented reality navigation for neurosurgery. *IEEE Transactions on Medical Robotics and Bionics*, 3(3):738–749, 2021.
- [24] Y. Li, X. Chen, N. Wang, W. Zhang, D. Li, L. Zhang, X. Qu, W. Cheng, Y. Xu, W. Chen, et al. A wearable mixed-reality holographic computer for guiding external ventricular drain insertion at the bedside. *Journal of Neurosurgery*, 131(5):1599–1606, 2018.
- [25] J. Martin, G. Regehr, R. Reznick, H. Macrae, J. Murnaghan, C. Hutchison, and M. Brown. Objective structured assessment of technical skill (OSATS) for surgical residents. *British Journal of Surgery*, 84(2):273–278, 1997.
- [26] A. Meola, F. Cutolo, M. Carbone, F. Cagnazzo, M. Ferrari, and V. Ferrari. Augmented reality in neurosurgery: A systematic review. *Neurosurgical Review*, 40(4):537–548, 2017.
- [27] A. Müns, J. Meixensberger, and D. Lindner. Evaluation of a novel phantom-based neurosurgical training system. *Surgical Neurology International*, 5, 2014.
- [28] X. A. Nguyen, D. Ljuhar, M. Pacilli, R. M. Nataraja, and S. Chauhan. Surgical skill levels: Classification and analysis using deep neural network model and motion signals. *Computer methods and programs in biomedicine*, 177:1–8, 2019.
- [29] H. Ofoma, B. Cheaney II, N. J. Brown, B. V. Lien, A. S. Himstead, E. H. Choi, S. Cohn, J. K. Campos, and M. Y. Oh. Updates on techniques and technology to optimize external ventricular drain placement: A review of the literature. *Clinical Neurology and Neurosurgery*, 213:107126, 2022.
- [30] D. D. Patil and S. G. Deore. Medical image segmentation: A review. *International Journal of Computer Science and Mobile Computing*, 2(1):22–27, 2013.
- [31] M. Schneider, C. Kunz, A. Pal’a, C. R. Wirtz, F. Mathis-Ullrich, and M. Hlaváč. Augmented reality-assisted ventriculostomy. *Neurosurgical Focus*, 50(1):E16, 2021.
- [32] S. Skyrman, M. Lai, E. Edström, G. Burström, P. Förander, R. Homan, F. Kor, R. Holthuizen, B. H. Hendriks, O. Persson, et al. Augmented reality navigation for cranial biopsy and external ventricular drain insertion. *Neurosurgical Focus*, 51(2):E7, 2021.
- [33] T. L. Steffensen, M. Auflem, H. N. Vestad, and M. Steinert. Embedded soft inductive sensors to measure arterial expansion of tubular diameters in vascular phantoms. *IEEE Sensors Journal*, 22(7):7240–7247, 2022.
- [34] B. L. Tai, D. Rooney, F. Stephenson, P.-S. Liao, O. Sagher, A. J. Shih, and L. E. Savastano. Development of a 3D-printed external ventricular drain placement simulator. *Journal of Neurosurgery*, 123(4):1070–1076, 2015.
- [35] X. Tan, D. Li, M. Jeong, T. Yu, Z. Ma, S. Afat, K.-E. Grund, and T. Qiu. Soft liver phantom with a hollow biliary system. *Annals of Biomedical Engineering*, pp. 1–11, 2021.
- [36] B. van Amsterdam, M. J. Clarkson, and D. Stoyanov. Gesture recognition in robotic surgery: A review. *IEEE Transactions on Biomedical Engineering*, 68(6), 2021.
- [37] F. Van Gestel, T. Frantz, C. Vannerom, A. Verhellen, A. G. Gallagher, S. A. Elprama, A. Jacobs, R. Buyl, M. Bruneau, B. Jansen, et al. The effect of augmented reality on the accuracy and learning curve of external ventricular drain placement. *Neurosurgical Focus*, 51(2):E8, 2021.
- [38] P. Woulfe, F. J. Sullivan, L. Byrne, A. Doyle, W. Kam, M. Martyn, and S. O’Keeffe. Optical fibre based real-time measurements during an LDR prostate brachytherapy implant simulation: Using a 3D printed anthropomorphic phantom. *Scientific Reports*, 11(1):1–8, 2021.
- [39] Z. Yi, B. He, Z. Deng, Y. Liu, S. Huang, and W. Hong. A virtual reality-based data analysis for optimizing freehand external ventricular drain insertion. *International Journal of Computer Assisted Radiology and Surgery*, 16:269–276, 2021.
- [40] S. Zhao, X. Xiao, Q. Wang, X. Zhang, W. Li, L. Soghier, and J. Hahn. An intelligent augmented reality training framework for neonatal endotracheal intubation. In *Proc. IEEE ISMAR*, 2020.