RedFame

Journal of Education and Training Studies
Vol. 4, No. 1; January 2016
ISSN 2324-805X E-ISSN 2324-8068
Published by Redfame Publishing
URL: http://jets.redfame.com

LewiSpace: an Exploratory Study with a Machine Learning Model in an Educational Game

Ramla Ghali[1], Sébastien Ouellet[1], Claude Frasson[1]

[1]Université de Montréal, Département d'informatique et de recherche opérationnelle, 2920 Chemin de la Tour, H3C 3J7 (QC), Canada

Correspondence: Ramla Ghali, Université de Montréal, Département d'informatique et de recherche opérationnelle, 2920 Chemin de la Tour, H3C 3J7 (QC), Canada

Received: August 6, 2015 Accepted: August 20, 2015 Online Published: October 19, 2015

doi:10.11114/jets.v4i1.1153 URL: http://dx.doi.org/10.11114/jets.v4i1.1153

Abstract

The use of educational games as a tool for providing learners with a playful and educational aspect is widespread. In this paper, we present an educational game that we developed to teach a chemistry lesson, namely drawing a Lewis diagram. Our game is a 3D environment known as LewiSpace and aims at balancing between playful and educational contents in order to increase engagement and motivation while learning. The game contains mainly five different missions aim at constructing Lewis diagram molecules which are organized in an ascending order of difficulty. We also conducted an experiment to gather data about learners' cognitive and emotional states as well as their behaviours through our game by using three types of sensors (electroencephalography, eye tracking, and facial expression recognition with an optical camera) and a self report personality questionnaire (the Big Five). Primary results show that a machine learning model namely logistic regression, can predict with some success whether the learner will success or fail in each mission of our game, and paves the way for an adaptive version of the game. This latter will challenge or assist learners based on some features extracted from our data. Feature extraction integrated into a machine learning model aims mainly at providing learners' with a real-time adaptation according to their performance and skills while progressing in our game.

Keywords: educational game, electroencephalogram, eye tracking, facial expression recognition, logistic regression model, big five questionnaire

1. Introduction

In Human Computer Interaction (HCI), the Intelligent Tutoring Systems (ITS) were among the first sophisticated learning systems. These systems are characterized by their capacity to provide a continuous feedback, hints, helps, etc. to learners. The adaptation was done mainly by using intelligence artificial techniques and more specifically machine learning algorithms. The use of the latter provides these systems with the 'intelligence' criterion. Nowadays, the ITS have progressed and moved to another type of environment since 2002: educational games or serious games (SGs). SGs are video games that aim to inform, test and train people while playing. They can be applied in several fields: military, government, education, business, health care, etc. Recently, they are more used for educational reasons (Conati 2002, Ghali et al. 2014, Jackson et al. 2012, Rowe et al. 2009) through their playful aspect, known as 'Game based Features' (McNamara et al. 2010).

Although this environment presented a very appropriate and quite moderate way of learning, the problem of user adaptation according to educational aspect remains of great importance. We think that researchers should focus more on this problem in this type of environment due to its difference with ITS. However, to our knowledge, only few works are interested to develop some adaptive user models in SGs and very few works are interested to provide learners' with real time adaptation which reacts while interacting with the game. However, we believe that this issue presents a very important factor to consider in SGs since that their main goal in educational applications is to improve and focus more on the educational and pedagogical content and not the playful aspect which does not contribute to learning improvement process. Therefore, on one hand we think that the adaptation must more focus on the type and nature of help to supplement and complement the educational content of the game. On the other hand, we think that this latter should be instantly and in real time. So, the resulting environments (SGs) should react immediately and/or according to

user's needs and learning capacity. Moreover, many works focus more on the playful aspect which increases motivation and pleasure but not necessary contributes to improving the learning process which depends mainly on the pedagogical contents. In addition, almost games present a serious problem when translating from game situations to non-game contexts. Furthermore, the students generally spend too much time in using entertainment and playful aspects rather than practicing educational contents.

In order to solve this problem and to develop more effective real time adaptive tools which consist of taking learners' differences (a Big Five personality was administrated in our work) and focus more of detecting when users need really more pedagogical help in SGs, we proposed in this paper a first version of a 3D puzzle game called **LewiSpace**. We hypothesize that this game will focus more on learning how to draw Lewis diagrams rather than playful features (which are available when navigating in a 3D environment, changing backgrounds' colors with the different places and gathering the requested atoms to construct complex molecules). Our goal with the study described in the current paper is to investigate whether it is possible to predict a learner's success and his desired level of help based on information gathered through different types of data: electroencephalography (The Affectiv Suite from EPOC Emotiv), eye tracking (the indice of workload extracted from pupil diameter), facial expression recognition (through FaceReader software) and self-report big five personality questionnaire. Since this is part of a larger project that aims to develop a game that will be able to adapt in real-time to learners, we first studied in this paper the descriptive results obtained, the importance of each sensor used and how it improves prediction of learner's success or fail in each mission of our game. We also studied the utility of combining different types of data and the necessity of using them to build the most appropriate real-time users' adaptation.

This paper is structured as follows: in the next section, we describe some related works and mention the disadvantages of the existing works. Next, we describe LewiSpace game that is designed to teach how to construct Lewid diagrams for some complex molecules. Next, we describe the experiment that we conducted in order to gather data and their pre-processing stage. Finally, the last section presents some descriptive results about learners' performance distribution in the different missions of our game. We also provide a comparison and a discussion about the different real time statistical machine learning techniques used in this study, how we extract the features from our multimodal kinds of data and how we select the best approach, more specifically concerning the real time machine learning algorithm, the features and the hyper parameters to take into consideration for this type of applications.

2. Previous Works

Recently, the use of educational games or serious games became widespread. These games are beneficial for learning because they incorporate two fundamental aspects: (1) educational aspect interested to learning content and strategies to present to learners, and (2) playful aspect that allows learners to play, explore, take rewards, control the environment, etc. In fact, researchers believe that this last aspect can increase learners' motivation and engagement (McNamara et al. 2010, Lester et al. 2014, Ghali et al. 2014). This aspect is also known as 'Game-Based Features' (McNamara et al. 2010). Moreover, Prensky, Johnson and Wu (Prensky 2001, Johnson et al. 2008) agree that educational games have not only playful aspects but several criteria and characteristics to increase exploration, immersion and motivation aspects.

According to McNamara and Jackson (McNamara et al. 2010), games based features could be grouped into five main categories: (1) **Feedback** which consists at providing learners with a specific, intelligent and motivational feedback; (2) **Incentives** which aims at promoting the aspects of bonuses and rewards. The latter are related to extrinsic motivation and have a direct effect on learners' self-efficacy, engagement and interest; (3) **Task difficulty** which consist at varying the difficulty of a task and adjust it according to learners' skills; (4) **Control** which allows the learner to monitor and manage the environment such as changing the color of background or avatar and finally (5) **Environment** which focuses at the design and the type of the environment.

Despite the last criteria proposed by (McNamara et al. 2010) to develop more effective educational games, the latter present several problems. Among them, we cite briefly the problem of spending too much time for playing instead of entertainment and learning, the problem of the construction and the order of pedagogical content, the problem of translating between playful and educational aspects, etc. To solve these problems, we suggest that more research will be done in the field of SGs and that this latter should be more intelligent. The intelligence criterion consists of offering to user a real time, continuous, and individualized adaptation according to learning content. We define this type of game as Intelligent Educational Games (IEG). Whereas, to date, only some works are interested to automatic (but not real time) user modeling and/or adaptation either in tutoring systems or educational games (Lester et al. 2014, D'Mello et al. 2012, Gobert et al. 2015, McQuiggan et al. 2006). The adaptation or modeling is usually done using some learners' criteria (such as emotions, engagement, motivation, workload, self-efficacy, performance, etc.). They could also be classified into two kinds of groups: (1) works based on the learner's interactions with the system and (2) works based on the electro-physiological sensors. Among these works, we cite as an example those of Gobert, Baker and his team which are

interested to automatically detect learner's disengagement (Gobert et al. 2015). They build an automatic machine learning detector of disengagement behavior. Their model is based on human labels of behaviors from log files and data mining techniques. Lester and colleagues (Lester et al. 2014) used Elliot and Pekrun model (Elliott et al. 2007) to automatically predict and adapt learners' emotions. This model has been empirically used with learners' interaction data with the system which are derived from a subjective method of self-assessment of emotions. Emotions are recorded from learners using a portable device (smartphone game device) every seven seconds. D'Mello and colleagues (D'Mello et al. 2012) have used eye tracking data to automatically detect emotions of boredom and disengagement among learners in interactions with a tutoring system. Automatic tracking of eye movements was integrated into a tutor that identify when a learner is bored, looking or zooming on the screen.

Recently, Jaques and colleagues (Jaques et al. 2014) used also gaze data features in order to predict two main emotions: boredom and curiosity. These emotions are predicted from several machine learning and feature selection algorithms collected from students' self-reported emotions in Meta tutor system (Azevedo et al. 2010). They obtained an accuracy of 69% for boredom and 73% for curiosity. Finally, (MCQuiggan et al. 2006) used decision trees and Bayesian networks to generate predictive models of self-efficacy. They obtained two families of models: (1) **static models** that are based on the demographics of the student from pre-test self-efficacy, and (2) **dynamic models** that combine static data model and physiological data (heart rate and skin conductance). The authors have shown that static models predict self-efficacy of students with an acceptable accuracy rate (73%). However, the dynamic models allow a prediction of self-efficacy with better accuracy rate (83%).

Although these works present a very important way to automatically detect some learners' negative behaviours or emotions which are not effective for learning, we were not interested in this paper to predict learners' emotions because our designed game LewiSpace is not emotionally engaged but focus more on educational aspect. The game also detects automatically learners' emotions through FaceReader software. We use this sensor to extract seven basic emotions (happy, sad, angry, surprised, scared, disgusted, and neutral) in addition to the valence and arousal of each emotion (Lewinski et al. 2014). In (Chaffar et al. 2006) we have also anticipated learners'emotional response using EEG techniques. We also complete the miss detection of some emotions due mainly to mouth occlusion by using the Affectiv Suite provided by Emotiv EEG sensor (Gheeguluscu et al. 2014). Whereas, the usage of gaze data (Tobii Tx300 sensor) more precisely pupil diameter is to measure learner's state of workload (Bartels et al. 2012) while interacting with our educational game. The following section describes LewiSpace, a 3D educational puzzle game.

3. LewiSpace: An Educational Puzzle Game

3.1 A Description and Design of the Environment

LewiSpace is an educational game which aims mainly to teach learners how to construct chemical structures of molecules using Lewis diagrams (Ghali et al. 2015). The game is mainly designed to be explored by college students who didn't have any knowledge about how to build Lewis' diagrams (a chemistry lesson). It has an exploratory environment (3D) developed using Unity 4.5 for the design of the game, integrating EEG and eye tracking sensors data using the Emotiv SDK v2.0 LITE and the Tobii SDK 3.0.

In the first user's interaction with the game, the player is simulated as an astronaut exploring a planet's surface and communicating with a non-player character known as Commander Arnold (figure 1). The player is told that he fell down into a cavern and that he has to explore the underground where he has each time to overcome obstacles (obstructions, lack of a useful resource, etc.) in order to progress in the game and find his lander, allowing him to return home. The player starts by exploring the environment which is mainly composed of five types of scenes leading to five different missions to accomplish during the game. The missions are constructed in ascending order of difficulty according to the complexity of molecules' structures and the player can't progress to the next mission before completing the latest one. By exploring our educational game, the player accumulates a certain number of atoms (which are hidden somewhere on the environment) that he adds to his inventory and can use them further in order to construct chemical compounds. The latter are used to unlock paths and move to another stage in the game. In the following, we will describe and present some screenshots for the different missions of our game.

Figure 1. The beginning of the game

3.1.1 Mission 1: Produce Water

After starting by exploring the environment and seeing tanks of hydrogen and oxygen, the player is invited to produce H_2O molecule (water) in order to refill his spacesuit's thermal regulation system. He has to place atoms in a grid by clicking on them and deduce the correct Lewis diagram structure (figure 2) according to rules that are presented to him throughout the game (see section 3.2), in the manner of a puzzle.

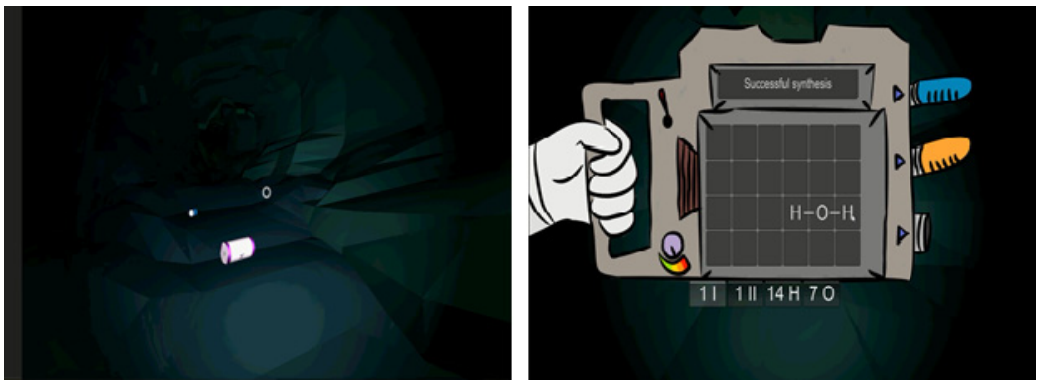

Figure 2. Atoms in the environment (left) and using a Lewis diagram tool to produce H2O molecule (right)

3.1.2 Mission 2: Warm and Destruct a Tunnel

In this mission, the player has to gather carbon and hydrogen in order to produce methane gas (CH_4). The player is provided with the information about the group of each atom and also the periodic table that represents the structures of atoms according to their group's number and atomic number of each one during all the missions. He has to find the number of electrons available on the last atom's layer (the valence shell) and to construct the right diagram according to the octet rule and the valence electrons indicated by the periodic table. After producing the methane, an animation is shown to the player, allowing the player to melt the tunnel's obstruction with a methane torch (figure 3).

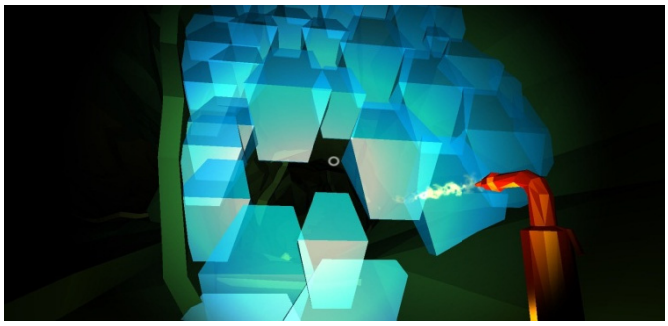

Figure 3. Melting a tunnel's obstruction using a methane torch

3.1.3 Mission 3: Dissolve a Metal Debris

The player has to construct a more complex molecule which lets him dissolve metal debris. The compound to produce is sulfuric acid (H_2SO_4). This structure has 32 electrons to distribute between the atoms and the player has to move some electrons in order to construct double bonds. We noticed that until this stage, all the three missions could be presented by a symmetrical diagram Lewis structure.

3.1.4 Mission 4: Craft a Refrigerant

Here, the player has to craft a refrigerant (C_2F_3Cl) and use it for his spacesuit to regulate his body temperature. This compound could be seen as less complex than the previous one, but it is the first one to present an asymmetrical structure.

3.1.5 Mission 5: Fill in the Fuel Cell with Ethanol

In the last mission, the player is out of the cavern. He finds his lander module on the surface but its fuel cell is empty. As a final task, the player has to gather and construct ethanol (C_2H_6O). As soon as this is done, the rocket takes off. At this stage, the game is over (see figure 4).

Figure 4. The end of the game

3.2 A Description of the Given Instructions and Rules

In this section, we will focus more on how we present the educational materials to learners while they progress on the game. Hence, to motivate the students with a playful aspect, we think that this latter is covered when they are navigating on a 3D environment. Every time, they have to look everywhere in the cavern structure to find the appropriate atoms to gather that allow them to construct the requested molecule. The design and the colors of the environments are also attractive and different from each other's. However, the educational aspect is covered by announcing the rules to use on some missions and the directives mentioned in order (see table 1). The learner has also the option to see the periodic table at any time while building a molecule by pressing a shortcut button that enables him to open or close this informative tool.

Table 1. Instructions and rules presented in LewiSpace mentionned according the missions

Missions	Instructions
Mission 1	- Hydrogen atoms can only bond once. This is because a single covalent bond involves one pair of electrons, and hydrogen needs 2 electrons to be full. This is an exception, as other atoms need 8 electrons. This is known as the octet rule, atoms tend to combine to satisfy it.
Mission 2	- Double covalent bonds involve 2 pairs of electrons. You can figu-re out if single or double bonds are needed with the octet rule and with the number of valence electrons the atoms have. - Open your Periodic Table by pressing I. Each column (except the pink-colored ones) group atoms by their number of valence elect-rons. For example, hydrogen has 1, calcium (Ca) has 2, aluminum (Al) has 3, fluorine (F) has 7. - When crafting a compound, each single bond you add represents 2 electrons, shared between two atoms. If an atom doesn't have 8 electrons after you sum up its lone electrons and those shared through bonds, you might have to add double bonds or to redraw the structure.
Mission 3	- It's often important to consider formal charges when drawing structures. You can calculate each atom's formal charge by sub-stracting each bond (1 for a single, 2 for a double one) and each lone electron from its initial number of valence electrons. - If the formal charge is not zero, you might be able to reconfigure the diagram (change the shape or the type of bonds). The octet rule can be violated in some cases. - Also, keep in mind that elements in the third row of the Periodic Table can sometimes hold more than 8 electrons
Mission 4	- No new rules
Mission 5	- No new rules

As we mentioned before, the learner is provided at any time he wants by an informative tool, the standard periodic table. This tool describes for each atom the symbol, the atomic number, the mass number or the number of nucleons and the group indicated on the top of each column.

4. Experiment and Data Preprocessing

4.1 Experiment

In order to gather data from eye tracking, electroencephalogram (EEG) sensors and learners' emotions, we conducted an

experiment where 40 participants (25 males and 15 females aged between 19 and 35 years) from Montreal University participated voluntarily in the study (with a compensation of 20$ for each participant). As a criterion for admissibility, we requested students who have no prior knowledge about Lewis Diagrams (a chemistry course supposed to be learned to college students). The study was held under strict laboratory conditions. Once we explained the whole process of the study and the participant signed the ethics agreement, the participant is invited to start our experiment. During the experiment, EEG is recorded with the Emotiv headset, which is communicating to the computer through Wi-Fi and only requires a saline solution for conduction. EEG is sampled at a rate of 128 Hz at a second and 14 channels could be measured using this device through TestBench. The headset was also communicating to the Affectiv suite (Gherguluscu 2014), which outputs five high-level features (short-term excitement, long-term excitement, meditation, frustration and boredom). Eye tracking was performed using Tobii Tx300 which is characterized with a high rating sampling frequency of 300Hz per second and its robustness to head movements and light variations. We extract from this sensor the pupil diameter in order to measure learner's workload (Bartels et al. 2012). Facial expression recognition was done using the FaceReader 6.0 software by recording the participant with a webcam. The face tracking process uses the popular Viola and Jones algorithm (Viola et al. 2001). This latter allows us to obtain a real time classification of seven basic emotions defined by Ekman (Ekman 1970): happy, sad, angry, surprised, scared, disgusted, and neutral with their valence and arousal (Lewinski et al. 2014).

The experiment process is composed of 7 steps: (1) installation of Emotiv EPOC headset, (2) calibration of Tobii eye tracker, (3) fill in a personality test (Big Five (Olivier et al. 1999)), (4) a pre-test that consist of constructing 3 molecules (CO_2, CCl_2F_2, C_2H_4), (5) exploring our 3D environment and learning Lewis diagram principles, (5) a post-test which is of a similar difficulty of the pre-test (this latter tests if the participant's understood all the instructions presented on the game and could apply them to solve other examples) and (7) finally, evaluation of our environment and self-reported difficulty of each presented mission of our game (easy, medium and hard ranging from 1 to 3) using a questionnaire based on a Likert's scale (see figure 5).

Figure 5. The experiment protocol

4.2 Data Pre-processing.

Given the data's sequential nature, the data stream was divided in individual sequences according to the learners' trials recorded by the game. For example, the learner could try and fail three times for the first task of the game, with a fourth successful trial. Four sequences would then be available for analysis. Each sequence was then reduced to a feature vector consisting of the 4 metrics, median, standard deviation, maximum, and minimum values for each feature gathered during the game session: short-term excitement, long-term excitement, meditation, frustration, boredom from the Affectiv suite from Emotiv (on a scale from 0 to 1), pupil diameter from the eye tracking sensors to measure learner's workload, arousal, valence and the seven emotions mentioned above from FaceReader (15 features at total). A total of 633 sequences (across 33 participants) 60-dimensional vectors (15 features multiplied by 4 metrics) were produced. 7 participants were ignored for analysis as technical errors (unrelated to the participants) corrupted data segments essential for a correct synchronisation of all data streams. So, only 33 participants out of 40 were taken into consideration for data analysis in the rest of this paper.

Other missing data values (e.g. the eye tracker failing to fit a model to the participant for a small amount of time) were replaced by the mean values for each feature.

Given the difficulty and the nature of the game, which encourages trials and errors, most sequences are labelled as failures. This presents a severely unbalanced dataset, and we used class weighting in order to address this issue. Class weighting inversely penalizes misclassifications according to the frequency of each class, and is implemented in Scikit-learn (Pedrogosa et al. 2011), the Python library we used to manipulate data and train machine learning models.

5. Results

In this part, we will first present some descriptive results gathered from our participants while interacting with our game,

LewiSpace. Second, we will show a suitable machine learning model to integrate with the appropriate features in real-time applications. Finally, we will discuss the importance of each sensor used for further learner's help adaptation.

5.1 Descriptive Results

As we mentioned before (experiment section), all the participants completed at the end a subjective evaluation questionnaire which contains questions about the difficulty of each mission of our game as well as the degree of game's appreciation (ranked from 1:not appreciated at all to 5: very appreciated). We noticed then that we obtained a mean of 3.083 (medium appreciated) and a standard deviation (SD) of 1.204. So, we can see clearly that a very big number of participants like the design and the content of our game. However, a small number didn't. This factor could be explained by the value of SD which is a little bit high. We also noticed in general that a small number of participants (20%: 8 from 40) succeed to complete the whole missions of the game because it's a difficult lesson for non-scientific people and needs more explications and examples (which is our main goal for an improved version of this game according to different types of learners and personalities).

Next, we calculated for each mission some statistics of the average number of failure (see figure 6 below). From this figure, we reported the mean and the standard deviation M(SD) for each mission. For example, mission 1 is ranked the second highest in term of average of failure although it's an easy mission (H_2O):10.04(9.49). Mission 2 has the lowest number of failure (6.15(5.75)) despite it is very similar to mission 1. Mission 3 has the highest number of failure: 14.55(8.55). However, mission 4 and mission 5 are about the same in term of mean of failure. Whereas, it should be noticed that is a large number of participants decided to quit in mission 3 (27 cases) and a small number of them completed mission 4 and more particularly mission 5 (8 cases) which is at the end of our game. Despite these results, we noticed that there are some learners have 0 failures in some missions and some ones have 29 errors or failures (the maximum number of trials which are wrong achieved for our sample of people).

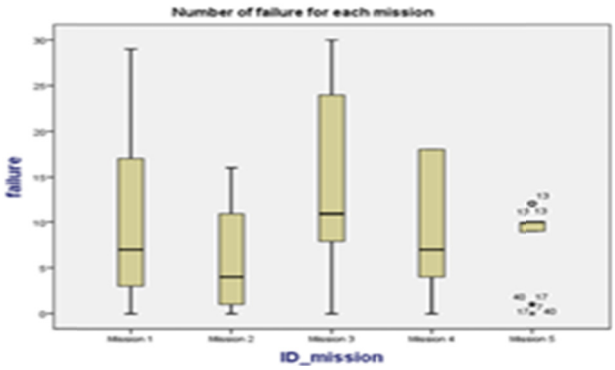

Figure 6. Statistics of failure per mission

Therefore, we calculated in the table below the means and standard errors of the average duration of success in seconds (s) for each mission to study again the difficulty of missions for our sample of students. We noticed from this table that mission 2 is the easiest one (433.80 s) and mission 3 is the hardest one (1433.26 s). However, missions 1 and 4 are about the same in term of difficulty according to the time spent in order to accomplish them. This result is very surprising for us as we assume that the mission 1 is very easy to accomplish.

Table 2. The average success time per mission

Missions	Total Duration
	Mean (Standard Error)
Mission 1	926.68 (69.83)
Mission 2	433.80 (24.37)
Mission 3	1433.26 (67.43)
Mission 4	902.03 (73.96)
Mission 5	626.72 (86.76)

Moreover, we realized an ANOVA to study if the missions are statistically different according to the number of failure and the duration. Results shows that missions depend only on the number of failure and not the duration ($F(4,666)=25.17;p=0.000^{**}<0.01$) but the results of three popular post hoc tests (Scheffe, LSD and Tukey) are not significant. This means that the missions are not totally different according to the failure factor. Finally, a paired samples t-test was conducted to compare the score improvement in the pre-test and post-test. There was a significant difference in the scores for the pre-test (M=12.28, SD=23.78) and the post-test (M=54.83, SD=29.42) conditions; $t(37)=-9.054$, $p = 0.000^{**}$. This result proves that our game contributes to improve learners' scores for drawing Lewis

structures after learning the lesson.

5.2 Selection of the Best Machine Learning Model

Support vector machine (SVM) with a Radial Basis Function (RBF) kernel and logistic regression models were tested with a grid search on values of gamma (for SVMs) and C to produce the highest balanced accuracy with a leave-one-participant-out scheme. This scheme was used in order to promote the selection of a model that can generalize well for a new participant from previous participants. Both algorithms performed similarly. For instance with a RBF SVM, the accuracy is about 54.9% with the hyper-parameters (C=1.0 and gamma=0.05) using all the features issued from 3 sensors. This value is very near to that of logistic regression (56.4%). In the following, we interest only on communicating results for the best logistic regression models, as the focus of this paper is not the comparison of machine learning algorithms. Logistic regression model in our case provides the highest accuracy in all cases as we will see in the next section (by taken into consideration all or some features).

5.3 Comparaison of the Importance of Each Sensor and the Big Five

After selecting the best ML model that allows us to detect if the user needs help or not, we focus on this section to study the features importance that contributes mainly to prediction. In what follows, we present the difference in accuracies in term of substracting each time one feature (sensor feature or Big Five questionnaire). Balanced accuracy is determined by the mean of correct classifications for each class while according both classes the same weight. Overall accuracy is the mean number of correct classifications with weighting for the number of samples (therefore giving more weight for the "failure" class), and the mean participant accuracy is the mean number of correct classifications per participant, ignoring whether or not a participant produced more or less samples in the dataset, similarly to the balanced accuracy.

Table 3. Feature selection through classification accuracies

	All features (3 sensors)	Ignores pupil diameter	Ignores Emotiv	Ignores facial expression recognition	Ignores Big Five Questionnaire
Balanced accuracy	0.564	0.564	0.501	0.564	0.584
Overall accuracy	0.603	0.603	0.256	0.603	0.564
Mean participant accuracy	0.593	0.593	0.312	0.593	0.549

Table 3 shows that ignoring the Emotiv has the highest impact on performance, whereas the other features do not seem to change the accuracies when ignored. Adding five features from the self-reported Big Five Questionnaire (the values along the five dimensions measured by the questionnaire before starting the game), we note that the balanced accuracy is highest, but at the cost of the overall accuracy, which means that the classifier predicts more often that a task will be successful but mispredicts more sequences in total. Ideally, the model should be balanced between those two measures of accuracy. A model that measures only the features from the Emotiv headset was therefore tested and produces the best results so far but still very similar to one which uses all features (using 3 sensors and ignoring the big five questionnaire), with a balanced accuracy of 0.570 , an overall accuracy of 0.635 and a mean participant accuracy of 0.609. However, this indicates that other features are not necessary and might even add noise to the dataset. Table 4 shows its confusion matrix for the logistic regression using only Emotiv Affectiv Suite (The most important feature), predicted values are shown vertically, and true values horizontally.

Table 4. Confusion matrix for a logistic regression model with Emotiv headset features

	Failure	Success	Total number
Failure	0.665	0.335	532
Success	0.525	0.475	101
Total number	407	226	633

From table 4, we can see clearly that failure is easiest to predict (with an accuracy of 66.5% which is higher that the random baseline of 50%), whereas, success is more difficult to predict (value of 47.5% which is less than 50%).

As to the relevance of the other features, we can speculate that brightness changes throughout the game (e.g. some scenes or actions producing various lighting effects) have a larger impact on the pupil diameter, a factor unaccounted during the experiment. We could improve this by controlling the brightness in real time. Facial expressions might also be more useful in a game which is more emotionally engaging which is not the goal of our game. LewiSpace focuses more on educational aspect and how to provide the adequate help to learners according the different situations encountered when playing it.

Finally, we present the Receiver Operating Characteristic curves known as ROC curves (figure 7) for each participant

compared with the random baseline using our best model, a logistic regression (C value of 0.1). It is a plot of the true positive rate against the false positive rate for the different possible cutpoints of a diagnostic test. Accuracy for these models was measured by the area under the ROC curve. An area of 1 represents a perfect test, whereas an area of 0.5 represents a worthless test. The figure below illustrates the width of the accuracies encountered due to individual differences between the participants. We noticed a lot of people with a good accuracy (under the curve) however a lot of cases are misclassified. This result suggests that we should more closely investigate individually trained models rather than generalized models.

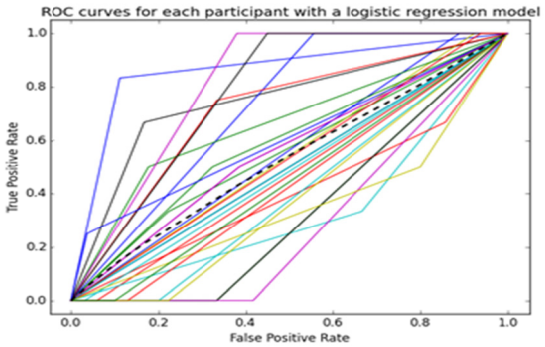

Figure 7. ROC curves for the 33 participants

6. Conclusions

In this paper, we presented our educational game aimed at teaching players how to draw Lewis diagrams known as LewiSpace. LewiSpace is a 3D environment integrating EEG and eyetracking SDKs and combining educational and playful aspects. Our study designed also to investigate the use of physiological data (electroencephalography, eye tracking, and facial expression recognition) and personality traits (the Big Five Questionnaire) in order to detect the performance of the learner and his need of help during progressing in our game. We described also some statistical results in term of failure and duration for each mission. From these results, we can clearly see that the game should be improved by providing more help and examples for students struggling to complete the tasks. This aim will be realized by collecting features from different types of physiological sensors and train a machine learning algorithm. Our finding shows that a logistic regression model using only Emotiv EPOC Affectiv Suite as features is the most suitable for detecting when learner have more difficulty (failure) and needs more help and examples to understand the lesson. We noticed also that personality traits as well as pupil diameter and emotions do not improve the accuracy of our model however they add more noise to our dataset due to the nature of our game.

Future work will involve developing a version of the game that reacts in real-time to the players' physiological data in order to help or challenge them accordingly. Before its development, we will however focus on models that will perform better on our current dataset and on the real-time data gathered in our next experiment.

Acknowledgements

We acknowledge the CRSH (Conseil de Recherche en Sciences Humaines, more precisely LEADS project) and NSERC for funding this work.

References

Azevedo, R., Johnson, A., Chauncey, A., & Burkett, C. (2010): Self-regulated learning with Meta-Tutor: Advancing the science of learning with MetaCognitive tools. In: New Science of Learning, pp. 225–247. http://link.springer.com/chapter/10.1007%2F978-1-4419-5716-0_11

Bartels, M., & Marshall, S. P. (2012). Measuring Cognitive Workload Across Different Eye Tracking Hardware Plateforms. ETRA 2012, Santa Barbara, CA. http://dl.acm.org/citation.cfm?id=2168582

Chaffar, S., & Frasson, C. (2006). Predicting Learners'Emotionl Response in Intelligent Distance learning Systems. The 19th International FLAIRS Conference, AAAI Press, Melbourne , FL , USA , May 15-17.

Conati, C. (2002). Probabilistic assessment of user's emotions in educational games. *Applied Artificial Intelligence 16*(7-8), 555-575. http://dx.doi.org/10.1080/08839510290030390

D'Mello, S., Olney, A., Williams, C., & Hays, P. (2012). Gaze tutor: A gaze-reactive intelligent tutoring system. International *Journal of Human-Computer Studies 70*(5), 377-398. http://dx.doi.org/10.1016/j.ijhcs.2012.01.004

Ekman, P. (1970). Universal facial expressions of emotion. *California Mental Health Re-search Digest, 8*, 151-158.https://www.paulekman.com/wp-content/uploads/2013/07/Universal-Facial-Expressions-of-Emotions1.pdf

Elliot, A. J., & Pekrun R. (2007), Emotion in the Hierarchical Model of Approach-Avoidance Achievement Motivation.EmotioninEducation, 57-74. http://dx.doi.org/10.1016/b978-012372545-5/50005-8

Ghali, R., Chaouachi, M., Derbali, L., & Frasson, C. (2014). Motivational Strategies to Support Engagement of Learners in Serious Games, The 6th International ICAART Conference, March 2014. http://dx.doi.org/10.1093/iwc/iwu013

Ghali, R., Ouellet, S., & Frasson, C. (2015). LewiSpace: An Educational Puzzle Game Combined with a Multimodal Machine Learning Environment. To appear in KI 2015: the 38th German Conference on Artificial Intelligence, Short paper.

Gherguluscu, I., & Muntean, C. H. (2014). A Novel Sensor-Based Methodology for Learner's Motivation Analysis in Game-Based Learning. *Interactive with Computers, 26*(4). http://iwc.oxfordjournals.org/content/early/2014/04/14/iwc.iwu013

Gobert, J. D., Baker, R. S., & Wixon, M. B. (2015). Operationalizing and Detecting Disengagement within Online Science Microworlds. *Educational Psychologist, 50*(1). http://dx.doi.org/10.1080/00461520.2014.999919

Jackson, G. T., Dempsey K. B., & McNamara D. S. (2012). Game based Practice in a Reading Strategy Tutoring System: Showdown in iSTART‐ME. Computer games, 115-138. http://dx.doi.org/10.1057/9781137005267.0013

Jaques, N., Conati, C., Harley, J., & Azevedo, R. (2014). Predicting Affect from Gaze Data During Interaction with an Intelligent Tutoring System. ITS conference 2014. http://www.cs.ubc.ca/~conati/my-papers/ITS-Natasha-2014.pdf.

Johnson, W. L., & Wu, S. (2008). Assessing Aptitude for Learning with a Serious Game for Foreign Language and Culture. Intelligent Tutoring Systems. B. Woolf, E. Aïmeur, R. Nkambou and S. Lajoie. *Springer Berlin Heidelberg, 5091*, 520-529. http://link.springer.com/chapter/10.1007%2F978-3-540-69132-7_55

Lester, J., Spires, H., Nietfeld, J., Minogue, J., Mott, W., & Lobene, E. (2014). Designing game-based learning environments for elementary science education: A narrative-centered learning perspective. *Information Sciences*, (264), 4-18. http://dx.doi.org/10.1016/j.ins.2013.09.005

Lewinski, P., den Uyl, T. M., & Butler, C. (2014). Automated facial coding: validation of basic emotions and facs AUs in FaceReader. Journal of Neuroscience. *Psychology, and Economics, 7*(4), 227-236. http://dx.doi.org/10.1037/npe0000028

McNamara, D. S., Jackson, G. T., & Graesser, A. (2010). Intelligent Tutoring and Games (ITaG). Ch3, Gaming for Classroom-Based Learning: Digital Role: Playing as a Motivator of Study. http://www.igi-global.com/chapter/intelligent-tutoring-games-itag/42686

McQuiggan, S., & Lester, J. (2006). Diagnosing Self-efficacy in Intelligent Tutoring Systems: An Empirical Study. Intelligent Tutoring Systems. M. Ikeda, K. Ashley and T.-W. Chan, Springer Berlin Heidelberg. *4053*, 565-574. http://link.springer.com/chapter/10.1007%2F11774303_56

Oliver, P. J., & Srivastava, S. (1999). The Big-Five Trait Taxonomy: History, Measurement, and Theoretical Perspectives, L., & Pervin et O. P. John (Eds.), Handbook of personality: Theory and research (2nd ed.). http://heckman.uchicago.edu/sites/heckman2013.uchicago.edu/files/uploads/OECD_Spencer_2015/John_Srivastava_1999_BIG%20FIVE%20TRAIT%20TAXONOMY.pdf

Pedregosa et al. (2011). Scikit-learn: Machine learning in Python, JMLR (12), 2825-2830. http://scikit-learn.org/stable/.

Prensky, M. (2001). Digital game based learning. New York: McGraw-Hill. http://www.amazon.fr/Digital-Game-Based-Learning-Marc-Prensky/dp/1557788634

Rowe, J., Mott, B., McQuiggan, S., Robison, J., Lee, S., & Lester J. (2009). Crystal island: A narrative-centered learning environment for eighth grade microbiology. Workshop on Intelligent Educational Games at the 14th International Conference on Artificial Intelligence in Education, Brighton, UK: 11-20. http://www.researchgate.net/publication/255576748_CRYSTAL_ISLAND_A_Narrative-Centered_Learning_Environment_for_Eighth_Grade_Microbiology

Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Kauai, HI, U.S.A., December 8-14. https://www.cs.cmu.edu/~efros/courses/LBMV07/Papers/viola-cvpr-01.pdf