big data and cognitive computing

MDPI

Article

Exploration of Feature Representations for Predicting Learning and Retention Outcomes in a VR Training Scenario

Alec G. Moore [1,*], Ryan P. McMahan [1,*] and Nicholas Ruozzi [2]

[1] Department of Computer Science, College of Engineering and Computer Science, University of Central Florida, Orlando, FL 32816, USA
[2] Department of Computer Science, Erik Jonsson School of Engineering and Computer Science, The University of Texas at Dallas, Richardson, TX 75080, USA; nicholas.ruozzi@utdallas.edu
* Correspondence: agm@knights.ucf.edu (A.G.M.); rpm@ucf.edu (R.P.M.)

Abstract: Training and education of real-world tasks in Virtual Reality (VR) has seen growing use in industry. The motion-tracking data that is intrinsic to immersive VR applications is rich and can be used to improve learning beyond standard training interfaces. In this paper, we present machine learning (ML) classifiers that predict outcomes from a VR training application. Our approach makes use of the data from the tracked head-mounted display (HMD) and handheld controllers during VR training to predict whether a user will exhibit high or low knowledge acquisition, knowledge retention, and performance retention. We evaluated six different sets of input features and found varying degrees of accuracy depending on the predicted outcome. By visualizing the tracking data, we determined that users with higher acquisition and retention outcomes made movements with more certainty and with greater velocities than users with lower outcomes. Our results demonstrate that it is feasible to develop VR training applications that dynamically adapt to a user by using commonly available tracking data to predict learning and retention outcomes.

Keywords: virtual reality; machine learning; intelligent tutoring systems; training

Citation: Moore, A.G.; McMahan, R.P.; Ruozzi, N. Exploration of Feature Representations for Predicting Retention-Session Performance in a VR Training Scenario. *Big Data Cogn. Comput.* **2021**, *5*, 29. https://doi.org/10.3390/bdcc5030029

Academic Editors: Achim Ebert, Peter Dannenmann and Gerrit van der Veer

Received: 25 May 2021
Accepted: 6 July 2021
Published: 12 July 2021

1. Introduction

Virtual Reality (VR) has been used extensively for education and training [1]. VR practitioners have developed educational VR environments for knowledge acquisition, such as learning about geometry [2], World War I [3], and how to inspect a haul truck [4]. VR has also been used for training psychomotor skills, such as how to visually scan for threats [5], use measurement tools [6], and put on personal protective equipment [7]. As consumer VR systems have become more easily available, these educational uses of VR have gained more attention [8].

The field of Intelligent Tutoring Systems (ITSs) is closely related to VR training and education. The goal of an ITS is to develop a deep understanding of each learner's behavior to allow for intelligent engagement in sustained learning activities [9]. While much work in the field of ITSs is focused on immediate knowledge, some research has made strides towards attempting to understand and predict longer-term retention of information, which is a better indicator of mastery than short-term retention [10]. ITSs incorporate Artificial Intelligence (AI) techniques to allow for tutoring systems that know "what they teach, who they teach, and how to teach it" through user models [11]. ITSs can use these user models to provide adaptive support for learners because they provide a representation of the learner in terms of relevant traits like learning behaviors and meta-cognitive ability [12]. This adaptive support can be implemented via proactive guidance (*scaffolding*) or reactive guidance (*feedback*) [13]. Most ITSs that create adaptive learning environments employ feedback [13], which is usually implemented with static rules for specific types of detected errors [14].

In this paper, we investigate using a machine learning (ML) approach to classify users of a VR training application into groups of low and high learning and retention outcomes, expanding upon our previous work which only sought to predict learning outcomes [15]. Our work in this paper makes use only of the head-mounted display (HMD) and handheld controller tracking data, as opposed to events or errors [14], which makes our approach viable for real-time scaffolding, before events or errors even occur. Unlike most prior approaches that are only capable of feedback [13], this tracking-based ML approach can potentially be used for both feedback and scaffolding, such as providing interaction cues [16], guiding the user's attention [17], or even simplifying the interactions [18]. Our work explores using HMD and controller tracking data as a means for predicting knowledge acquisition, knowledge retention, and performance retention.

After discussing related work, we present a user study in which participants used a VR training application to learn the procedure for troubleshooting a surgical robot [19], then returned a week later to perform this task again in VR. Participants were administered a knowledge test after the initial training session and before the week-later retention session to evaluate knowledge acquisition and retention, respectively. Errors were tracked in the retention simulation to evaluate performance retention. For each learning and retention outcome, we used the results to separate participants into two groups: high-performance participants scoring at least one standard deviation above the mean and low-performance participants (i.e., all remaining participants that did not score at least one standard deviation above the mean).

We then present the results of an ML experiment that employed support-vector machines (SVMs) [20] to predict which group each participant belonged to, based on their tracking data. For the experiment, we compared six different sets of input features based whether the data represented positions or velocities and three different combinations of HMD and controller data: (a) linear-and-angular features for both the HMD and controllers, (b) linear-and-angular features for the HMD and linear-only features for the controllers, and (c) linear-only features for both the HMD and controllers. We compared the accuracy (i.e., correctness of the prediction) and the Matthews correlation coefficient (MCC) (i.e., a measure of the quality of binary classifications [21]) of each model learned from these six different sets of input features. For the feature representations, we applied Principal Component Analysis (PCA) [22]. The results of our ML experiment indicate that this approach can yield high degrees of accuracy for predicting learning and retention outcomes, with our maximum observed overall accuracy at 96.7%. However, we note varying accuracy across our different input features, which reinforces the usefulness of exploring these features as hyperparameters when building models for similar educational purposes.

In order to understand how the SVMs classified participants into their high and low-performance groups with such high degrees of accuracy, we have visually inspected tracking data segments that were correctly identified by the SVMs. Visualizations show that participants with high learning and retention outcomes moved with better economies of motion than those participants with lower outcomes. These results show that it is feasible to use HMD and controller tracking data to develop new real-time scaffolding and feedback techniques for VR training applications. We anticipate that our work will be useful in the development of systems that predict learning and retention outcomes, and subsequently respond dynamically when learning can be improved. Such systems may make use of real-time classifiers to provide scaffolding and feedback that better support the user's learning. In this paper, we present the following research activities:

- A study that collected VR tracking data and results pertaining to knowledge acquisition, knowledge retention, and VR-based performance retention from 60 participants across two VR training sessions, separated by one week.
- An ML experiment investigating six sets of input features for predicting the three different outcomes. This is the first such experiment to investigate both learning and retention, particularly psychomotor retention.

We also present the following research results:

- Our results indicate that our velocity-based ML models generally outperformed our position-based models for predicting all three outcomes.
- Our results also indicate that VR tracking data can be better used to predict psychomotor-based outcomes than cognitive-based outcomes.

2. Related Work

Machine learning has been applied for predicting both cognitive and psychomotor learning outcomes. Much of this research has focused on creating "online" classifiers that can evaluate or predict performance in real time. There has also been some research more recently looking at predicting knowledge retention at least a few days after the last learning session.

2.1. Predicting Cognitive Learning Outcomes

There has been some work in the past on predicting cognitive learning outcomes in training and education environments, particularly in the realms of declarative and procedural knowledge. While the field of ITSs has many examples of predicting cognitive learning outcomes based on system and contextual events, we focus our discussion on research that makes use of physiological data for prediction.

The prediction of declarative knowledge gains based on physiological data has been an active area of research for some time. Schneider and Blikstein [23] developed an ear anatomy trainer with a tangible user interface and used ML methods to predict the learning gains of students. Based on a "rough" median split of their participants, they were able to achieve 100% accuracy with their SVM using a Multilayer Perceptron kernel [24]. Although they did not find a direct correlation between knowledge acquisition and posture, they determined that student posture had predictive value for classifiers by showing that it was a useful feature for their system. Won et al. [25] also researched prediction of declarative knowledge acquisition, specifically looking at the learning gains of environmental principals in a teacher–student dyad. They made use of skeletal data in the form of joint angles tracked from a Microsoft Kinect, a full-body motion-sensing device that makes use of structured light and time of flight , to predict the learning gains. While the system showed little predictive power over all dyads, decision trees yielded a classification accuracy of 85.7% when comparing the top seven and bottom seven teacher–student dyads.

Beyond these approaches for predicting declarative knowledge gains, researchers have also investigated the prediction of procedural knowledge gains based on physiological data. Amershi and Conati [12] designed a desktop tutor that taught mathematics using a clustering approach that used distilled eye-tracking features, such as gaze shifts and interface events. This ML model was able to classify students based on high and low knowledge acquisition with an average accuracy of 86.3%. Recently, we also developed an ML approach that partitions users into high and low knowledge acquisition groups [15]. We made use of velocity data derived from HMD and controller tracking data to predict the learning gains from a robotic operating room training scenario based on a post-test knowledge test and found a mean accuracy of 93.1% among our examined approaches. Unlike our recent work, in this paper, we investigate both position and velocity data, and we also investigate using ML models to predict retention outcomes, in addition to learning outcomes.

The research studies above indicate that physiologically based data, such as body postures [23,25], eye tracking [12], and body movements [15], can be used to predict cognitive learning outcomes, such as declarative and procedural knowledge acquisition.

2.2. Predicting Psychomotor Learning Outcomes

In addition to predicting cognitive learning outcomes, there has been research into predicting psychomotor learning outcomes from physiological data.

DeMoraes et al. [26] developed an online training assessment system for a bone marrow harvesting simulator that made use of a Gaussian Naïve Bayes classifier. This system was able to match expert classification with a Kappa coefficient of 80% by utilizing position, velocity, forces, and time as features. Ultimately, their system made mistakes in only 20 of 150 cases. Similarly, Sewell et al. [27] performed classification in a mastoidectomy simulator via a Hidden Markov Model. This system used position, distance, tool force, and suction position features, and yielded 85% correct classification for both novices and experts. In more recent work that attempts to predict learning of gross motor movements, dos Santos [28] investigated rhythm skill prediction in a dancing trainer. This system made use of data from the student's phone and the song tempo in beats per minute (BPM) to classify their movements as faster, correct, slower, or mixed and was able to correctly identify 74% of sessions with an F1 score of 0.79.

The works by DeMoraes et al. [26] and Sewell et al. [27] indicate that motion data collected during training can be used to predict fine motor skill learning of complex tasks. Additionally, dos Santos [28] has shown that gross motor movement learning can also be predicted and used for feedback. These results indicate that psychomotor outcomes can be estimated by using positional tracking data.

2.3. Predicting Cognitive Retention Outcomes

In addition to predicting the immediate learning outcomes, research has also been conducted to predict longer-term knowledge retention. Wang and Beck [10] examined using logistic regression to predict knowledge retention for a period of five to ten days after the student's last practice. Their approach makes use of features representing prior performance, when the student last practiced, and student response time. Work by Choffin et al. [29] has looked at modeling student learning and forgetting of skills by leveraging the Knowledge Tracing Machines framework [30] to embed features from multiple skills together. Their results suggest that incorporating both item-skill relationships and forgetting effects provides better learner models than only incorporating one or the other. Li et al. [31] found speed of mastery to be a relevant feature for predicting performance in week-later assessments within their Automatic Reassessment and Relearning System intelligent tutor.

By looking at the retention after a period of time, these works attempt to measure the "broader notion of knowing a skill" [10]. These works all found that the retention of knowledge appears to vary by skill and is possibly unique to each student [10,29,30] and generally requires additional considerations beyond prediction of knowledge acquisition.

2.4. Novelty of Current Work

To the best of our knowledge, our work is the only one that attempts to predict retention outcomes for psychomotor outcomes, and one of few works analyzing prediction of retention outcomes for procedural knowledge. This comparison can be seen in Table 1.

Table 1. An overview of related work that have predicted learning and retention outcomes.

Prediction	Paper	System	Input	Model
Cognitive learning	[12]	Math tutor	Eye tracking, Context events	K-means
	[25]	Classroom monitor	Skeletal tracking	Decision tree
	[23]	Anatomy tutor	Skeletal tracking, Context events	SVM
	[15]	Surgical simulation	Head tracking, Hand tracking	SVM
Psychomotor learning	[27]	Surgical simulation	Instrument tracking	HMM
	[26]	Surgical simulation	Instrument tracking	Naive Bayes
	[28]	Dance monitor	Phone accelerometer, Song BPM	Decision tree, kNN, Logistic regression Naive Bayes, Neural network, Random forest, SVM
Cognitive retention	[10]	Math tutor	System events	Logistic regression
	[31]	Math tutor	System events	Logistic regression
	[29]	Math tutor	System events	Logistic regression
Cognitive learning, Cognitive retention, Psychomotor retention	Ours	Surgical simulation	Head tracking, Hand tracking	SVM

3. VR Learning and Retention User Study

We made use of an existing VR application designed for training first assistants how to troubleshoot a faulted arm on a surgical robot [19]. This robotic operating room (OR) application makes use of virtual hand-based selections and manipulations [32,33] to support multiple equipment interactions, as well as communicating with a virtual surgeon and staff member by selecting dialog options. In order to complete the scenario, the user must perform an ordered set of steps involving these interactions while moving around the virtual OR by physically walking [34] and periodically looking at a vision cart screen. Table 2 shows a list of these subtasks and the types of interactions that they require.

We slightly varied the virtual training application between the learning and retention sessions. During the learning session, it provided interaction cues, which convey actions to take [35], for each step. These interaction cues consisted of verbal instructions and visual animations showing perceived affordances and feedforward information. Selection cues used semi-transparent green controller models that continuously linearly interpolated from the user's controller to the target dialog option or object to be selected (see Figure 1). Manipulation cues used semi-transparent green copies of the objects being manipulated and linearly interpolated these copies to the target positions for the manipulations. Finally, for travel cues, these animations consisted of semi-transparent green boots that linearly interpolated from the user's position to a icon representing the travel destination that read "Stand Here".

In the learning version of our VR training application, the interaction cues described were preemptively presented to the users, in order to demonstrate and train how to perform the troubleshooting task. However, for the retention-session version, these cues were not presented, unless the user committed an error or was inactive for 30 s.

Table 2. The subtasks and their required interactions involved in the VR training simulation.

#	Subtask	Interaction
1	Check error message	Walk + Look
2	Consult Surgeon	Select (dialog)
3	Ask to press power down	Select (dialog)
4	Ask to press power up	Select (dialog)
5	Ask to call support	Select (dialog)
6	Ask for release wrench	Select (dialog)
7	Grab release wrench	Select (wrench)
8	Ask for emergency stop	Select (dialog)
9	Hold instrument carriage	Select (carriage)
10	Insert wrench	Position (wrench)
11	Rotate wrench	Rotate (wrench)
12	Check vision monitor	Look
13	Remove wrench	Position (wrench)
14	Remove instrument	Position (instrument)
15	Give instrument to staff	Position (instrument)
16	Ask to recover fault	Select (dialog)
17	Ask to disable arm	Select (dialog)
18	Check error message	Walk + Look
19	Use cannula lever	Select (lever)
20	Use instrument clutch	Position (clutch)
21	Use port clutch	Position (clutch)
22	Ask to confirm disable	Select (dialog)
23	Check error message	Look
24	Ask to press recover fault	Select (dialog)

Figure 1. A first-person perspective of the VR training application used for our research.

3.1. Materials

The VR hardware for this study was HTC Vive system, consisting of an HMD and two handheld controllers, which were used to interact with the VR training application. The display of the Vive HMD has a resolution of 1080 × 1200 pixels per eye, a 90 Hz refresh rate, and affords a 110° diagonal field of view (FOV). We fitted the HMD with the Vive audio strap that integrates over-the-ear headphones. The VR application maintained 90 frames per second to match the Vive's refresh rate and was developed in Unity. The input data from the Vive was processed with the SteamVR plugin. For every frame in the VR training

application, the HMD and controller tracking data was logged. This data consisted of the global positions and quaternions of both the HMD and the controllers, as well as the frame's timestamp.

3.2. Procedure

The following procedure was reviewed and approved by the University of Texas at Dallas Institutional Review Board (IRB).

The human subjects study consisted of one learning and one retention session for each participant. The duration of the sessions was approximately 60 min for the learning session and 30 min for the retention session. The retention session occurred one week after the learning session.

Participants first gave informed consent, then the learning session began with a background survey on the demographics, education, and technology experience of the participant. To train the participant on how to use the HTC Vive, the experimenter would help the participant put on the HMD and run the SteamVR tutorial. After this, the experimenter would then run the participant through the VR training application. Once the participant finished, they were then administered a number of questionnaires regarding their VR experience. Finally, participants were administered a knowledge test consisting of multiple-choice questions pertaining to the training scenario.

One week later (restricted to the same day of the week to avoid confounds), a participant would begin the retention session with the experimenter administering the same knowledge test to measure knowledge retention. After completing the test, the experimenter would help the participant to put on the HTC Vive, and the participant would then experience the retention version of the VR application. After completing the retention application, the participant was given a free-response exit survey and compensated $15 USD.

3.3. Participants

A total of 61 participants were recruited through university mailing lists and completed the initial training session. However, one participant did not return to complete the retention session. Thus, our data consists of 60 participants (11 females, 49 males). None of our participants had prior knowledge of or experience with surgical robots or the training task. The mean age of the participants was 22.6 ± 4.2 years.

4. Machine Learning Experiments

In this section, we discuss the learning and retention outcomes, experimental design, input features, input feature representation, hyperparameters, and scoring method.

4.1. Learning and Retention Outcomes

For this research, we are concerned with three learning and retention outcomes: Knowledge Acquisition, Knowledge Retention, and Performance Retention. We measure Knowledge Acquisition and Knowledge Retention with the knowledge tests that were administered at the end of the learning session and beginning of the retention session, respectively. We measure Performance Retention by tracking errors and completion time at the subtask level. If participants made no errors and completed a subtask within 30 s (i.e., the period of inactivity allowed before an interaction cue was presented), they are regarded as having successfully completed that subtask. In order to predict these different outcomes in a consistent way, we choose to implement a similar approach to that by Moore et al. [15]. For each outcome factor, we fit a Gaussian distribution to the observed scores, then classify all scores one standard deviation above the mean as high performance, and all others as low performance. These splits are described in Table 3.

Table 3. Evaluation metrics and categorizations of participants.

Metric	Value Range	Mean Score	SD	# High	# Low
Knowledge Acquisition	$[0, 20]$	8.45	3.04	10	50
Knowledge Retention	$[0, 20]$	10.50	3.60	9	51
Performance Retention	$[0, 24]$	8.47	3.05	9	51

(# High and # Low indicate the number of participants identified with high- and low-performance.)

4.2. Experimental Design

In order to evaluate the performance of our models, we conduct an exhaustive grid-search with four-fold, participant-level cross-validation across all hyperparameter options. This is performed by first segmenting our data into an 80/20 train/test split for later evaluation. On the training data, we create four approximately equal partitions to iteratively train on three of the partitions and evaluate on the fourth, retaining the average score for comparison against other hyperparameter configurations. We then select the configuration with the highest average score, train on the entirety of the training data, and then test on both the training and testing data, separately and overall.

4.3. Input Features

Our data consists of the position and rotation of the tracked HMD and handheld controllers at a rate of 90 Hz, the frame rate of the training application. We encode this tracking data in the form of a three-value vector for position and a four-value quaternion for rotation. To mitigate possible noise introduced by including the rotation data, we compare three sets of position and rotation data: (a) linear-and-angular features for both the HMD and controllers, (b) linear-and-angular features for the HMD and linear-only features for the controllers, and (c) linear-only features for both the HMD and controllers. Because of recent success seen in making use of velocity rather than position for predicting success [15,36], we also look at the linear and angular velocities derived from the position and rotation data, thus resulting in six total conditions for each evaluation metric.

4.4. Input Feature Representation

In our analysis, we take the 90 Hz data and use a sliding window approach to chunk it into spans that are more meaningful than an instant of data. This window is made by selecting a time span for the segment length, and another for the shift size. Depending on these parameters, there can be a significant amount of overlap between subsequent windows. The frames of data within a window are concatenated to create a high-dimensional vector that encodes the information. These vectors are then processed with Principal Component Analysis (PCA) to extract salient features. This process greatly reduces the amount of data and forms the input vectors used for training our models.

4.5. Hyperparameters

We consider several hyperparameters for both the representation of the input and the support vector machine (SVM). The six described options for Input Feature Representation can be regarded as hyperparameters within an evaluation metric prediction. However, in this paper, we will treat them as separate experiments as we are interested in discussing the differences in their performance. The sliding window approach for chunking our data is determined by two hyperparameters: the segment length and shift size. The PCA feature extraction also has a hyperparameter for determining the number of features extracted. Finally, the SVM that we use in our methodology has two hyperparameters: penalty and expressivity. The range of values that we explore for these five hyperparameters are shown in Table 4.

Table 4. Value range of hyperparameters considered.

Hyperparameter	Seg. Length	Shift Size	PCA Features	Penalty	Expressivity
Min Value	30	5	10	1	0.001
Max Value	120	13	80	10,000	0.090

4.6. Scoring Method

We fit SVMs [20] with Gaussian kernels to the featurized data. The resulting linear separator will yield a classification of low or high for each window segment. One possible way to extend the segment classification to an entire user is to determine the user's label by majority vote over the predicted labels of the user's velocity segments. This approach is reasonable, but doesn't accommodate well for cases where several vectors are incorrectly classified with low confidence. Thus, we weight each vote based on its distance from the decision boundary.

Let $l_{u,i}$ and $d_{u,i}$ denote the predicted label and corresponding distance for the ith segment of participant p. We define the participant-level confidence that p belongs to class c as:

$$\text{conf}_{p,c} = \frac{\sum_i \text{ st. } {}_{l_{p,i}=c}\, d_{p,i}}{\sum_{i=1}^{N} d_{p,i}},$$

where N is the number of segments that belong to participant p. Then, the predicted label of the user is determined by selecting the class c that maximizes the confidence.

5. Machine Learning Results

In this section, we will detail the results of our ML experiments across our learning metrics. We choose to analyze the performance of each model in terms of the Matthews correlation coefficient (MCC) score. This value is a special case of Pearson's correlation coefficient for two categories and is valued from -1 (full counter-correlation) to 1 (full correlation). Unlike metrics such as accuracy, MCC takes into account the relative size of each category when assigning a score, thus balancing the weight of disparately sized categories [21]. This is particularly useful for our experiments in which the relative category sizes have approximately a 5:1 ratio of low-performance to high-performance participants.

5.1. Knowledge Acquisition Results

Table 5 shows the full results of our Knowledge Acquisition experiment. First, it is clear from the discrepancies in high-performance accuracy between the training (High Acc = 1.000) and testing (High Acc = 0.000) datasets and the MCC scores for the testing data (MCC = -0.135) that the first and third sets of input features, both position-based, suffered from overfitting the high-performance training data. Similarly, the sixth set, the velocity-based, linear-only input features set, also suffered from overfitting the high-performance training data, resulting in a low MCC score (MCC = 0.000).

Interestingly, the two best performing models were the second and fifth sets of input features (i.e., the position-based and velocity-based linear-and-angular HMD features and linear-only controller features, respectively). The velocity-based fifth set yielded the best test MCC of 0.674 and a better overall accuracy of 0.833 than the position-based second set, which yielded a test MCC of 0.400 and an overall accuracy of 0.783. This position-based second set was more conservative in identifying low-performance participants in both the training (Low Acc = 0.725) and testing (Low Acc = 0.900) datasets than the velocity-based fifth set, which was more accurate for both training (Low Acc = 0.825) and testing (Low Acc = 1.000).

Table 5. Knowledge Acquisition prediction results. For each of the six sets of input features, the accuracy of the low-performance and high-performance predictions and the MCC are shown for both the training and testing data. Accuracy and MCC are also shown for each set of input features across the entire dataset (i.e., both the training and testing data).

Set	Position/ Velocity	HMD	Controllers	Train/ Test	Low Acc	High Acc	MCC	Overall Acc	Overall MCC
1	Position Position	Lin + Ang Lin + Ang	Lin + Ang Lin + Ang	Train Test	0.800 0.900	1.000 0.000	0.632 −0.135	0.817	0.513
2	Position Position	Lin + Ang Lin + Ang	Linear Linear	Train Test	0.725 0.900	1.000 0.500	0.553 0.400	0.783	0.516
3	Position Position	Linear Linear	Linear Linear	Train Test	0.829 0.900	1.000 0.000	0.665 −0.135	0.836	0.541
4	Velocity Velocity	Lin + Ang Lin + Ang	Lin + Ang Lin + Ang	Train Test	0.700 0.700	0.750 0.500	0.346 0.158	0.700	0.309
5	Velocity Velocity	Lin + Ang Lin + Ang	Linear Linear	Train Test	0.825 1.000	0.750 0.500	0.482 0.674	0.833	0.493
6	Velocity Velocity	Linear Linear	Linear Linear	Train Test	1.000 1.000	0.625 0.000	0.762 0.000	0.917	0.674

5.2. Knowledge Retention Results

Table 6 shows the complete results of our Knowledge Retention experiment. Unlike the Knowledge Acquisition experiment, there are no clear results indicating that any of the six models overfitted the training data for low- or high-performance outcomes. However, it is also the case that none of the six models performed as well on the Knowledge Retention results, as they performed on the Knowledge Acquisition results. For Knowledge Acquisition, five of the six models yielded overall MCC scores ranging from 0.493 to 0.674. However, for Knowledge Retention, the six models only yielded overall MCC scores ranging from 0.327 to 0.430. In terms of test MCC scores, two of the models, the first and third sets, yielded chance-like results (i.e., near zero MCC scores).

Table 6. Knowledge Retention prediction results. For each of the six sets of input features, the accuracy of the low-performance and high-performance predictions and the MCC are shown for both the training and testing data. Accuracy and MCC are also shown for each set of input features across the entire dataset (i.e., both the training and testing data).

Set	Position/ Velocity	HMD	Controllers	Train/ Test	Low Acc	High Acc	MCC	Overall Acc	Overall MCC
1	Position Position	Lin + Ang Lin + Ang	Lin + Ang Lin + Ang	Train Test	0.732 0.600	0.857 0.500	0.435 0.076	0.717	0.358
2	Position Position	Lin+Ang Lin + Ang	Linear Linear	Train Test	0.732 0.700	0.857 0.500	0.435 0.158	0.733	0.377
3	Position Position	Linear Linear	Linear Linear	Train Test	0.756 0.500	0.857 0.500	0.459 0.000	0.717	0.358
4	Velocity Velocity	Lin + Ang Lin + Ang	Lin + Ang Lin + Ang	Train Test	0.854 0.800	0.714 0.500	0.477 0.258	0.817	0.430
5	Velocity Velocity	Lin + Ang Lin + Ang	Linear Linear	Train Test	0.610 0.400	0.857 1.000	0.331 0.316	0.617	0.327
6	Velocity Velocity	Linear Linear	Linear Linear	Train Test	0.805 0.800	0.714 0.500	0.412 0.258	0.783	0.380

The best performing model for predicting Knowledge Retention was the fourth set of velocity-based, linear-and-angular HMD and controller input features. This model yielded the best overall MCC of 0.430 and the second best test MCC of 0.258, which would have

dramatically improved had it identified both high-performance participants instead of just one (i.e., High Acc = 0.500). The sixth set of velocity-based, linear-only HMD and controller input features performed similarly with an overall MCC of 0.380 and the same test MCC of 0.258. Finally, the fifth set of velocity-based, linear-and-angular HMD and linear-only controller features performed well with an overall MCC of 0.327 and the best test MCC of 0.316. However, this model was overly generous with high-performance labels, particularly in the testing data (High Acc = 1.000), which yielded poor low-performance identification (Low Acc = 0.400).

5.3. Performance Retention Results

Table 7 shows the complete results of our Performance Retention experiment. Like the Knowledge Retention results, there are no clear results indicating that any of the six models overfitted the training data for low- or high-performance outcomes. However, unlike the Knowledge Retention results, the six models performed relatively well on the Performance Retention results, yielding overall MCC scores ranging from 0.341 to 0.869 and test MCC scores ranging from 0.258 to 0.529.

The best performing model for predicting Performance Retention in the VR training application was the sixth set of velocity-based, linear-only HMD and controller input features. It yielded the highest overall MCC score of 0.869 (with an overall accuracy of 0.967), and it yielded one of the top test MCC scores of 0.400, incorrectly predicting exactly one low-performance participant and one high-performance participant. The third and fourth sets of input features yielded the best test MCC scores of 0.529. However, these models were generous with high-performance labels in the testing data (High Acc = 1.000), which yielded conservative low-performance identification (Low Acc = 0.700).

Table 7. Performance Retention prediction results. For each of the six sets of input features, the accuracy of the low-performance and high-performance predictions and the MCC are shown for both the training and testing data. Accuracy and MCC are also shown for each set of input features across the entire dataset (i.e., both the training and testing data).

Set	Position/ Velocity	HMD	Controllers	Train/ Test	Low Acc	High Acc	MCC	Overall Acc	Overall MCC
1	Position	Lin+Ang	Lin+Ang	Train	0.878	1.000	0.716	0.867	0.620
	Position	Lin + Ang	Lin + Ang	Test	0.800	0.500	0.258		
2	Position	Lin + Ang	Linear	Train	0.927	1.000	0.805	0.900	0.685
	Position	Lin + Ang	Linear	Test	0.800	0.500	0.258		
3	Position	Linear	Linear	Train	0.780	1.000	0.584	0.800	0.572
	Position	Linear	Linear	Test	0.700	1.000	0.529		
4	Velocity	Lin + Ang	Lin + Ang	Train	0.683	0.714	0.290	0.700	0.341
	Velocity	Lin + Ang	Lin + Ang	Test	0.700	1.000	0.529		
5	Velocity	Lin + Ang	Linear	Train	0.780	1.000	0.584	0.800	0.517
	Velocity	Lin + Ang	Linear	Test	0.800	0.500	0.258		
6	Velocity	Linear	Linear	Train	1.000	1.000	1.000	0.967	0.869
	Velocity	Linear	Linear	Test	0.900	0.500	0.400		

6. Visual Inspection of Results

Given the highly accurate results of our ML experiments, we decided to conduct visual inspections of the motions of participants, in order to try and understand how the SVMs are classifying participants with such high degrees of accuracy. We use an approach similar to our recent work [15], rendering the positions and high-velocity moments of the tracking data within the environment, in a top-down orthographic view. We show the HMD tracking data in yellow, and the left and right controller data in blue and red, respectively. To convey high-velocity moments, green lines connecting the three tracked traces are rendered for frames in which the velocity of any tracked device exceeded 1 m/s,

with a minimum 15 frame break since the last rendered line to prevent visual clutter. To convey time, we adjust the brightness of all four traces from dim to bright over the course of the segment. For example, Figure 2 shows the visualizations of the same set of actions for a participant with low performances on all three metrics (i.e., Knowledge Acquisition, Knowledge Retention, and Performance Retention) and a participant with high performances on all three metrics.

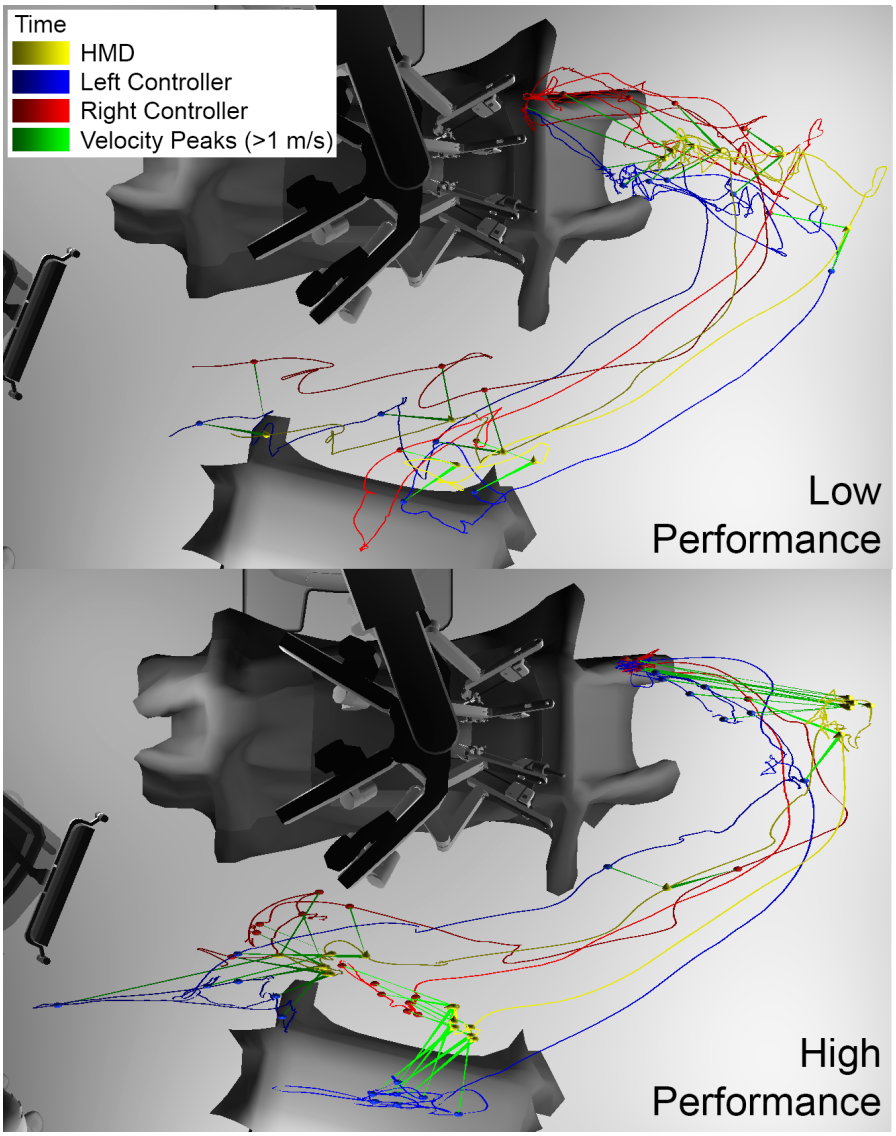

Figure 2. Visualizations of motion data from a low-performance participant and a high-performance participant, based on all three metrics, for the same set of actions.

Through our visual inspections, we have noticed that the motions of low-performance participants generally have more directional changes and return more often to prior positions than the motions of high-performance participants. Additionally, we have observed that low-performance participants exhibit fewer velocity peaks (i.e., moments in which velocity exceeds 1 m/s) than high-performance participants. Together, these results suggest that low-performance participants moved more haphazardly and with less certainty. In contrast, high-performance participants moved more smoothly and with greater certainty, and likely, intention. These results are similar to prior results from the robotic surgery domain, in which less-experienced robotic surgeons often demonstrate worse economies of motion (i.e., excessive motions) than more-experienced robotic surgeons [37,38].

Considering that our visualizations indicated that low-performance participants moved more haphazardly and timidly, we decided to investigate whether prior VR experience had a significant effect on whether a participant would be a low- or high-performance learner. We conducted a Mann–Whitney U test for each metric to compare the distribution differences of low and high performances among participants with and without prior VR experiences. Table 8 shows the results of these tests. The results clearly indicate that prior VR experience did not have a significant effect on Knowledge Acquisition, Knowledge Retention, or even Performance Retention.

Table 8. Results of Mann–Whitney U tests comparing the distribution differences of low and high performances among participants with or without prior VR experiences.

Metric	Prior VR Exp Low: High Ratio	No VR Exp Low: High Ratio	U	*p*
Knowledge Acquisition	28:6	22:4	432.0	0.888
Knowledge Retention	29:5	22:4	439.0	0.968
Performance Retention	27:7	24:2	385.0	0.401

7. Discussion

In this section, we will discuss our results and what knowledge has been gained through this work.

7.1. Position-Based versus Velocity-Based Models

Overall, we found that our velocity-based models generally outperformed our position-based models. For all three metrics, we found that one of our velocity-based models yielded the best results. Our velocity-based, linear-and-angular HMD with linear-only controller model was the best for predicting Knowledge Acquisition. Our velocity-based, linear-and-angular HMD and controller input features had the best results for predicting Knowledge Retention. Finally, our velocity-based, linear-only HMD and controller input features were the best for predicting Performance Retention.

In contrast, we observed that our position-based models produced some of the worst results. For predicting Knowledge Acquisition, we found that two of the position-based models suffered from overfitting the high-performance training data. Similarly, we found that the same two position-based models, the linear-and-angular HMD and controller input features and the linear-only HMD and controller input features, yielded chance-like results for predicting Knowledge Retention. However, all three position-based models did perform moderately well for predicting Performance Retention, which is based on psychomotor skills as opposed to cognitive-only skills.

These results strongly support future research into the use of velocity-based models for predicting learning and retention outcomes, both cognitive and psychomotor ones. On the other hand, particularly if computing resources are limited, future researchers can likely omit position-based models for predicting cognitive learning and retention outcomes, like our Knowledge Acquisition and Knowledge Retention metrics. However, we believe that position-based models are still viable for predicting psychomotor outcomes, based on the results of our Performance Retention experiment.

7.2. Linear-and-Angular versus Linear-Only Models

In general, we did not find any results indicating that linear-and-angular or linear-only models performed better than the other. In our experiments, we investigated three different combinations of these features: (a) linear-and-angular features for both the HMD and controllers, (b) linear-and-angular features for the HMD and linear-only features for the controllers, and (c) linear-only features for both the HMD and controllers. We found that a velocity-based version of each combination performed the best for one of our three metric predictions. We found that the velocity-based, linear-and-angular HMD and controller combination performed the best overall for predicting Knowledge Retention. We found that

the velocity-based, linear-and-angular HMD and linear-only controller model performed the best overall, without overfitting, for predicting Knowledge Acquisition. Finally, we found that the velocity-based, linear-only HMD and controller features performed the best overall for predicting Performance Retention.

These results suggest that there is still much research to be conducted into the investigation of these linear and angular features. It is likely that one combination is not superior to the others, and that researchers should investigate each to identify the best model for their prediction metric. Furthermore, it is important to note that we did not investigate angular-only features, which might be viable models for some types of prediction, such as simulator sickness [39]. There are also other types of features that should be investigated, such as the linear distances between the HMD and handheld controllers [40].

7.3. Cognitive versus Psychomotor Models

We found that our models produced the best results for the psychomotor-based Performance Retention metric, as opposed to the cognitive-based Knowledge Acquisition and Knowledge Retention metrics. The Performance Retention models yielded generally higher overall MCC scores, with five of the six models ranging from 0.517 to 0.869. In contrast, four of the six Knowledge Acquisition models had overall MCC scores ranging from 0.309 to 0.516 and the best Knowledge Retention model had an overall MCC score of 0.430. The psychomotor-based Performance Retention metric also yielded the best overall accuracy of 0.967, out of all 18 models evaluated, and the only perfect MCC score of 1.000 for the velocity-based, linear-only HMD and controller input features model.

These results indicate that VR tracking data can be better used to predict psychomotor-based learning and retention outcomes than to predict cognitive-based outcomes. This is intuitive as VR tracking data is directly generated by psychomotor-based actions. Hence, it is a more-direct representation of the psychomotor-based mental models that participants have.

These results also imply that VR technologies are most likely more useful for training psychomotor-based skills, such as troubleshooting a surgical robot, as opposed to cognitive-only skills, such as mathematical calculations. However, more research is necessary to investigate these potential differences in usefulness. In particular, real-world efficacy evaluations of skills transfer is necessary. In our research, we were only able to assess psychomotor retention by having participants use the retention version of our VR training application. Ideally, we would have assessed psychomotor retention by having participants demonstrate a real-world transfer of those psychomotor-based skills to a physical surgical robot and OR environment. However, this was not feasible due to the limited availability of such robotic ORs in hospitals [19].

7.4. Limitations

In our ML experiments, we chose to omit predicting performance during the learning session of our study. This decision was made because while there may be utility in such a classifier, it would be trying to predict values derived in part from actions the participant has already undertaken, and wouldn't be as directly comparable.

Additionally, in this study, we performed PCA to reduce the dimensionality of our input vectors before using them to train the SVM models. While this is a generally accepted practice, for completeness, a future study would ideally also treat rotation algorithms for the PCA, as well as other feature extraction techniques such as convex-hull representations [15], and other types of ML models as hyperparameters, and evaluate performance among those. The decision not to explore those here was due in part to the computational complexity of such a broad exploration.

Finally, the positional data was originally encoded in terms of VR world space, and while this remained consistent with the real-world space in terms of scale, orientation, and being stationary, it decreased the model's ability to encode salient features such as head-to-hand distance. Directly encoding such salient features would likely have been beneficial

for our models' predictive power as Pfeuffer et al. [40] found. However, evaluating all possible salient features would be intractable, and selecting a subset of features to evaluate over may have introduced bias.

8. Conclusions

In this paper, we explored the feasibility of employing ML models based on different sets of VR tracking features to predict learning and retention outcomes from a VR training application. Our results show that such models can be used to predict such educational outcomes with high degrees of accuracy. Furthermore, our results indicate that velocity-based models are likely better predictors of learning and retention outcomes, particularly cognitive-based outcomes, than position-based models. However, our results did not indicate that any particular combination of linear-and-angular or linear-only conditions yielded better results. Hence, we generally recommend investigating different linear and angular combinations of input features, in addition to considering some features not investigated in this work, such as head-to-hand distances. Finally, our results clearly indicate that VR tracking features can be better used to predict psychomotor outcomes than cognitive ones.

Author Contributions: Conceptualization, A.G.M., R.P.M., and N.R.; methodology, N.R.; software, A.G.M.; validation, A.G.M.; formal analysis, R.P.M.; investigation, A.G.M.; resources, R.P.M.; data curation, A.G.M.; writing—original draft preparation, A.G.M.; writing—review and editing, R.P.M. and N.R.; visualization, A.G.M.; supervision, R.P.M. and N.R.; project administration, R.P.M.; funding acquisition, R.P.M. All authors have read and agreed to the published version of the manuscript.

Funding: This material is based on work partially supported by the National Science Foundation grant number 2021607—"CAREER: Leveraging the Virtualness of Virtual Reality for More-Effective Training" as well as by the Defense Advanced Research Projects Agency Explainable Artificial Intelligence (XAI) program grant number N66001-17-2-4032.

Institutional Review Board Statement: This study was conducted according to the guidelines of the Declaration of Helsinki, and approved by the Institutional Review Board of the University of Texas at Dallas (protocol #15-102; approved on 18 June 2019).

Informed Consent Statement: Informed consent was obtained from all subjects involved in this study.

Data Availability Statement: Participant data unavailable due to privacy and ethical restrictions. Code available at github.com/tapiralec/LearningandRetentionOutcomes (accessed on 25 June 2021).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Bowman, D.A.; McMahan, R.P. Virtual Reality: How Much Immersion Is Enough? *Computer* **2007**, *40*, 36–43. [CrossRef]
2. Lai, C.; McMahan, R.P.; Kitagawa, M.; Connolly, I. Geometry explorer: Facilitating geometry education with virtual reality. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*; Springer International Publishing: Cham, Switzerland, 2016; Volume 9740, pp. 702–713; [CrossRef]
3. Yu, R.; Duer, Z.; Ogle, T.; Bowman, D.A.; Tucker, T.; Hicks, D.; Choi, D.; Bush, Z.; Ngo, H.; Nguyen, P.; Liu, X. Experiencing an Invisible World War I Battlefield Through Narrative-Driven Redirected Walking in Virtual Reality. In Proceedings of the 2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR), Tuebingen/Reutlingen, Germany, 18–22 March 2018 ; pp. 313–319; [CrossRef]
4. McMahan, R.P.; Bowman, D.A.; Schafrik, S.; Karmis, M. Virtual Environment Training for Preshift Inspections of Haul Trucks to Improve Mining Safety. In *First International Future Mining Conference and Exhibition*; The Australasian Institute of Mining and Metallurgy (AusIMM): Sydney, Australia, 2008; pp. 167–174.
5. Ragan, E.D.; Bowman, D.A.; Kopper, R.; Stinson, C.; Scerbo, S.; McMahan, R.P. Effects of Field of View and Visual Complexity on Virtual Reality Training Effectiveness for a Visual Scanning Task. *IEEE Trans. Vis. Comput. Graph.* **2015**, *21*, 794–807. [CrossRef]
6. Bertrand, J.; Brickler, D.; Babu, S.; Madathil, K.; Zelaya, M.; Wang, T.; Wagner, J.; Gramopadhye, A.; Luo, J. The role of dimensional symmetry on bimanual psychomotor skills education in immersive virtual environments. In Proceedings of the 2015 IEEE Virtual Reality (VR), Arles, France, 23–27 March 2015; pp. 3–10; [CrossRef]
7. Eubanks, J.C.; Somareddy, V.; McMahan, R.P.; Lopez, A.A. Full-body portable virtual reality for personal protective equipment training. In *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*; Springer International Publishing: Cham, Switzerland, 2016; Volume 9740, pp. 490–501; [CrossRef]

8. Carruth, D.W. Virtual reality for education and workforce training. In Proceedings of the 2017 15th International Conference on Emerging eLearning Technologies and Applications (ICETA), Stary Smokovec, Slovakia, 26–27 October 2017; pp. 1–6; [CrossRef]
9. Corbett, A.T.; Koedinger, K.R.; Anderson, J.R. Chapter 37 - Intelligent Tutoring Systems. In *Handbook of Human-Computer Interaction*, 2nd ed.; Helander, M.G., Landauer, T.K., Prabhu, P.V., Eds.; North-Holland: Amsterdam, The Netherlands, 1997; pp. 849–874; [CrossRef]
10. Wang, Y.; Beck, J.E. Using Student Modeling to Estimate Student Knowledge Retention. In Proceedings of the International Conference on Educational Data Mining (EDM), Chania, Greece, 19–21 June 2012.
11. Nwana, H.S. Intelligent tutoring systems: An overview. *Artif. Intell. Rev.* **1990**, *4*, 251–277. [CrossRef]
12. Amershi, S.; Conati, C. Unsupervised and supervised machine learning in user modeling for intelligent learning environments. In Proceedings of the International Conference on Intelligent User Interfaces, Proceedings IUI, Honolulu, HI, USA, 28–31 January 2007; pp. 72–81; [CrossRef]
13. VanLehn, K. The Relative Effectiveness of Human Tutoring, Intelligent Tutoring Systems, and Other Tutoring Systems. *Educ. Psychol.* **2011**, *46*, 197–221. [CrossRef]
14. Gutierrez, F.; Atkinson, J. Adaptive feedback selection for intelligent tutoring systems. *Expert Syst. Appl.* **2011**, *38*, 6146–6152. [CrossRef]
15. Moore, A.G.; McMahan, R.P.; Dong, H.; Ruozzi, N. Extracting Velocity-Based User-Tracking Features to Predict Learning Gains in a Virtual Reality Training Application. In Proceedings of the 2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR), Porto de Galinhas, Brazil, 9–13 November 2020; IEEE Computer Society: Los Alamitos, CA, USA, 2020; pp. 881–890; [CrossRef]
16. Hu, X.; Moore, A.G.; Eubanks, J.C.; Aiyaz, A.A.; McMahan, R.P. The Effects of Delayed Interaction Cues in Virtual Reality Training. In Proceedings of 2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW), Atlanta, GA, USA, 22–26 March 2020; pp. 63–69; [CrossRef]
17. Danieau, F.; Guillo, A.; Doré, R. Attention guidance for immersive video content in head-mounted displays. In Proceedings of the 2017 IEEE Virtual Reality (VR), Los Angeles, CA, USA, 18–22 March 2017; pp. 205–206; [CrossRef]
18. McMahan, R.P.; Herrera, N.S. AFFECT: Altered-fidelity framework for enhancing cognition and training. *Front. ICT* **2016**, *3*, 29. [CrossRef]
19. Moore, A.G.; Hu, X.; Eubanks, J.C.; Aiyaz, A.A.; McMahan, R.P. A Formative Evaluation Methodology for VR Training Simulations. In Proceedings of the 2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW), Atlanta, GA, USA, 22–26 March 2020; pp. 125–132; [CrossRef]
20. Boser, B.E.; Guyon, I.M.; Vapnik, V.N. A training algorithm for optimal margin classifiers. In Proceedings of the Fifth Annual Workshop on Computational Learning Theory, Pittsburgh, PA, USA, 27–29 July 1992; pp. 144–152.
21. Chicco, D.; Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genom.* **2020**, *21*, 6. [CrossRef] [PubMed]
22. Wall, M.E.; Rechtsteiner, A.; Rocha, L.M. Singular value decomposition and principal component analysis. In *A Practical Approach to Microarray Data Analysis*; Springer: Boston, MA, USA, 2003; pp. 91–109; doi:10.1007/0-306-47815-3_5 [CrossRef]
23. Schneider, B.; Blikstein, P. Unraveling students' interaction around a tangible interface using multimodal learning analytics. *J. Educ. Data Min.* **2015**, *7*, 89–116.
24. Pal, S.; Mitra, S. Multilayer perceptron, fuzzy sets, and classification. *IEEE Trans. Neural Netw.* **1992**, *3*, 683–697. [CrossRef] [PubMed]
25. Won, A.S.; Bailenson, J.N.; Janssen, J.H. Automatic detection of nonverbal behavior predicts learning in dyadic interactions. *IEEE Trans. Affect. Comput.* **2014**, *5*, 112–125. [CrossRef]
26. De Moraes, R.M.; Dos Santos Machado, L. Online training assessment in virtual reality simulators based on Gaussian Naive Bayes. In Proceedings of the 8th International FLINS Conference, Madrid, Spain, 21–24 September 2008.
27. Sewell, C.; Morris, D.; Blevins, N.H.; Dutta, S.; Agrawal, S.; Barbagli, F.; Salisbury, K. Providing metrics and performance feedback in a surgical simulator. *Comput. Aided Surg.* **2008**, *13*, 63–81. [CrossRef] [PubMed]
28. dos Santos, A.D.P. Using Motion Sensor and Machine Learning to Support the Assessment of Rhythmic Skills in Social Partner Dance: Bridging Teacher, Student and Machine Contexts. Ph.D. Thesis, University of Sydney, Sydney, Australia, 2019.
29. Choffin, B.; Popineau, F.; Bourda, Y.; Vie, J.J. DAS3H: modeling student learning and forgetting for optimally scheduling distributed practice of skills. *arXiv* **2019**, arXiv:1905.06873.
30. Vie, J.J.; Kashima, H. Knowledge tracing machines: Factorization machines for knowledge tracing. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 750–757.
31. Li, S.; Xiong, X.; Beck, J. Modeling Student Retention in an Environment with Delayed Testing. Available online: https://core.ac.uk/download/pdf/213000909.pdf (accessed on 25 June 2021).
32. Moore, A.G.; Kodeih, M.; Singhania, A.; Wu, A.; Bashir, T.; McMahan, R.P. The Importance of Intersection Disambiguation for Virtual Hand Techniques. In Proceedings of the 2019 IEEE International Symposium on Mixed and Augmented Reality (ISMAR), Beijing, China, 14–18 October 2019; pp. 310–317.
33. McMahan, R.P.; Kopper, R.; Bowman, D.A. Principles for designing effective 3D interaction techniques. In *Handbook of Virtual Environments*; CRC Press: Boca Raton, FL, USA, 2014; pp. 299–325.

34. LaViola, J.J., Jr.; Kruijff, E.; McMahan, R.P.; Bowman, D.; Poupyrev, I.P. *3D User Interfaces: Theory and Practice*; Addison-Wesley Professional: Boston, MA, USA, 2017.
35. Hu, X.; Moore, A.G.; Coleman Eubanks, J.; Aiyaz, A.; P. McMahan, R. Evaluating Interaction Cue Purpose and Timing for Learning and Retaining Virtual Reality Training. In Proceedings of the 2020 Symposium on Spatial User Interaction (SUI), Virtual Event, Canada, 30 October–1 November 2020; Article 5, pp. 1–9. doi:10.1145/3385959.3418448 [CrossRef]
36. Padmanaban, N.; Ruban, T.; Sitzmann, V.; Norcia, A.M.; Wetzstein, G. Towards a machine-learning approach for sickness prediction in 360 stereoscopic videos. *IEEE Trans. Vis. Comput. Graph.* **2018**, *24*, 1594–1603. [CrossRef]
37. Ershad, M.; Koesters, Z.; Rege, R.; Majewicz, A. Meaningful assessment of surgical expertise: Semantic labeling with data and crowds. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*; Springer: Cham, Switzerland, 2016; pp. 508–515.
38. Lendvay, T.S.; Brand, T.C.; White, L.; Kowalewski, T.; Jonnadula, S.; Mercer, L.D.; Khorsand, D.; Andros, J.; Hannaford, B.; Satava, R.M. Virtual reality robotic surgery warm-up improves task performance in a dry laboratory environment: a prospective randomized controlled study. *J. Am. Coll. Surg.* **2013**, *216*, 1181–1192. [CrossRef] [PubMed]
39. Kennedy, R.S.; Lane, N.E.; Berbaum, K.S.; Lilienthal, M.G. Simulator Sickness Questionnaire: An enhanced method for quantifying simulator sickness. *Int. J. Aviat. Psychol.* **1993**, *3*, 203–220. [CrossRef]
40. Pfeuffer, K.; Geiger, M.J.; Prange, S.; Mecke, L.; Buschek, D.; Alt, F. Behavioural Biometrics in VR: Identifying People from Body Motion and Relations in Virtual Reality. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Glasgow, Scotland, UK, 4–9 May 2019; Association for Computing Machinery: New York, NY, USA, 2019; CHI'19, pp. 1–12. [CrossRef]