Comunicar, n. 76, v. XXXI, 2023 | Media Education Research Journal | ISSN: 1134-3478; e-ISSN: 1988-3478
www.comunicarjournal.com

Towards learner performance evaluation in iVR learning environments using eye-tracking and machine-learning

Hacia una metodología de evaluación del rendimiento del alumno en entornos de aprendizaje iVR utilizando eye-tracking y aprendizaje automático

Dr. Ana Serrano-Mamolar. Postdoctoral Research Associate, Department of Computer Engineering, University of Burgos (Spain) (asmamolar@ubu.es) (https://orcid.org/0000-0002-0027-7128)
Ines Miguel-Alonso. Predoctoral Fellow, Department of Computer Engineering, University of Burgos (Spain) (imalonso@ubu.es) (https://orcid.org/0000-0001-8882-7587)
Dr. David Checa. Assistant Professor, Department of Computer Engineering, University of Burgos (Spain) (dcheca@ubu.es) (https://orcid.org/0000-0001-6623-3614)
Dr. Carlos Pardo-Aguilar. Senior Lecturer, Department of Computer Engineering, University of Burgos (Spain) (cpardo@ubu.es) (https://orcid.org/0000-0003-1424-1318)

ABSTRACT

At present, the use of eye-tracking data in immersive Virtual Reality (iVR) learning environments is set to become a powerful tool for maximizing learning outcomes, due to the low-intrusiveness of eye-tracking technology and its integration in commercial iVR Head Mounted Displays. However, the most suitable technologies for data processing should first be identified before their use in learning environments can be generalized. In this research, the use of machine-learning techniques is proposed for that purpose, evaluating their capabilities to classify the quality of the learning environment and to predict user learning performance. To do so, an iVR learning experience simulating the operation of a bridge crane was developed. Through this experience, the performance of 63 students was evaluated, both under optimum learning conditions and under stressful conditions. The final dataset included 25 features, mostly temporal series, with a dataset size of up to 50M data points. The results showed that different classifiers (KNN, SVM and Random Forest) provided the highest accuracy when predicting learning performance variations, while the accuracy of user learning performance was still far from optimized, opening a new line of future research. This study has the objective of serving as a baseline for future improvements to model accuracy using complex machine-learning techniques.

RESUMEN

Actualmente, el uso de los datos del seguimiento de la mirada en entornos de aprendizaje de Realidad Virtual inmersiva (iVR) está destinado a ser una herramienta fundamental para maximizar los resultados de aprendizaje, dada la naturaleza poco intrusiva del eye-tracking y su integración en las gafas comerciales de Realidad Virtual. Pero, antes de que se pueda generalizar el uso del eye-tracking en entornos de aprendizaje, se deben identificar las tecnologías más adecuadas para el procesamiento de datos. Esta investigación propone el uso de técnicas de aprendizaje automático para este fin, evaluando sus capacidades para clasificar la calidad del entorno de aprendizaje y predecir el rendimiento de aprendizaje del usuario. Para ello, se ha desarrollado una experiencia docente en iVR para aprender el manejo de un puente-grúa. Con esta experiencia se ha evaluado el rendimiento de 63 estudiantes, tanto en condiciones óptimas de aprendizaje como en condiciones con factores estresores. El conjunto de datos final incluye 25 características, siendo la mayoría series temporales con un tamaño de conjunto de datos superior a 50 millones de puntos. Los resultados muestran que la aplicación de diferentes clasificadores como KNN, SVM o Random Forest tienen una alta precisión a la hora de predecir alteraciones en el aprendizaje, mientras que la predicción del rendimiento del aprendizaje del usuario aún está lejos de ser óptima, lo que abre una nueva línea de investigación futura. Este estudio tiene como objetivo servir como línea de base para futuras mejoras en la precisión de los modelos mediante el uso de técnicas de aprendizaje automático más complejas.

KEYWORDS | PALABRAS CLAVE

Virtual environment, game-based learning, machinelearning, eye-tracking, feature extraction, neuroeducation.
Entorno virtual, aprendizaje basado en juegos, aprendizaje automático, registro de mirada, extracción de características, neuroeducación.

Received: 2022-12-26 | Reviewed: 2023-01-25 | Accepted: 2023-02-23 | OnlineFirst: 2023-05-23 | Published: 2023-07-01
DOI https://doi.org/10.3916/C76-2023-01 | Pages: 9-19

1. Introduction and state of the art

Over the past decade, lower neuro sensor costs and simpler dataacquisition and analysis techniques within different sectors have widened the scope of many final applications. Eye-tracking systems, for example, incorporate many of those techniques. The expensive customized solutions of advanced medical and even advertising research (Duchowski, 2002) have evolved into reliable commercial solutions such as high-end laptops and reasonably priced Virtual Reality Head Mounted Displays (Shadiev & Li, 2022). Compared with other neuro sensors, eye-tracking provides stable signals that describe gaze behavior, one of the main doors to the analysis of human behavior in both education and psychology (Rodero & Larrea, 2022), to name a few. Besides, eye-tracking has a powerful advantage in terms of final user acceptance: its low intrusiveness. For instance, the user can freely perform varied tasks wearing only a lightweight pair of glasses fitted with eye-tracking technology. This neurosensory device also has a drawback: it only records data on eye fixation pupil dilation and constriction. In other words, no cerebral responses to external visual objects that might cause the eye to react in one way or another are monitored.

Two promising fields of application for eye-tracking are education (García Carrasco et al., 2015) and training (Gardony et al., 2020). Eye-tracking can help to answer many questions: How do we look at learning materials depending on their multimedia presentation? How easily are we distracted? Which activities focus our attention more than others? For how long we can concentrate on a certain issue? etc. (Farran et al., 2016; Glennon et al., 2020). The answers to these questions can help teachers and trainers to better understand how we learn and how to optimize learning and training experience, to maximize learning and training outcomes. Eye-tracking can help to solve these questions in both 2D environments, i.e., screens (Añaños-Carrasco, 2015), and in 3D environments, i.e., real world and immersive Virtual Reality (iVR).

IVR environments present some challenging advantages for learning and training (Checa & Bustillo, 2020). Firstly, they offer hands-on learning: learner-centered rather than teacher-led interactive experiences. Secondly, the students learn in autonomous ways at their own pace, unlike standardized learning experiences that, in many cases, reduce learning outcomes. Thirdly, real-life difficulties may be simulated for both students and workers: from reorienting attention and dwell time in city environments (Lapborisuth et al., 2021) to awareness, prevention and detection of anxiety or depression in students (Martinez et al., 2021).

Finally, users of iVR environments have no feeling of being under observation: as the immersiveness of the experience increases within the iVR environment after a couple of minutes, the feeling of being observed decreases, prompting natural behavior. As the iVR experience can be recorded and closely monitored, user performance is more closely evaluated than it is, for example, in exam-based learning experiences. The analysis of behavior metrics can also be used for learner assessment in iVR. This VR simulation (Wismer et al., 2022) used for the assessment of compliance and physical laboratory skills accurately predicted (77%) both the expert and the novice status of the user. Collecting relevant behavioral data in VR, e.g., head and eye movement tracking, and behavior metrics data will yield more accurate results. Eye-tracking and iVR environments are, therefore, new technologies for learning and training with a challenging future, available to the general public and to specialists alike. The new Head Mounted Displays (HMDs) for immersive experiences within high-quality iVR environments record eye-tracking data in a non-intrusive way.

Up until now, eye-tracking has been used for basic actions: movements within iVR environments when physical room is limited (Sun et al., 2018), hands-free interaction within the iVR environment, such as text typing (Ma et al., 2018) and moving virtual objects (Tanaka et al., 2021). Some examples of complex tasks are prioritizing a scene according to user gaze (Patney et al., 2016) and measuring cognitive workload by means of eye-tracking, which was first investigated for a very specific task: training surgeons during analogous vesicourethral anastomosis tasks (Cowan et al., 2021). Leveraging eye-tracking technology within VR presents a novel approach to studying learner attention and motivation, while potentially improving teaching effectiveness and serving as a valuable assessment tool (Rappa et al., 2022). However, some major problems must be overcome before it can be fully implemented in learning environments that apply eye-tracking.

Firstly, efficient processing of massive iVR datasets of eye-tracked learning experiences must be demonstrated. Secondly, assuming that useful information could be identified in those datasets: could we identify the best way of learning depending on the available iVR contents? Thirdly, the most accurate techniques for extracting this hidden information should be established, considering that learning is a changing and customized process for each human being. All these questions should be answered for 3D eye-tracking, a more complex task than traditional screen-based 2D eye-tracking (Gardony et al., 2020). Eye-tracking technology has the potential to complement other data collection tools and provide distinct data sets that can enhance learning in virtual reality environments. For this purpose, machine-learning techniques might be one of the most promising solutions for all these tasks and questions (Gardony et al., 2020).

Machine-learning implies data-driven techniques used to learn from big datasets that describe complex tasks. The application of machine-learning techniques to eye-tracking datasets recorded in iVR learning environments can be for different tasks (Gardony et al., 2020). Firstly, machine-learning can perform a task commonly known as feature extraction, which is used to identify the main features of those datasets where key information is concentrated. For instance, hierarchical discriminant component analysis, a machine-learning technique, has been successfully used for eye-tracking and EEG-dataset feature extraction for gaze and attention reorientation across different gaze events (Lapborisuth et al., 2021).

Secondly, machine-learning can classify a user attention exercise and the quality of a learning environment; furthermore, on the same basis, it can predict user learning performance by comparison with previous patterns. Asish et al. (2022) roposed the use of deep learning (Convolutional Neural Networks) to classify attention in 3 exercises during an iVR learning experience based on a labelled eye-gaze dataset. Thirdly, ML may be used in a more complex architecture to adapt the learning iVR environment to the specific needs and pace of each individual user.

On the basis of the three tasks mentioned above, machine-learning can help the design of eye-tracked iVR experiences. In this research, the second task is addressed. Different machine-learning techniques were used to classify learning-environment quality and to attempt to predict user learning performance. These two objectives were then tested on a huge dataset (>50M data points) of real experiences within a realistic learning scenario where 63 students repeated a defined task and improved their performance.

Compared to a previous huge dataset (Asish et al., 2022), a labelled eye-gaze dataset with 3.4M data points, the one in this study is 15 times the size and has greater dataset diversity (different expertise levels and environmental conditions), increasing the complexity of the proposed task: from user identification to learning quality classification and the prediction of user-learning performance. Finally, the question to be answered in this research is whether eye-tracking-based datasets from iVR learning environments are suitable for the evaluation of learning conditions and learner performance by means of machine-learning. It should be outlined that this research does not aim to find a reliable and robust solution for these tasks, but a first approach that will provide a baseline for future improvements in this research strategy.

2. Material: An iVR learning environment

It is advisable to follow three steps (Figure 1) in the development of an effective iVR educational experience: pre-design, design, and evaluation (Checa & Bustillo, 2020). The first step, pre-design, establishes a scenario in which learning is enhanced through the introduction of iVR technologies. In this research, an iVR environment for learning how to operate a bridge crane has been created. The bridge crane is used in many industrial and transport-related processes. Remote-control operation means that iVR simulators can closely mirror industrial tasks. An iVR training experience acquires user performance data during exercises to test expertise and is designed to be short, easy to learn, and repeatable.

Once the learning objectives are fixed, it is necessary to apply a pedagogical approach and to take learning theories into account during the design phase. Learning theories provide guidelines on student motivations, learning processes, and outcomes (Pritchard, 2017). This experience seeks to promote learning by linking iVR to a fusion of principles from multiple pedagogical perspectives. There are many learning theories developed for use in iVR experiences or that can be easily accommodated for use in these new technologies. Four learning theories were considered for this research.

Figure 1. Stages followed for developing the iVR learning experience

(I) Pre-design: Establishing a scenario in which learning is enhanced by the introduction of iVR technologies. Advantages of virtual reality versus traditional learning (Checa & Bustillo 2022a):

(II) Design: The experience is designed to achieve the highest degree of user immersion. (A) Photorealistic environment to interact with the overhead crane controller. (B) Teleport point for fixation movement. (C) Virtual actor guiding the user. (D) User performing the experience with an HMD.

(III) Evaluation: (A) Analyze the handling skills of an overhead crane in novice users under ideal conditions. (B) Clocked environment (visual and auditory). (C) Environment with dim lighting (visual). (D) Environment with high operator traffic (visual and auditory).

Firstly, the theory of situated learning (Huang et al., 2010) that employs a constructivist approach, in so far as students learn professional skills by actively participating in an iVR experience. Secondly, the technological perspective of the 3D Virtual Learning Environments (Dalgarno & Lee, 2010), according to which students learn through autonomous interaction, hands-on learning, and problem solving.

Thirdly, the embodied cognition framework (Wilson, 2002) where there is a connection between our motor and visual senses; therefore, the more explicit the connection, as within iVR experiences, the easier the learning becomes. Finally, the theoretical underpinning of Dale's cone of experience (Dale, 1946) holds that students learn best when they go through a real experience, or the experience is realistically simulated. The proposed iVR learning environment offers a realistic experience in which to practice these principles and a safe environment where some mistakes can be corrected.

The second step of this methodology is the design phase. The experience is designed to achieve the highest degree of user immersion. Immersion is the subjective impression of participating in a realistic experience and involves the willing suspension of disbelief. The design of immersive learning experiences that induce this disbelief draws on 1) sensory, 2) action-oriented, and 3) symbolic factors (Dede, 2009). Related to sensorial factors, the goal is to replace real-world sensory information with synthetic stimuli, such as 3D visual imagery, spatialized sound, and force, or tactile responses (Bowman & McMahan, 2007). Related to action-oriented factors, action immersion is a way of empowering the participant in an experience where actions can be initiated that replicate those of the real world. The experience is designed to allow intuitive and natural actions. These interactions were developed with the support of a previously created framework (Checa et al., 2020). The framework simplifies the development

process with functions and services that are pre-programmed for their effective reuse. Remote control of an overhead crane is the primary means of interaction between the user and the application. The user can grab the controller with either hand and press the buttons that control the movement of the bridge crane with the other hand as shown in the video presentation of the simulator (Checa & Bustillo, 2022). Furthermore, the user is able to move within the available space of its current reality, approximately 3x3 meters. However, it was found that the user required additional space to complete the proposed exercise, so a movement system based on fixations was created. Four teleportation points were arranged as shown in Figure 1 (II-B).

Finally, considering the symbolic factors, the activation of semantic and psychological associations is essential for symbolic immersion of the participant in the content of the experience. A real situation that is recreated in a digital version deepens the immersive experience. In this case, in order to encourage these associations, the scenario, shown in Figure 1(II), was designed to be photorealistic. Unreal Engine, a graphics game engine compatible with the selected HMD, was used for the creation of this educational iVR experience.

The evaluation is the last phase of the development of this educational iVR experience. In this research, the skills of novice users when operating a bridge crane under ideal conditions and with external aspects that affect visual or auditory performance were analyzed. To do so, different environments were created in which the task to be performed was always the same, changing only certain external aspects that affected performance. The proposed task consisted of moving a bridge crane hook towards a barrel at a starting position, hooking up the barrel, and completing the proposed course within the shortest possible time, while trying not to knock down any cones. Figure 1 (III) shows the different factory premises where the task was performed. Figure 1(III-A) corresponds to ideal conditions, Figure 1(III-B) to the clocked environment (visual and auditory), Figure 1(III-C) to an environment with dim lighting (visual), and Figure 1(III-D) to an environment with high operator traffic (visual and auditory). It must be mentioned that a short experience with simple objectives was designed where different unforeseen factors could easily be introduced as disturbances. With this strategy, users can test the experience more than once within a short time, recording different levels of expertise as they quickly learn, by repetition and under different learning conditions, as the number of disturbances increased. Different data types, presented in Section 3.2. were automatically collected for this evaluation.

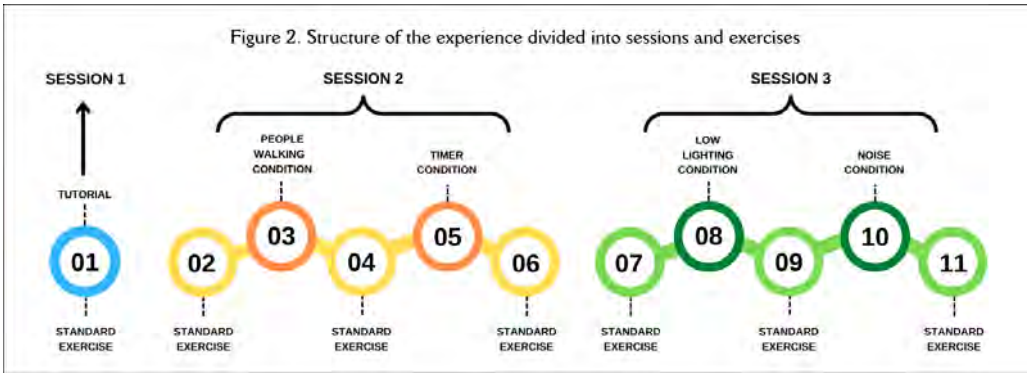
3. Learning experiences and dataset as the method

In this section, the participants and their learning experiences, as well as the data on the learning experiences are described.

3.1. Learning experiences

The learning experiences were split into 3 sessions performed in consecutive weeks for data collection. The structure of the entire experience is shown in Figure 2.

Figure 2. Structure of the experience divided into sessions and exercises



In the first session (Session 1 in Figure 2), the participants performed an iVR tutorial to learn to use the basic controls of the bridge crane and to become familiar with the iVR environment. They then

completed the standard exercise of the educational iVR experience described in Section 2. In this exercise, participants had to operate the bridge-crane so that the barrel was hooked up and transported through a circuit between cones without the load falling and without knocking down any cones. The exercise ended when the user left the load at the end of the circuit. This standard exercise was repeated in the following exercises to improve the skills of the participants at controlling the bridge-crane.

A week later, the second session took place, which consisted of 5 exercises (Session 2 in Figure 2), the first, third, and fifth of which were standard exercises of an educational iVR experience. In the second one, the user controlling the bridge-crane had to follow safety procedures when operatives were walking through the factory. In the fourth exercise, the sound of a factory bell was included that might be stressful for operator performance.

Finally, the last 5 exercises (Session 3 in Figure 2) formed the third session. The standard routine was repeated in the first, third, and fifth exercises. In the second one, lighting conditions worsened, which meant operating the bridge crane was more difficult. Finally, potentially stressful background noises within the factory while operating the bridge crane were added that could affect performance in the fourth exercise. Furthermore, to finish the whole experience, all participants were invited to complete a satisfaction survey. The purpose of gathering this information was to study whether the above-mentioned factors influenced the results of the participants.

The sample consisted of 63 students (56% female) of third-year Audio-visual Communication Degree or first-year Communication and Multimedia Design Master's Degree. The mean age of the sample was 22.3 years old (SD=2.15), and all participants performed the three sessions under the same conditions.

The setup used for the three sessions consisted of three desktop computers equipped with Intel Core i7-10710U, 32GB RAM and NVIDIA GTX 2080 graphics cards connected to HTC Vive Pro Eye HMDs and their hand-controllers (see Figure 1D). These experiences were all performed while following Spanish regulations to prevent the transmission of COVID-19. The approved Burgos University Bioethics Committee protocol was followed for data collection in compliance with data protection (Reference Number: UBU 01/2022).

3.2. Dataset description

A dataset was created to collect data from the experience described in Section 2. It included two types of data: 1) global data from each exercise; and 2) user performance data. The dataset is summarized in Table 1 (https://bit.ly/3nOpd5G). For the global data, the selected attributes were: user identifier (ID); time (T) spent on task; collision faults (F); and number of times two buttons simultaneously pressed on crane control (Pb).

The user performance data consisted of 15 inputs or attributes within the iVR environment related to position and rotation of: the crane ($Cp_{x,y,z}$ and $Cr_{y,z}$), the load ($Lp_{x,y,z}$ and $Lr_{x,y,z}$), and the user's head ($Hp_{x,y,z}$ and Hr_x). Moreover, 10 inputs were extracted from the eye-tracking system: gaze focus position ($Fp_{x,y,z}$); distance between user and focal point (D); eye openness (EL_o and ER_o); and pupil position ($PLp_{x,y}$ and $PRp_{x,y}$). Those last 25 inputs were temporal series acquired at 120 Hz.

Figures in the left column of Table 1 show the temporal evolution of one input (Lp_x) for all users for the 2^{nd}, 8^{th}, and 11^{th} exercises, showing no possibility for traditional data analysis to extract immediate information from them. The Experience number (Xn) and the user Performance (P) results are also shown in bold in Table 1, variables that will be considered as outputs or classes for the prediction models, as will be explained in Section 4.1.

4. Analysis and findings

Having recorded the data from the learning experiences, totalling 693 exercises, the machine-learning modelling was performed in several stages. First, the data were labelled. Then, the data underwent pre-processing (data encoding, handling missing values and outliers and normalization), visualization, and feature selection, before input into the machine-learning algorithms. Finally, the application of classification techniques was tested. Figure 3 summarizes these stages.

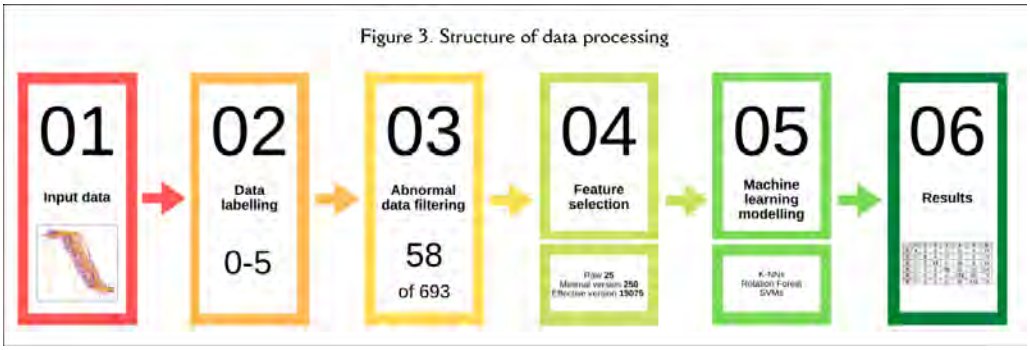


Figure 3. Structure of data processing

4.1. Data labelling

In real-world environments, objective monitoring of learner performance to determine whether a learner is ready to perform a certain task or whether they need more preparation can be difficult. However, metrics such as completion time and accuracy can be objectively recorded in a virtual environment. These metrics have been used in this study as a learner performance measure.

The selected learner performance measure, labelled as an integer value between 0 and 5, was based on two parameters: 1) completion time; and 2) collision faults. The faults were calculated from the number of cones knocked over during the exercise and the number of times that several buttons of the controller were pressed at the same time. Both parameters were rated from 0 to 5, and the minimum of both was assigned as the final performance metric. As for the labels used for the learning context, they correspond to those used to identify each exercise and have been described in Figure 2.

Each exercise described in Section 3.1 for each user was considered a single sample, so a performance evaluation was assigned for each user and exercise performed. Each user performed 11 sessions and 63 users participated in the experiment, so the original dataset was composed of 693 samples. The distribution of each dataset label in the sample was as follows: label 0, 3%; label 1, 5%; label 2, 12%; label 3, 26%; label 4, 35%, and label 5, 19%. A clear unbalance, especially in classes 0 and 1, was observed- a natural result considering that users quickly learn the proposed task, improving their score after the first couple of sessions.

4.2. Data preparation and feature extraction

During the data-capture stage, errors may occur that are difficult to detect during the experience, and it is crucial to filter them out so that they do not introduce noise into the dataset. Software capture failures can occur due to buffer saturation, momentary sensor failure, and even circumstances such as reflections in the glasses or misalignment of the HMD. These are errors that can be detected by data visualization and then filtered.

Data pre-processing was therefore used to filter out abnormal data, prior to the machine-learning tasks. Several libraries that are widely used in the field of data science were selected for this task. On the one hand, for data visualization, the Pandas library (McKinney, 2011) and the tslearn package (Tavenard et al., 2020) were selected, due to their special design for time series analyses. As a result, 58 samples were removed from the original dataset of 693, because they showed very high abnormalities or unusual user behaviours. On the other hand, the samples were of different duration, as each learner completed the exercises within a different amount of time. Research on time series classification is usually focused on the case of uniform length series. As this work is intended to provide a baseline for future research, the time series were normalized, in this case to the maximum length (4326), making use of the Timeseries Resample function of the tslearn library which performs linear interpolation. The time series were then resampled to the longest duration (4326 datapoints), until they were all of the same length.

Secondly, feature selection of the raw data was performed. The objective of this task was to explore the amount of useful information hidden in each dataset. For this purpose, the FRESH algorithm (Christ et al., 2016) was used from the tsfresh package (Christ et al., 2018). Its library includes a wide variety of features that can be derived from raw time series data; in this case, 19075 features for each time series were

extracted. The set of features can include basic statistical attributes (peaks, highs, lows, etc.), correction measures, and evolution of a time series (white noise, trend, seasonality, autocorrelation, etc.). There are some library pre-defined dictionaries, two of which were used in this study. A lighter version, called "minimal", and a more complete one called "efficient". Feature extraction was carried out in both modes, thus obtaining two new datasets: one with the minimal version and the other with all the features, called the efficient version.

4.3. Learner performance and modelling process

The machine-learning algorithms were then used to predict learner performance and learning environment quality. The three different datasets proposed in Section 4.2 were tested: 1) the original raw data, 2) the minimal version of feature extraction and 3) the complete version (efficient version).

Three machine-learning techniques, each of a very different nature, were tested for this task: 1) k-nearest neighbours, a simple yet efficient clustering algorithm that uses proximity to make classifications or predictions about the grouping of an individual data point. The value of k defines how many neighbours will be checked (in this case k was set to 1). 2) Support Vector Machines (SVM), a complex well-established algorithm that defines a hyperplane in an N-dimensional space, with N as the number of features that distinctly classify the given data points. And 3) Random Forest (RF) an optimal diversity algorithm, which builds decision trees on different samples and uses majority voting for classification and averages for regression. The aim is to evaluate which one best predicts performance and the most suitable dataset for that classification task. The three algorithms were evaluated using the WEKA library (Hall et al., 2008).

A cross-validation scheme was selected, due to its statistical invariance for the selection of those subsets, to split the dataset into training instances and validation instances. A 10-fold cross-validation was selected due to the dataset size. The selected quality indicator was accuracy, representing the proportion of correctly classified observations over the number of total instances that were evaluated.

4.4. Results

Table 2 shows the results obtained for each of the mentioned experiments. The best results are highlighted in bold. The minimal version dataset obtained better results than the other two datasets, showing the necessity of feature selection in datasets of this sort. As for the algorithms, Random Forest was the one that clearly performed the best in both tasks.

Table 2. Environmental quality and learner performance classification results

	Exercises (Accuracy (%))			Learner Performance (Accuracy %)		
Algorithm / Dataset	RF	1-NN	SVM	RF	1-NN	SVM
Raw data	**43.56**	26.17	35.41	**42.38**	35.24	40.74
Minimal version	**44.29**	31.12	42.34	**59.31**	48.20	51.98
Efficient version	**40.11**	25.84	40.72	**59.12**	48.05	51.72

Some issues should be outlined. First, the poor performance of kNN showed that the algorithms that were used to search for previous experiences with a strong similarity to the one to be predicted were unsuitable for these sorts of tasks. So, this result outlines that different levels of expertise and learning conditions increase the complexity of predicting learner performance. It is a fascinating challenge where machine-learning techniques that are especially designed for complex data structures will play a central role. Second, all the feature-extraction techniques and machine-learning algorithms that were tested provided medium-to-low prediction performance, which was hardly highly accurate, revealing a future research line for improvement. Finally, average performance values are shown in Table 2, while the performance of all classes (performance levels or exercises) was not shown. The confusion matrix for the best method, Random Forest, and both classification tasks are shown in Table 3 to analyze this issue in detail, including the percentage of correctly predicted instances on the right of each confusion matrix. The confusion matrix for the classification of the experiences (on the left of Table 3) showed that the experiences with some kind of learning limitation (noise, time pressure…) achieved high levels of accuracy (78% on average compared

with the 26% for the standard exercises); those were exercises 3, 5, 8, and 10, marked with an asterisk in Table 3. Regarding learner performance, although the model failed to give the right classification, it tended to predict classes that were close to the right ones; therefore, the system was able to classify novice students and expert students correctly. The classifications of the models were significantly better for classes 3, 4, and 5 (medium-high good performance) than for classes 0, 1, and 2 (low performance). A result that was also foreseeable, given the imbalance of the classes outlined in Section 4.1.

Table 3. Confusion Matrix for classification of experiences (left) and performance (right) - RF model

	1	2	3*	4	5*	6	7	8*	9	10*	11	%
1	**21**	11	0	3	0	0	4	0	0	0	0	54
2	12	**26**	5	2	4	3	7	0	0	0	0	44
3*	6	9	**39**	0	5	3	4	0	5	0	0	55
4	3	11	9	**10**	3	7	5	0	9	0	13	14
5*	0	0	9	0	**43**	0	0	0	0	0	0	83
6	3	7	2	15	10	**4**	3	0	7	0	17	6
7	5	8	7	5	3	2	**15**	0	10	0	5	25
8*	0	0	0	0	0	0	0	**65**	0	0	0	100
9	0	2	0	9	2	9	8	0	**7**	0	21	12
10*	0	0	4	0	5	2	3	0	2	**46**	0	74
11	0	2	0	7	2	9	11	0	7	0	**15**	28

	0	1	2	3	4	5	%
0	**4**	2	0	3	8	4	19
1	0	**9**	4	11	13	2	23
2	0	5	**13**	9	35	5	19
3	0	0	0	**78**	52	15	54
4	0	0	0	4	**172**	62	72
5	0	0	0	0	35	**112**	76

5. Discussion and conclusions

Current iVR systems generally use standardized learning methods that do not adapt to the individual characteristics of each learner. This leads to high levels of demotivation, passive attitudes, boredom, low engagement, and frustration among trainees. Eye-tracking data can play an important role in monitoring these environments and as a complement to other data collection tools, e.g., behavior metrics. The use of AI techniques on datasets extracted from iVR training environments can be the desired solution, to adapt learning iVR environments to the different backgrounds and characteristics of each learner. In this study, the way in which basic machine-learning techniques can be applied to achieve that goal has been examined, specifically to evaluate learning conditions and learner performance, within areas where the existing bibliography is specially limited. To do so, an iVR environment and a testing experience have been designed, in such a way that the students were expected to repeat a simple short task while exposed to different disturbances, learning quickly and generating a dataset with a high diversity of exercises for the expertise of each user and under different environmental conditions. Different machine-learning techniques were then tested for two tasks: 1) quality classification of the learning environment; and 2) prediction of learner performance. Well-established data-science methods were followed to test the following techniques: data labelling, data filtering, feature extraction, and machine-learning modelling under a cross-validation scheme. Among the algorithms that were tested, Random Forest showed the best accuracy for both tasks. While high accuracy was achieved for classifying abnormal learning conditions (78%), the results were not so good for prediction of learner performance (59%). It should be outlined that the aim of this research is not to find a reliable and robust solution for these tasks, but it is a first approach that will provide a baseline for future improvements for the use of machine-learning in iVR learning environments.

Compared with the existing bibliography, similar accuracy levels were achieved for quality evaluation of the environment. While in this study the expert or novice status of the user could be predicted to an accuracy of 77% in an iVR simulation (Wismer et al., 2022) for the measurement of laboratory skills and learner assessment and compliance using behavior metrics, accuracy levels of 78% were achieved while rating the quality of the learning environment. Compared to the evaluation of attention or distraction (Asish et al., 2022), model accuracy was lower; a difference that arises from the definition of classes in both works: while up to 6 levels were used in this research, Asish et al. (2022) used a binary classification, that usually yields higher levels of accuracy. Finally, compared with the classification of driving (Deng et al., 2020), some common conclusions have been achieved in this work: the stability and high accuracy of

ensemble techniques, like Random Forest, over other classical algorithms, like kNN, or SVMs. Again, the high accuracy achieved in this work (up to 89%) might come from the selection of only 3 classes and the strong difference in behavior between drivers in each class. As was also outlined in those previous works, the extension of the datasets, in terms of learners and conditions, is required to achieve higher accuracy. Nevertheless, the suitability of machine-learning for the performance of such tasks has been confirmed in this research, in so far as one of the largest datasets more than 50M data points was processed far more efficiently than conventional human-based data-processing techniques.

Future studies could be focused on improving the accuracy of prediction models for learning evaluation in iVR environments. An aim that could be achieved by expanding the dataset to include experiences from new users, improving the labelling methodologies, and utilizing balancing techniques for highly unbalanced classes (such as the SMOTE algorithm). Additionally, alternative machine-learning techniques could be tested, such as Hidden Markov Models with proven results for time series, in order to capture the dynamic trends of learner performance. Furthermore, the results have motivated the need to add session-related information to the dataset, so that intra- and inter-session learner performance patterns could be extracted.

Authors' Contribution

Idea, C.P.A.; Literature review (state of the art), D.C.; Methodology, I.M.A., D.C.; Data analysis, A.S.M., C.P.A.; Results, A.S.M.; Discussion and conclusions, A.S.M., I.M.A.; Writing (original draft), C.P.A., I.M.A.; Final revisions, C.P.A., I.M.A.; Project design and sponsorship, D.C., A.S.M.

Funding Agency

This work was partially supported by the ACIS Project (Reference Number: INVESTUN/21/BU/0002) of the Consejeria de Empleo e Industria of the Junta de Castilla y León (Spain), the Erasmus+ RISKREAL Project (2020-1-ES01-KA204-081847) of the European Commission, the HumanAid Project (TED2021-129485B-C43) of the Spanish Ministry of Science and Innovation, and the Margarita Salas program of the Spanish Ministry of Universities funded by NextGenerationEU.

References

Añaños-Carrasco, E. (2015). Eyetracker technology in elderly people: How integrated television content is paid attention to and processed. [La tecnología del «EyeTracker» en adultos mayores: Cómo se atienden y procesan los contenidos integrados de televisión]. *Comunicar*, *45*, 75-83. https://doi.org/10.3916/C45-2015-08

Asish, S.M., Kulshreshth, A.K., & Borst, C.W. (2022). Detecting distracted students in educational VR environments using machine learning on eye gaze data. *Computers & Graphics*, *109*, 75-87. https://doi.org/10.1016/j.cag.2022.10.007

Bowman, D.A., & Mcmahan, R.P. (2007). Virtual reality: How much immersion is enough? *Computer*, *40*(7), 36-43. https://doi.org/10.1109/MC.2007.257

Checa, D., & Bustillo, A. (2020). A review of immersive virtual reality serious games to enhance learning and training. *Multimedia Tools and Applications*, *79*, 5501-5527. https://doi.org/10.1007/s11042-019-08348-9

Checa, D., & Bustillo, A. (2022). *Grua Rv*. http://3dubu.Es/En/Cranevr/

Checa, D., Gatto, C., Cisternino, D., De Paolis, L.T., & Bustillo, A. (2020). A Framework for Educational and Training Immersive Virtual Reality Experiences. In L. T. de Paolis, & P. Bourdot (Eds.), *Augmented reality, virtual reality, and computer graphics* (pp. 220-228). Springer International Publishing. https://doi.org/10.1007/978-3-030-58468-9_17

Christ, M., Braun, N., Neuffer, J., & Kempa-Liehr, A.W. (2018). Time series feature extraction on basis of scalable hypothesis tests (tsfresh - A Python package). *Neurocomputing*, *307*, 72-77. https://doi.org/10.1016/j.neucom.2018.03.067

Christ, M., Kempa-Liehr, A., & Feindt, M. (2016). Distributed and parallel time series feature extraction for industrial big data applications. *ArXiv*, *1*. https://doi.org/10.48550/arXiv.1610.07717

Cowan, A., Chen, J., Mingo, S., Reddy, S.S., Ma, R., Marshall, S., Nguyen, J.H., & Hung, A.J. (2021). virtual reality vs dry laboratory models: Comparing automated performance metrics and cognitive workload during robotic simulation training. *Journal of Endourology*, *35*(10), 1571-1576. https://doi.org/10.1089/end.2020.1037

Dale, E. (1946). *Audiovisual methods in teaching*. Dryden Press. https://bit.ly/42aW03X

Dalgarno, B., & Lee, M.J.W. (2010). What are the learning affordances of 3-D virtual environments? *British Journal of Educational Technology*, (1), 41-41. https://doi.org/10.1111/j.1467-8535.2009.01038.x

Dede, C. (2009). Immersive interfaces for engagement and learning. *Science*, *323*(5910), 66-69. https://doi.org/10.1126/science.1167311

Deng, Q., Wang, J., Hillebrand, K., Benjamin, C.R., & Soffker, D. (2020). Prediction performance of lane changing behaviors: A study of combining environmental and eye-tracking data in a driving simulator. *IEEE Transactions on Intelligent Transportation Systems*, *21*(8), 3561-3570. https://doi.org/10.1109/TITS.2019.2937287

Duchowski, A.T. (2002). A breadth-first survey of eye-tracking applications. *Behavior Research Methods, Instruments, & Computers*, *34*(4), 455-470. https://doi.org/10.3758/BF03195475

Farran, E., Formby, S., Daniyal, F., Holmes, T., & Herwegen, J. (2016). Route-learning strategies in typical and atypical development; eye-tracking reveals atypical landmark selection in Williams syndrome: Route-learning and eye-tracking. *Journal*

of Intellectual Disability Research, *60*(10), 933-944. https://doi.org/10.1111/jir.12331

García-Carrasco, J., Hernández-Serrano, M.J., & Martín-García, A.V. (2015). Plasticity as a framing concept enabling transdisciplinary understanding and research in neuroscience and education. *Learning, Media and Technology*, *40*, 152-167. https://doi.org/10.1080/17439884.2014.908907

Gardony, A.L., Lindeman, R.W., & Brunyé, T.T. (2020). Eye-tracking for human-centered mixed reality: Promises and challenges. *Proc.SPIE*, *11310*, 113100T. https://doi.org/10.1117/12.2542699

Glennon, J.M., Souza, H., Mason, L., Karmiloff-Smith, A., & Thomas, M.S.C. (2020). Visuo-attentional correlates of Autism Spectrum Disorder (ASD) in children with Down syndrome: A comparative study with children with idiopathic ASD. *Research in Developmental Disabilities*, *104*, 103678. https://doi.org/10.1016/j.ridd.2020.103678

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. (2008). The WEKA data mining software: An update. *SIGKDD Explor. Newsl*, *11*(1), 10-18. https://doi.org/10.1145/1656274.1656278

Huang, H.M., Rauch, U., & Liaw, S.S. (2010). Investigating learners' attitudes toward virtual reality learning environments: Based on a constructivist approach. *Computers and Education*, *55*(3), 1171-1182. https://doi.org/10.1016/j.compedu.2010.05.014

Lapborisuth, P., Koorathota, S., Wang, Q., & Sajda, P. (2021). Integrating neural and ocular attention reorienting signals in virtual reality. *Journal of Neural Engineering*, *18*(6). https://doi.org/10.1088/1741-2552/ac4593

Ma, X., Yao, Z., Wang, Y., Pei, W., & Chen, H. (2018). Combining brain-computer interface and eye-tracking for high-speed text entry in virtual reality. In *IUI '18: 23rd International Conference on Intelligent User Interfaces* (pp. 263-267). https://doi.org/10.1145/3172944.3172988

Martinez, K., Menéndez-Menéndez, M.I., & Bustillo, A. (2021). Awareness, prevention, detection, and therapy applications for depression and anxiety in serious games for children and adolescents: Systematic review. *JMIR Serious Games*, *9*(4). https://doi.org/10.2196/30482

Mckinney, W. (2011). *pandas: A foundational Python library for data analysis and statistics*. Python High Performance Science Computer.

Patney, A., Salvi, M., Kim, J., Kaplanyan, A., Wyman, C., Benty, N., Luebke, D., & Lefohn, A. (2016). Towards foveated rendering for gaze-tracked virtual reality. *ACM Trans. Graph*, *35*(6), 1-12. https://doi.org/10.1145/2980179.2980246

Pritchard, A. (2017). *Ways of learning: Learning theories for the classroom*. Routledge. https://doi.org/10.4324/9781315460611

Rappa, N.A., Ledger, S., Teo, T., Wong, K.W., Power, B., & Hilliard, B. (2022). The use of eye-tracking technology to explore learning and performance within virtual reality and mixed reality settings: A scoping review. *Interactive Learning Environments*, *30*(7), 1338-1350. https://doi.org/10.1080/10494820.2019.1702560

Rodero, E., & Larrea, O. (2022). Virtual reality with distractors to overcome public speaking anxiety in university students; [Realidad virtual con distractores para superar el miedo a hablar en público en universitarios]. *Comunicar*, *72*. https://doi.org/10.3916/C72-2022-07

Shadiev, R., & Li, D. (2022). A review study on eye-tracking technology usage in immersive virtual reality learning environments. *Computers & Education*. https://doi.org/10.1016/j.compedu.2022.104681

Sun, Q., Patney, A., Wei, L.Y., Shapira, O., Lu, J., Asente, P., Zhu, S., Mcguire, M., Luebke, D., & Kaufman, A. (2018). Towards virtual reality infinite walking: Dynamic saccadic redirection. *ACM Transactions on Graphics*, *37*(4), 1-13. https://doi.org/10.1145/3197517.3201294

Tanaka, Y., Kanari, K., & Sato, M. (2021). Interaction with virtual objects through eye-tracking. In *International Workshop on Advanced Imaging Technology (IWAIT) 2021* (pp. 1176624). SPIE. https://doi.org/10.1117/12.2590989

Tavenard, R., Faouzi, J., Vandewiele, G., Divo, F., Androz, G., Holtz, C., Payne, M., Yurchak, R., Rußwurm, M., Kolar, K., & Woods, E. (2020). Tslearn, A Machine-learning Toolkit for Time Series Data. *J. Mach. Learn. Res*, *21*, 1-6.

Wilson, M. (2002). Six views of embodied cognition. *Psychonomic Bulletin and Review*, *9*, 625-636. https://doi.org/10.3758/BF03196322

Wismer, P., Soares, S.A., Einarson, K.A., & Sommer, M.O.A. (2022). Laboratory performance prediction using virtual reality behaviometrics. *PloS One*, *17*(12). https://doi.org/10.1371/journal.pone.0279320