Université de Montréal

Environnement d'adaptation pour un jeu sérieux

par Sébastien Ouellet

Département d'information et de recherche opérationnelle,
Faculté des Arts et des Sciences

Mémoire présenté en vue de l'obtention du grade de Maîtrise en informatique

Juin 2016

Résumé en français

Nous avons développé un jeu sérieux afin d'enseigner aux utilisateurs à dessiner des diagrammes de Lewis. Nous l'avons augmenté d'un environnement pouvant enregistrer des signaux électroencéphalographiques, les expressions faciales, et la pupille d'un utilisateur. Le but de ce travail est de vérifier si l'environnement peut permettre au jeu de s'adapter en temps réel à l'utilisateur grâce à une détection automatique du besoin d'aide de l'utilisateur ainsi que si l'utilisateur est davantage satisfait de son expérience avec l'adaptation. Les résultats démontrent que le système d'adaptation peut détecter le besoin d'aide grâce à deux modèles d'apprentissage machine entraînés différemment, l'un généralisé et l'autre personalisé, avec des performances respectives de 53.4% et 67.5% par rapport à un niveau de chance de 33.3%.

Résumé en anglais

We developed a serious game in order to teach users how to draw Lewis diagrams. We integrated an environment able to record electroencephalographic signals, facial expressions, and pupil diameters with the serious game. The goal of this work is to determine whether such an environment enabled the serious game to detect in real-time whether or not the user needs help and adapt itself accordingly, and if the experience is more enjoyable for the users if the game tries to adapt itself. Results show that two approaches were promising in order to detect the level of help needed, both training a machine learning models but one using a general data set and the other a personalized (to the user) data set, with their respective performances being 53.4% and 67.5% compared to a chance baseline of 33.3%.

Table des matières

Table des figures

Liste des tableaux

1 Introduction

1.1 Problématique

Avec la popularité grandissante des jeux vidéos dans des contextes autres que ludiques [1], le domaine de l'éducation s'intéresse aux possibilités technologiques offertes par les jeux vidéos éducatifs, où le terme jeux sérieux est en vogue afin de les désigner [2]. Le terme est également appliqué à des jeux ayant d'autres buts, tels que sensibiliser les joueurs à des problèmes sociaux, les entraîner à accomplir des tâches spécifiques (p.ex. programmes militaires), ou les informer sur des pratiques de santé [3]. Les jeux sérieux ont généralement tous en commun une emphase sur la résolution de problème grâce à des compétences acquises au cours du jeu, ce qui est semblable à plusieurs jeux vidéos ludiques à l'exception que les problèmes rencontrés dans un jeu sérieux sont généralement applicables à des situations qui ne sont pas virtuelles [4].

Le développement de plus en plus de jeux sérieux peut s'expliquer par les avantages attribués à l'apprentissage de compétences et concepts académiques combinés aux bénéfices apportés par l'utilisation de jeux vidéos typiques, tels que le développement cognitif des joueurs, leur bien-être psychologique, et la croissance de liens sociaux dans le cas de jeux multijoueurs [5]. D'intérêt particulier est la motivation générée par plusieurs jeux vidéos en comparaison aux méthodes d'instruction traditionnelles utilisées par les écoles.

Toutefois, bien que les jeux sérieux aient démontré qu'ils permettent aux

utilisateurs de mieux saisir et retenir des concepts académiques lorsque comparés à des cours conventionnels, il n'est pas encore clair qu'ils motivent davantage les utilisateurs [6], ce qui demeure un problème par rapport à ce qui est envisagé pour les jeux sérieux. L'un des avantages envisagés d'un jeu sérieux par rapport à une méthode traditionnelle est la possibilité que l'apprenant soit davantage motivé à poursuivre son apprentissage dans un contexte domestique (i.e. non pas à l'intérieur d'un cours). Si l'aspect ludique du jeu sérieux est suffisamment intéressant, l'apprenant peut décider de compléter un niveau du jeu comme si c'était n'importe quel autre jeu vidéo puisque l'expérience de jeu est agréable, sans se préoccuper que le défi rencontré communique du contenu éducatif.

1.2 Objectifs

Afin d'adresser ce problème de perception des jeux sérieux par les apprenants, nous avons amorcé le projet de recherche suivant : le développement d'un jeu enseignant un concept de chimie (la construction de diagrammes de Lewis) auquel est intégré un environnement de mesures psychophysiologiques apte à détecter le besoin d'aide de l'utilisateur, indiquant ainsi au jeu comment s'adapter au joueur. Nous détections le besoin d'aide de l'utilisateur grâce à des modèles d'apprentissage machine entraînés sur les mesures suivantes : l'électroencéphalographie, le suivi du regard, et les expressions faciales. Les modèles associaient les valeurs des différentes mesures à des niveaux de besoin d'aide indiqués par l'utilisateur lui-même. Cette méthode

produisait donc des modèles capables de lire les mesures d'un utilisateur puis de prédire son niveau de besoin d'aide.

La motivation est largement affectée par la perception d'un jeu éducationnel (i.e. s'il est agréable ou non), et un aspect majeur de la perception d'un apprenant par rapport à un tel jeu est la difficulté des tâches à accomplir [7]. Le jeu sérieux développé pour ce projet se voulait capable d'adapter ses tâches de sorte que l'utilisateur ne se trouvait pas dépassé (tâche trop difficile) ou ennuyé (tâche trop facile) lors d'une session. Afin d'ajuster la difficulté des tâches, le jeu sérieux adaptait la quantité d'information communiquée au joueur. L'information consistait d'indices reliés aux tâches, tels que des exemples de solutions similaires.

Ce projet comptait donc deux objectifs de recherche. Un premier objectif était d'identifier une méthode nous permettant de détecter le besoin d'aide d'un apprenant grâce aux mesures psychophysiologiques choisies : l'électroencéphalographie, le suivi du regard, et les expressions faciales. Le second objectif était de développer un jeu sérieux pouvant se servir de cette méthode afin de tester si la difficulté des tâches pouvait être adaptée à l'apprenant en temps réel.

2 Revue de littérature

2.1 Apprentissage machine

Afin de produire un modèle capable de prédire le besoin d'aide d'un apprenant, nous avons décidé d'utiliser des algorithmes d'apprentissage machine. L'apprentissage machine peut se résumer par la définition d'un modèle mathématique qui trouve lui-même ses paramètres grâce aux propriétés statistiques d'un ensemble de données [8]. Afin de produire un modèle, nous n'avons donc qu'à choisir un type d'algorithme adéquat aux données disponibles, des hyperparamètres décidant de comment ses paramètres sont ajustés, et une banque de données formées d'attributs (i.e. un nombre quelconque de valeurs discrètes ou continues associées à des caractéristiques du processus qui génère les données) et d'étiquettes (i.e. une valeur que le modèle prédit).

D'intérêt particulier au présent projet de recherche, nous avons concentré nos efforts sur des algorithmes répondant à certaines propriétés : rapides à exécuter (dû à la nature à temps réel du jeu sérieux) et robustes lorsqu'entraînés sur un petit nombre d'échantillons (dû au fait que nous devions amasser nous-mêmes les données à travers des sessions de jeu dans un temps limité). Les machines à vecteurs de support (SVM), les modèles de plus proches voisins (k-NN), et les modèles de "Random Forest" répondent adéquatement à ces conditions [8, 9].

Les SVMs trouvent un minimum global dans l'espace des paramètres maximisant une marge entre les échantillons appartenant à une catégorie

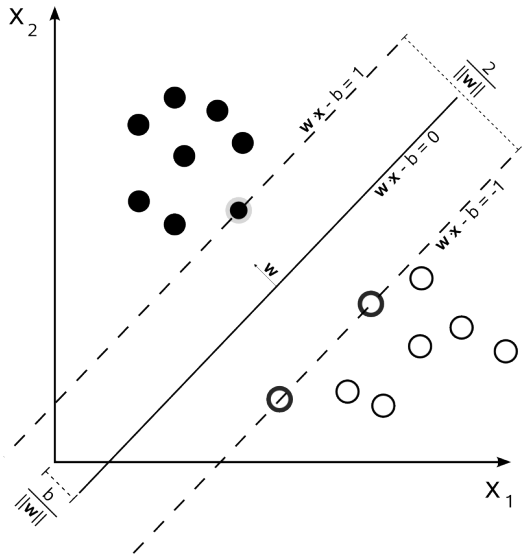

FIGURE 1 – Illustration de la définition de vecteurs de supports grâce à des échantillons se trouvant près de la frontière entre deux les catégories. La figure provient de Wikimédia, produite par Cyc.

par rapport aux échantillons d'une autre catégorie. Lors de la prédiction, le modèle de SVM va indiquer de quel côté de la marge l'échantillon se retrouve [10]. Les hyperparamètres décident des propriétés de la marge, si elle se déforme plus ou moins pour les échantillons ne se trouvant pas du bon côté, par exemple. La figure 1 illustre la maximisation de la marge définie par des vecteurs.

En terme d'applications, les modèles de régression logistiques sont souvent comparés aux SVMs car dans les deux cas, ces modèles furent conçus pour une classification binaire et linéaire [11]. La régression logistique a l'avantage, en terme d'interprétation, de retourner la probabilité qu'un échantillon appartienne à l'une ou l'autre des catégories étudiées.

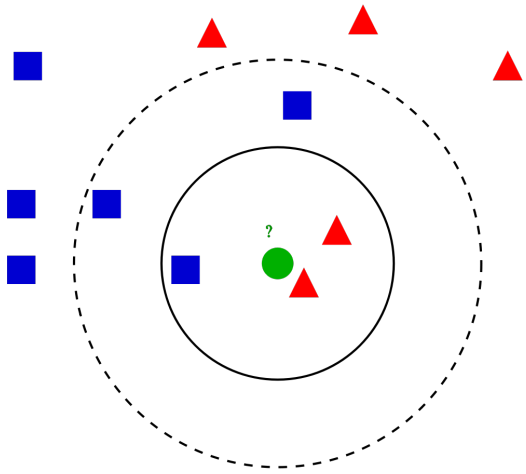

FIGURE 2 – Un échantillon de catégorie inconnue (cercle) et les deux catégories (carrés et triangles). La figure provient de Wikimédia, produite par Antti Ajanki.

Les modèles de plus proches voisins ne sont pas entraînés au préalables, mais lorsque que la prédiction de la catégorie d'un échantillon inconnu est nécessaire, le modèle trouve un nombre d'échantillons qui sont les plus près de l'échantillon inconnu (en terme de distance entre les attributs) et prend un vote de ceux-ci afin de décider de la catégorie. L'hyperparamètre particulièrement important ici est le nombre de voisins que le modèle choisit (allant de un seul à potentiellement des dizaines ou plus). L'importance du nombre de voisins est visible dans la figure 2, où un choix de deux voisins indique une classification vers la catégorie illustrée par des triangles, alors que choisir cinq voisins indique une classification vers la catégorie illustrée par des carrés.

Les modèles de "Random Forest" constituent un ensemble d'arbres de décisions, une technique développée afin de prévenir le surentraînage, un

phénomène où un modèle d'apprentissage machine peut très bien prédire les échantillons sur lesquels il fut entraîné, mais ne se généralise pas à de nouveaux échantillons [9]. Les hyperparamètres importants décident de la taille des arbres (nombre de décisions que chacun des arbres prend) ainsi que leur nombre dans l'ensemble.

2.2 Systèmes tutoriels intelligents

La littérature à propos des systèmes tutoriels intelligents est également pertinente au problème discuté ici, puisqu'on peut considérer un jeu sérieux comme un système tutoriel intelligent intégrant des mécaniques de jeu vidéo [12]. Les systèmes tutoriels intelligents se démarquent de logiciels éducatifs typiques par leur utilisation de méthodes d'intelligence artificielle [13]. L'intelligence artificielle est intégrée dans ces systèmes afin de promouvoir l'apprentissage de l'utilisateur, puisque le système tutoriel intelligent possède un modèle capable de changer le contenu ou la présentation d'une tâche éducative dépendemment du type d'utilisateur.

Un type d'intelligence artificielle utilisé dans les systèmes tutoriels intelligents est un modèle pouvant prédire le niveau de réussite d'une tâche. Ramla Ghali, et al. ont démontré la possibilité d'utiliser des signaux psychophysiologiques afin de prédire le score d'un participant lors de la résolution de problèmes mathématiques [14]. Les signaux étaient des indices calculés grâce à un casque EEG, une méthode qui nous a inspiré pour nos expériences. Un modèle de régression linéaire utilisant des informations reliées aux tâches

(temps d'exécution, complexité, type de tâche) et les indices EEG atteignait une corrélation R de 0.501, alors qu'un second modèle ignorant les informations reliées aux tâches atteignait un indice de corrélation R de 0.162 entre les indices produits par le casque EEG (Workload et Engagement) et les scores, et que les deux indices étaient négativement corrélés aux scores. La faible corrélation du second modèle nous a indiqué qu'un modèle plus sophistiqué qu'un modèle de régression linéaire était nécessaire lorsque la seule information disponible étaient les signaux psychophysiologiques.

2.3 Jeux sérieux

Les jeux sérieux se définissent comme des jeux vidéos combinant le divertissement d'un jeu typique à d'autres buts, tels qu'enseigner du contenu éducatif ou sensibiliser les joueurs à des situations particulières (p.ex. la simulation d'un individu vivant sous le seuil de la pauvreté) [4]. Les principaux développeurs de jeux sérieux sont généralement des scientifiques explorant des problèmes culturels, sociaux, ou pédagogiques, ou des organismes produisant des logiciels afin d'entraîner du personnel spécialisé, p.ex. des infirmières ou des soldats [15]. La plupart de ces jeux se présentent sous forme de simulations qui ont l'avantage de pouvoir être exécutées à répétition en changeant les situations pertinentes afin de combler les compétences du personnel, un exemple étant la possibilité de permettre à des pompiers de combattre virtuellement différents types d'incendie.

Les jeux sérieux présentent un avantage par rapport aux systèmes tu-

toriels intelligents, puisque le but principal des mécaniques de jeu dans ce contexte est d'accroître la motivation d'un joueur, ce qui peut être adressé par des mécaniques offrant un sentiment de succès (accomplir une tâche est suivi d'une action agréable dans l'univers du jeu), d'autonomie (plusieurs actions peuvent être accomplies de façon non-linéaire), et d'intérêt (considérant qu'il existe une raison dans l'univers du jeu afin d'accomplir des tâches) [12]. Ces aspects pourraient donc être utilisés par les éducateurs afin de produire des jeux sérieux plus motivants que les systèmes tutoriels intelligents.

Les jeux sérieux incluant des mesures psychophysiologiques ont fait l'objet d'une méta-étude récente [16], indiquant que ce type de données apportait de nombreuses nouvelles façons de mesurer l'ennui, l'attention, ou d'autres états mentaux associés à un session de jeu. La majorité des travaux utilisent ces mesures afin d'observer l'effet d'événements du jeu sur l'état du participant, plutôt que de s'en servir afin d'adapter l'expérience à l'utilisateur, sauf dans des cas où le contrôle du jeu (ou de l'environnement virtuel) est exécuté à travers ces mesures, permettant à l'utilisateur d'interagir avec l'environnement virtuel grâce à des états mentaux spécifiques.

Chanel et al. a démontré la détection d'anxiété et d'ennui basée sur des signaux physiologiques et d'électroencéphalographie dans un contexte de jeu vidéo ludique (Tetris) [17]. Toutefois, le jeu ne se modifiait pas par rapport à l'état détecté du joueur. Ils ont conduit une expérience avec 14 participants dans un environnement qui mesuraient les signaux suivants : EEG, la conductance cutanée, le pouls, le volume sanguin relié au pouls, l'expansion

du torse, et la température de la peau. Leur modèle de classification pouvait prédire trois classes (ennui, engagement, et anxiété) avec une précision de 63%.

Plus communément, les jeux sérieux ne s'adaptent pas ou s'adaptent non pas grâce à des capteurs physiques, mais avec des modèles se basant sur des mesures directement reliées au jeu et à l'interaction entre le jeu et l'utilisateur (p.ex. score, nombres de clics, déplacement dans l'environnement virtuel) [1, 18]. Un exemple est Crystal Island [19], un jeu où un apprenant se retrouve dans une communauté insulaire et doit interagir avec les personnages et les lieux afin de diagnostiquer une épidémie. Le jeu essaie de détecter l'état émotionnel d'un utilisateur afin de modifier comment les agents dans l'environnement virtuel communiquent avec l'utilisateur. Ce type de modèle est décrit en détails en [20], celui-ci pouvant fonctionner avec ou sans capteurs.

Ramla Ghali et al. ont publié une description du jeu sérieux présenté dans cette thèse et un modèle entraîné sur les signaux psychophysiologiques qui prédisait si un apprenant allait réussir la tâche en cours ou non [21]. Le modèle pouvait prédire correctement 57% des échantillons, une hausse de performance de 14% par rapport à un modèle basé sur le hasard (50%).

3 Expérimentation

Afin d'amasser des données, nous avons conduit deux expériences grâce à un environnement conçu spécifiquement à cette fin. L'environnement en question consistait en un jeu vidéo combiné à du matériel spécialisé accumulant des données au cours d'une session. La première expérience servait à obtenir les données nécessaires afin d'entraîner des modèles d'apprentissage machine pouvant prédire le besoin d'aide des apprenants. La seconde expérience implémentaient ces modèles afin de tester si le jeu pouvait adapter la difficulté des tâches en temps réel et accumulaient davantage de données, nous permettant de raffiner notre analyse par rapport à la détection d'un niveau de besoin d'aide.

3.1 Implémentation du jeu sérieux

3.1.1 Aspects conceptuels

Le contexte du jeu sérieux était une aventure où l'apprenant jouait le rôle d'un astronaute explorant une planète inhabitée. L'histoire débutait au moment où le personnage contrôlé par le joueur finissait une mission et retournait vers sa fusée afin de retourner en orbite. Le joueur était accompagné d'un autre personnage, Commandant Arnold, qui communiquait avec lui par radio. Lors de son déplacement, le joueur se retrouvait coincé dans une caverne où il devait alors surmonter des obstacles afin de retrouver sa fusée, ce qui marquait la fin du jeu. Surmonter les obstacles consistait essentiellement

à combiner les objets trouvés sur le chemin en des molécules spécifiques. Le but ludique du jeu était de permettre à un joueur d'explorer un environnement et de progresser au-delà de défis mis sur son chemin, et le but éducatif était d'enseigner au joueur à construire correctement les molécules nécessaires à son progrès.

Le jeu en entier était composé d'un grand environnement à explorer, avec de nombreux endroits inaccessibles au début. Au cours du jeu, le joueur accumulait des éléments chimiques et assemblait des molécules lui permettant d'accéder à de nouveaux endroits et d'obtenir davantage d'éléments. Le jeu était donc une combinaison de jeux d'aventure et de réflexion.

L'exploration est un aspect agréable de nombreux jeux vidéos, motivant aisément les joueurs à progresser vu que plusieurs sont curieux de découvrir de nouveaux endroits et ce qu'ils contiennent. L'exploration était accomplie par le mouvement fluide en trois dimensions et à la première personne du personnage, avec la possibilité de sauter, pour permettre différentes altitudes dans un environnement.

La collecte d'objets est également populaire, et permet d'intégrer une interaction supplémentaire offerte au joueur. Dispersés dans l'environnement, le joueur peut trouver des éléments (représentés par des atomes, tels que des atomes d'oxygène, de carbone, d'hydrogène, etc.) et les amasser dans son inventaire. La collecte de tels éléments est effectuée en cliquant sur des emplacements spécifiques dans l'espace de jeu qui contiennent l'élément désiré, tel qu'illustré par la figure 3. On y voit deux sources de carbone attachées au

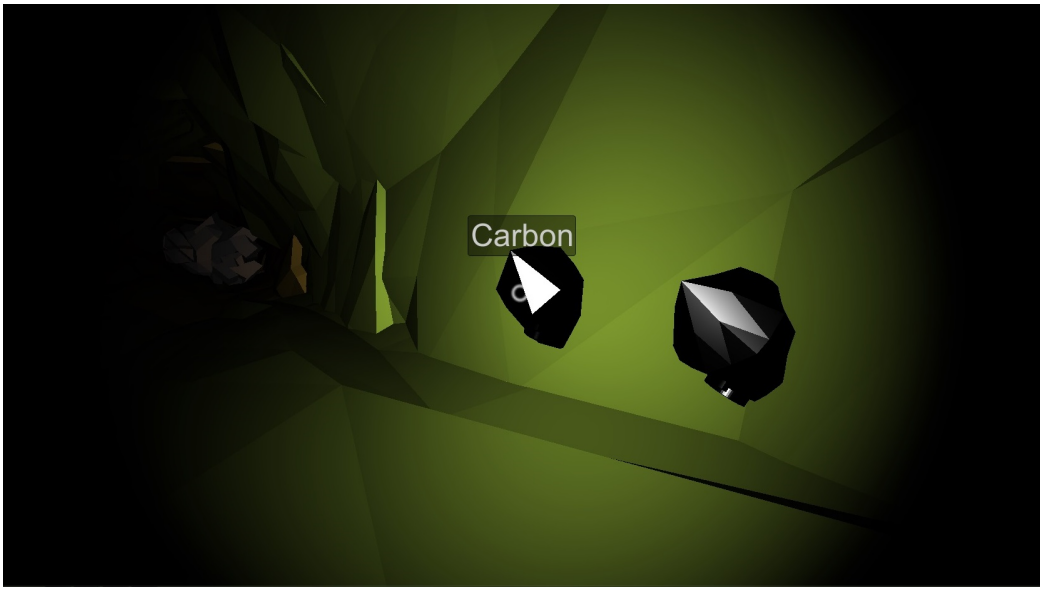

FIGURE 3 – Une capture d'écran illustrant la collecte d'un élément.

mur de la caverne, sur lesquels le joueur pouvait cliquer afin de les déposer dans son inventaire.

Les éléments amassés pouvaient être ensuite assemblés afin de produire des molécules de plus en plus complexes qui permettaient au joueur de continuer à progresser. Par exemple, le joueur avait besoin d'un élément qui était hors de portée et devait composer du méthane (avec le carbone et l'hydrogène déjà disponible) afin d'utiliser sa torche qui lui ouvrait un passage dans un mur. Le joueur devait donc trouver le carbone s'il ne l'avait pas déjà à ce point-ci, et les assembler de sorte à ce qu'il puisse activer sa torche, tel qu'illustré dans la figure 4. Ce principe était la mécanique centrale du jeu et était appliquée tout au cours de la session.

L'assemblage des éléments devait suivre la structure des molécules dési-

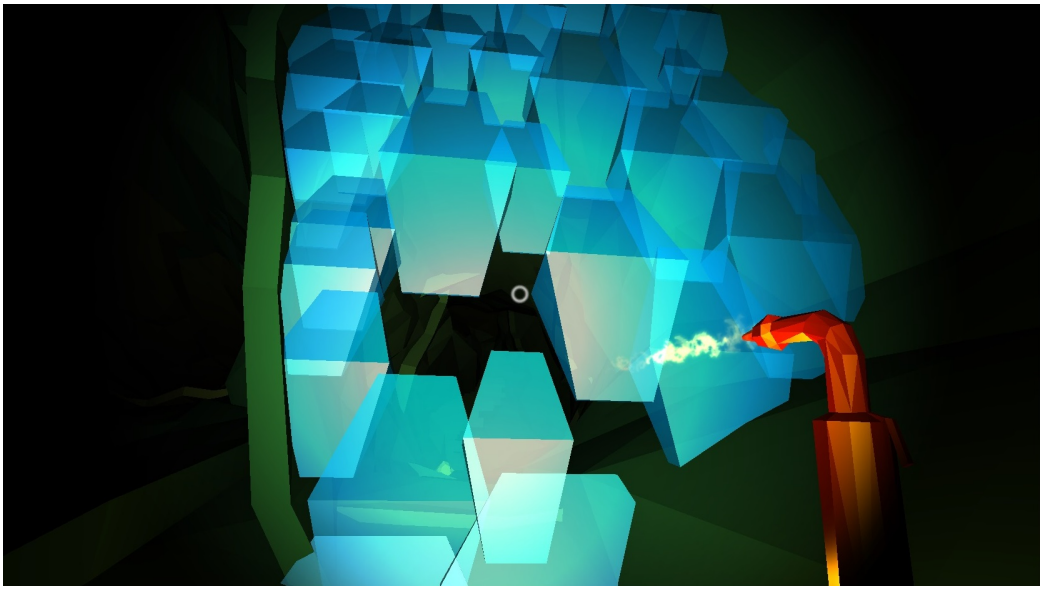

FIGURE 4 – Une capture d'écran illustrant un événement produit par une solution correcte.

rées, représentée par un diagramme de Lewis [22]. Ce type de diagramme est une représentation en deux dimensions d'une molécule, demandant au joueur de disposer des éléments chimiques selon une forme appropriée et de les relier dépendamment des électrons de valence disponibles. Le joueur était poussé à expérimenter avec les éléments à sa portée dans le but d'apprendre la composition de plusieurs molécules et les principes qui ordonnent les liaisons possibles entre les atomes. Il était également nécessaire d'assembler et de réutiliser les mêmes molécules plusieurs fois, ce qui motivait le joueur à ne pas oublier les structures déjà découvertes. Afin de composer des molécules, le joueur disposait d'un outil lui permettant d'accéder à un inventaire des éléments qu'il avait amassés, illustré par la Figure 5. Nous avons conçu la

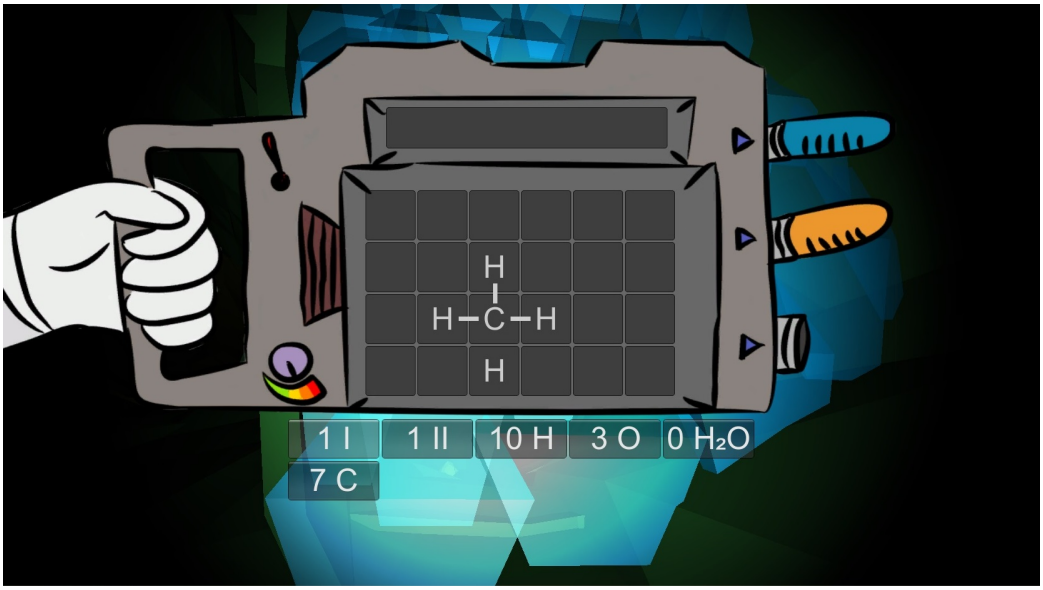

FIGURE 5 – Une capture d'écran montrant l'outil utilisé par le joueur.

méthode d'inventaire et d'assemblage de sorte qu'elle soit similaire à celle utilisée par de nombreux jeux populaires, p.ex. Minecraft [23].

Nous avons choisi la formation de diagrammes de Lewis comme contenu éducatif pour son originalité, car il est difficile de trouver des jeux vidéos enseignant ce concept, pour sa similitude vis-à-vis l'interface de jeux vidéos ludiques populaires [23] et pour le raisonnement logique requis lors des tâches. Également, les éducateurs se questionnent activement sur les méthodes d'enseignement de ce concept [24].

3.1.2 Étapes et tâches rencontrées par le joueur

Le joueur devait accomplir cinq tâches différentes afin de compléter le jeu. Chacune d'elle demandait au joueur de dessiner le diagramme de Lewis cor-

FIGURE 6 – Une capture d'écran où le personnage non-joueur communique avec le joueur.

respondant au composé communiqué au joueur. L'information était donnée textuellement, et le joueur avait accès à un historique du texte qui lui était fourni au cours du jeu.

Le joueur était introduit comme un astronaute en cours de mission, parcourant la surface d'une planète. Un personnage non-joueur, Commandant Arnold, se trouvait en orbite et communiquait fréquemment avec le joueur afin de clarifier les situations dans lesquelles il se trouvait. C'est lui qui indiquait quel composé chimique est nécessaire au progrès, tel qu'illustré par la figure 6.

Au cours des tâches, le jeu présentait le contenu éducatif relié à la formation de diagrammes de Lewis. Plutôt que décrire tous les principes au départ,

des règles furent introduites individuellement, suivant la progression d'un jeu vidéo de réflexion où les problèmes demandent de combiner de plus en plus de règles. Ainsi, les premiers obstacles nécessitaient moins de connaissances que ceux qui suivaient, similairement à un jeu vidéo ludique où le joueur peut se débrouiller dans les premiers niveaux, mais doit apprendre de nouvelles techniques afin de continuer. Les tâches rencontrées par le joueur sont donc plus faciles à aborder au départ, puis grandissent en complexité.

Pré-tâches pour le calibrage - expérience numéro 2 Nous avons ajouté cette section lors de la seconde expérience, afin de calibrer un modèle d'apprentissage spécifiquement pour l'apprenant actuel. Elles se passent alors que le joueur marchait à la surface de la planète, en vue de sa fusée, et qu'il avait accès à plusieurs éléments. La première pré-tâche indiquait au joueur de former une molécule composée de 2 éléments, un défi particulièrement simple afin de produire un sentiment de contrôle chez le joueur (i.e. le joueur devrait se sentir à l'aise avec la difficulté du jeu). La seconde lui demandait ensuite de former une molécule considérablement plus compliquée afin de produire un sentiment opposé à celui de la première pré-tâche.

Tâche 1 L'astronaute explorait à pied le terrain près de sa fusée. En marchant, il tombait dans une crevasse et se retrouvait dans une caverne. Arnold l'informait alors qu'il doit d'abord assurer sa survie avant d'explorer la caverne afin de remonter à la surface. Des bonbonnes d'oxygène et d'hydrogène pouvaient être ramassées, puisqu'elles étaient tombées avec lui de la surface,

afin de produire de l'eau, utile au système de réfrigération de sa combinaison (formule : H_2O). Le joueur était ici introduit au concept de liaison dans le cas de l'hydrogène. Le jeu l'informait de la règle d'octet (i.e. généralement, qu'un atome sera entouré de 8 électrons dans une molécule, et 2 électrons forment une liaison, illustrée par la figure 8), et que l'hydrogène est une exception puisqu'il ne peut supporter qu'une seule liaison avec un autre élément.

FIGURE 7 – Solution pour la tâche 1

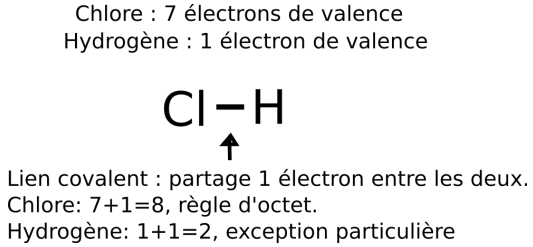

FIGURE 8 – Illustration de la règle d'octet

Tâche 2 Le chemin était bloqué par une formation cristalline, mais le Commandant Arnold informait le joueur qu'après analyse, le matériel pouvait fondre avec la torche qu'il possédait. Le joueur devait récolter du carbone, qui se trouvait naturellement dans la caverne, afin de produire du carburant pour sa torche qui fonctionne au méthane (formule : CH_4). À ce point-ci, le jeu informait le joueur qu'il est possible de consulter un tableau périodique : en appuyant une touche spécifique du clavier, un tableau apparaissait dans

l'écran de jeu. Le jeu fournissait également des instructions afin de trouver le nombre d'électrons de valence des éléments désirés. Le joueur apprenait également l'existence de liaisons simples (formées par une paire d'électrons) et doubles (formées par deux paires), ce qui lui permettait, avec la règle d'octet, de dessiner des diagrammes de Lewis en tenant compte des nombres d'électrons.

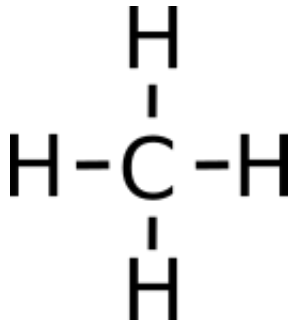

FIGURE 9 – Solution pour la tâche 2

Tâche 3 Des débris métalliques empêchaient l'astronaute d'avancer. Arnold l'informait que produire de l'acide sulfurique avec le sulfure se trouvant dans la caverne lui permettrait de dissoudre les débris (formule : H_2SO_4). Une règle supplémentaire était indiquée au joueur, l'informant que les éléments de la troisième rangée du tableau peuvent parfois dépasser la contrainte de la règle d'octet et supporter plus d'électrons de valence (illustrée par la figure 11).

Tâche 4 La température d'une zone de la caverne était particulièrement élevée, et la combinaison spatiale du joueur avait besoin d'un meilleur ré-

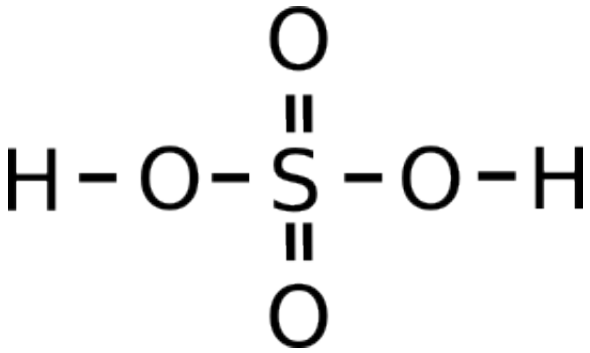

FIGURE 10 – Solution pour la tâche 3

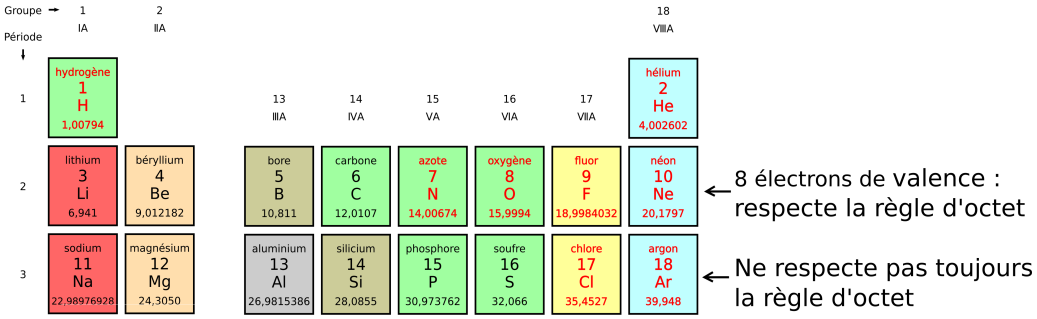

FIGURE 11 – Illustration de la troisième rangée

frigérant afin de pouvoir la traverser. Arnold lui suggérait de produire un type de fréon (CTFE ou chlorotrifluoroéthylène), avec du chlore et du fluor. afin de l'utiliser dans son système de réfrigération (formule : C_2F_3Cl). À ce point-ci, aucune nouvelle règle était introduite car le joueur possédait les connaissances nécessaires pour résoudre cette tâche. La difficulté de la tâche venait des nouveaux éléments nécessaires et de la structure asymétrique du diagramme de Lewis requis.

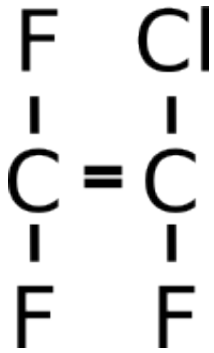

FIGURE 12 – Solution pour la tâche 4

Tâche 5 Le joueur retrouvait sa fusée et doit alimenter sa pile à combustible avec de l'éthanol afin de décoller et de retourner en orbite (formule : CH_3CH_2OH). Cette tâche présentait le composé demandant le plus grand nombre d'éléments. Lorsque l'éthanol était utilisée par le joueur, le joueur perdait le contrôle du personnage, puis la caméra se déplaçait et montrait une animation du décollage de la fusée. Cet événement marquait la fin du jeu.

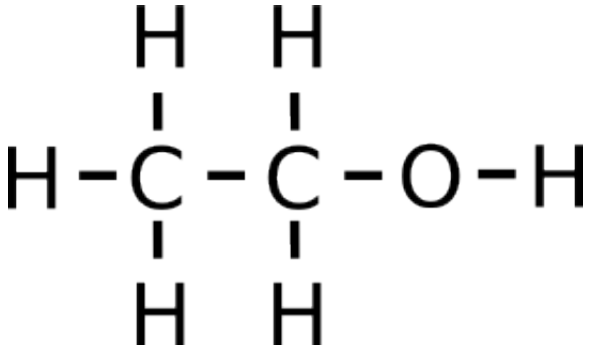

FIGURE 13 – Solution pour la tâche 5

3.1.3 Aspects techniques

Nous avons conçu le jeu sérieux grâce au logiciel Unity®, version 4.5, un moteur et éditeur de jeux vidéos offert par Unity Technologies [1]. Les ressources audiovisuelles furent développées et éditées grâce à Blender, version 2.6 [25], et Audacity®, version 2 [2]. Nous avons trouvé des fichiers audio sur http ://freesound.org, et la liste d'attribution se trouve, ainsi que le code source et les ressources audiovisuelles, dans un dépôt libre d'accès à l'adresse suivante : https ://github.com/Zebreu/louisjvi. Nous avons utilisé le langage de programmation C# afin de faciliter l'intégration de matériel permettant le suivi de regard ainsi que celui d'électroencéphalographie (EEG).

3.2 Environnement d'adaptation

Nous avons intégré au jeu un ensemble de technologies permettant l'acquisition et l'analyse de données en temps réel. Trois types de données pouvaient être récoltées : les expressions faciales, grâce à une caméra installée sur le moniteur de l'ordinateur utilisé, un électroencéphalogramme, grâce à un casque, et le regard ainsi que les dimensions des pupilles grâce à un ensemble de caméras infrarouges spécialisées pour le suivi du regard.

Les expressions faciales étaient reconnues automatiquement par FaceReader, offert par Noldus [26]. Les catégories utilisées lors de cette étude furent les suivantes : triste, joyeux, fâché, surpris, effrayé, dégoûté, et neutre, définies

1. http ://unity3d.com/
2. http ://audacityteam.org/

par Ekman selon des groupes d'unités d'actions du visage [27]. FaceReader fonctionne selon un algorithme de Active Appearance Model [28] afin de calculer des attributs pouvant être utilisés par un classifieur afin d'identifier, grâce à un entraînement supervisé, le type d'émotion illustrée par le visage du participant. Également, deux autres catégories étaient offertes par FaceReader et furent prises en compte : Valence et Arousal, i.e. le degré de positivité ou de négativité et le niveau d'excitation.

Nous avons utilisé un casque EPOC fabriqué par Emotiv comme outil d'électroencéphalographie [29]. Le casque est vendu à la fois pour des buts ludiques et de recherche, permettant aux utilisateurs de contrôler des jeux vidéos et d'enregistrer des états mentaux associés à l'ennui, l'excitation (à long et court terme), la méditation et la frustration (mesurés grâce à un indice variant de 0 à 1). Nous avons utilisé le SDK développé par Emotiv afin d'intégrer au jeu la capacité d'enregistrer et d'amasser un historique de ces indices [30]. Des évaluations du casque le comparant avec les résultats offerts par de l'équipement dispendieux [31] ainsi qu'avec des méthodes évaluant les états mentaux [32] ont déterminé que l'équipement d'Emotiv était suffisamment fiable bien que la qualité des enregistrements étaient inférieurs (i.e. les signaux étaient semblables mais le casque d'Emotiv montrait davantage de bruit).

Le Tobii TX300 offert par Tobii est un système de caméras infrarouges détectant les yeux d'un utilisateur et calculant où le regard se porte sur une surface (généralement l'écran d'un ordinateur) [33]. Considérant que les

tâches du jeu sont d'un type nécessitant du raisonnement logique plutôt que de l'attention visuelle, nous avons seulement pris en compte le diamètre de la pupille du participant dans l'environnement d'adaptation et non pas où le regard du participant se trouvait. Le diamètre de la pupille est considéré comme un indicateur de l'effort cognitif exercé par un utilisateur et de la difficulté d'un problème à résoudre [34].

Un programme auxiliaire au jeu lisait ces trois sources de données en temps réel. Ce programme évaluait si le participant trouvait la tâche trop difficile (les méthodes étant détaillées dans la section suivante) et indiquait au jeu d'offrir de l'aide. L'aide supplémentaire se trouvait majoritairement sous forme textuelle, informant le participant d'astuces afin de dessiner des diagrammes de Lewis, d'indices portant sur la tâche actuelle à accomplir, et d'exemples graphiques décrivant pas à pas la construction d'un diagramme, sans résoudre les diagrammes spécifiquement requis par le jeu afin de ne pas offrir une solution même si le participant est incapable d'avancer. Des exemples de cette aide pouvait être les phrases suivantes : " Plusieurs diagrammes sont symétriques " et " La formule moléculaire parfois indique la disposition de ses éléments ". Nous permettions donc à un joueur d'échouer et d'arrêter la session de jeu sans compléter toutes les tâches.

3.3 Méthodologie

Nous avons conduit deux expériences. La première s'est effectuée sans adaptation et a permis de recueillir des données qui furent ensuites analysées

afin de préparer la seconde expérience. Nous avons effectué la majorité de l'analyse en Python grâce à Scikit-learn, une bibliothèque logicielle d'apprentissage machine [35]. Deux expérimentateurs prirent part aux expériences, Ramla Ghali et Sébastien Ouellet (moi-même).

3.3.1 Procédure des sessions commune aux deux expériences

Nous avons conduit chacune des expériences à l'Université de Montréal avec 20 étudiants de l'université qui se sont portés volontaires (donc 40 au total, pour les deux expériences). Nous avons annoncé les expériences grâce à des affiches placées sur le campus. Une session typique pouvait avoir une durée de 60 à 90 minutes. Les conditions que les participants devaient remplir étaient les suivantes :

- Ne pas être apte à produire des diagrammes de Lewis.
- Être familier avec des jeux vidéos.
- Être capable de lire l'anglais.

Nous avons offert une compensation de 15$ pour la première expérience, et de 20$ pour la seconde (la différence reflétant le budget alloué à l'étude plutôt qu'une différence entre les deux expériences).

Les sessions débutaient par la signature d'un formulaire de consentement éthique, puis par la description de l'équipement utilisé, informant à la fois le participant de ce qui était enregistré au cours du jeu. Le jeu était brièvement introduit, rassurant le participant que des instructions lui seront données au cours du jeu. Le participant complète ensuite un test de personnalité basé sur

le modèle des Big Five [36] disponible librement sur le web [37]. Le casque EPOC de Emotiv est ensuite installé sur la tête du participant, vérifiant à la fois si la qualité du signal reçu par les électrodes grâce à l'interface Emotiv Control Panel offerte par la même compagnie. Nous démarrions ensuite l'enregistrement du vidéo pour les expressions faciales. Nous avons enregistré le vidéo pour la première expérience sans effectuer d'analyse immédiate, alors que pour la seconde expérience, nous avons utilisé le SDK pour le logiciel FaceReader de Noldus afin de démarrer à la fois l'enregistrement du vidéo, l'analyse en ligne de FaceReader, et l'écriture d'un fichier d'historique des résultats de l'analyse.

Nous partions ensuite le jeu, débutant par le calibrage du Tobii TX300 afin d'adapter le modèle de suivi du regard au participant actuel. Le SDK fourni par Tobii fut utilisé afin d'intégrer le calibrage dans la première scène du jeu. Le jeu présentait ensuite de courtes explications sur l'interface graphique et démarrait un pré-test consistant en 3 diagrammes de Lewis à dessiner afin d'obtenir un niveau de compétence de départ. Lorsque complété, le participant pouvait à partir de ce point progresser dans le jeu à son propre rythme, pouvant quitter le jeu à tout moment s'il est à court de temps ou incapable d'accomplir une tâche. Lors du jeu, l'expérimentateur ne fournissait pas de contenu éducatif si le participant posait une question, mais pouvait répondre à toute autre question (p.ex. un problème survenant avec le clavier). Lorsque la session de jeu était complétée (que ce soit par retrait volontaire ou complétion de toutes les tâches), nous présentions un post-test de difficulté

semblable au pré-test, évaluant le niveau de compétence actuel du participant. À ce point-ci, toutes les mesures d'enregistrement étaient arrêtées. La dernière étape était la complétion d'un questionnaire posant les questions suivantes au participant (reproduites telles que données aux participants). Dans le questionnaire, la notion de segment était définie comme la tâche à accomplir, ou plus précisément, une période de temps débutant à la fin d'une tâche allant jusqu'à la complétion de la tâche suivante.

- Pour les 5 parties suivantes, indiquez-nous si vous désiriez de l'aide supplémentaire ou un défi supplémentaire lors d'un segment de jeu. Écrivez un chiffre de 1 à 3, 1 signifiant un segment trop facile ou ennuyant pour vous, 2 un segment convenable, et 3 un segment trop difficile.
  - Premier segment, avec l'eau (H2O)
  - Second segment, avec le mur de glace et le méthane (CH4)
  - Troisième segment, avec les débris métalliques et l'acide (H2SO4)
  - Quatrième segment, avec le réfrigérant juste avant d'arriver à la surface (C2F3Cl)
  - Dernier segment, avec l'éthanol sur la surface de la planète (C2H6O)
- En général, à quel point avez-vous apprécié jouer ? Écrivez un chiffre entre 1 et 5, 1 signifiant un jeu que vous n'avez pas aimé du tout, et 5 un jeu que vous avez beaucoup apprécié.

Lorsque complété, l'expérimentateur remettait la compensation au participant et était libre de répondre à toutes ses questions (p.ex. la discussion

d'une solution au jeu).

3.3.2 Méthode d'analyse et seconde expérience

Entre les deux expériences, nous avons analysé les données amassées par la première expérience afin de produire un modèle capable d'adapter par la suite le jeu par rapport au besoin de chacun des participants.

Les données consistaient en 15 signaux différents : le diamètre de la pupille en millimètres, les cinq indices du Affectiv Suite (excitation à court terme, excitation à long terme, méditation, ennui, et frustration) entre 0 et 1, et neuf indices d'expressions faciales de FaceReader (neutre, triste, joyeux, fâché, surpris, effrayé, dégoûté, arousal, et valence) entre 0 et 1. Nous avons rassemblé les signaux en séquences délimitées par les essais pour chacune des tâches. Par exemple, un participant peut avoir essayé quatre solutions potentielles pour la première tâche avant de réussir à trouver une cinquième qui est, elle, correcte. Ce participant aurait alors cinq séquences de données associées à la première tâche. Ces séquences furent réduites à des attributs calculés par des mesures statistiques : la moyenne, l'écart-type, le minimum, et le maximum. Ainsi, chacune des séquence est représentée par un vecteur de 60 attributs (4 mesures statistiques pour chacun des 15 signaux). Les signaux furent également normalisés selon une échelle allant de -1 à 1. Dans le cas de données manquantes (p.ex. le suivi du regard n'a pas détecté de yeux pendant quelques secondes, ou une électrode du casque a perdu contact), nous avons substitué la valeur manquante par la moyenne du signal. Cette

substitution fut nécessaire pour 9.6% des données analysées.

Nous avons ensuite entraîné des algorithmes d'apprentissage machine supervisés sur la banque de données, utilisant comme étiquette les réponses aux questionnaires indiquant le niveau d'aide requis par les participants dépendamment des tâches. Dans un contexte d'apprentissage machine, une étiquette se définie comme une valeur représentant la catégorie d'une observation. Nous avons donc associé les séquences des signaux psychophysiologiques à des étiquettes qu'on pouvait ensuite prédire en se basant seulement sur les observations. Pour la première expérience, le nombre de séquences auxquelles les participants ont attribué un niveau de difficulté *Trop facile* était trop bas et cette catégorie de réponses fut combinée à *Adéquate*, obtenant ainsi une banque de données plus équilibrée. Nous avons sélectionné les modèles de classifieur par validation croisée (les stratégies étant spécifiés dans la section suivante) et en cherchant des hyperparamètres par recherche exhaustive en grille (dans le cas de plusieurs hyperparamètres, où par exemple γ et C sont modifiés à tour de rôle pour un modèle de machine à vecteurs de support). Ce type de sélection s'effectue en créant un modèle avec des hyperparamètres spécifiques puis en l'entraînant sur plusieurs sous-ensembles de données afin de le tester sur un sous-ensemble mis à part. Ceci est répété plusieurs fois en changeant les hyperparamètres afin de trouver le modèle le plus performant.

Deux approches (discutées plus en détails dans la section suivante) émergèrent pour le système d'adaptation adopté pour la seconde expérience : un modèle général, entraîné sur la banque de données amassés lors de la pre-

mière expérience, et un modèle personnalisé, qui est produit au début du jeu sur le participant actuel. Nous avons accompli l'entraînement du modèle personnalisé grâce à deux nouvelles pré-tâches se trouvant au début du jeu (avant la première tâche avec l'eau), les deux demandant encore de dessiner des diagrammes de Lewis. La seconde molécule demandée était toutefois considérablement plus complexe que la première, offrant donc au classifieur des séquences avec des étiquettes pour des tâches faciles et difficiles. Ces tâches étaient particulières dans le sens où une erreur au niveau du diagramme n'empêchait pas le joueur d'avancer puisqu'à ce point dans le jeu, le participant ne connaissait pas toutes les règles requises pour la construction de diagrammes. Une fois le modèle personnalisé entraîné (un court moment suffisant seulement, ne ralentissant pas le participant), le jeu reprend tel que dans la première expérience et chacun des modèles décide individuellement à toutes les deux secondes si les données reçues au cours du jeu suggèrent que le participant a besoin d'aide ou non, puis le jeu offre de l'aide supplémentaire si dans un intervalle de temps donné (40 secondes), plus de la moitié des décisions faites par les modèles suggèrent que le participant nécessite davantage d'aide.

4 Résultats et discussion

Les résultats sont présentés chronologiquement, où nous démontrons l'analyse que nous avons exécutée entre les deux expériences et le processus de notre préparation pour la seconde expérience compte tenu des résultats de la première. Particulièrement d'intérêt, nous mesurions la précision de classification des mesures psychophysiologiques afin de détecter le besoin d'aide des participants et leur niveau de réussite pour une tâche en cours.

4.1 Première expérience

Tel que décrit précédemment, nous avons découpé et réduit les données en des vecteurs d'attributs, produisant au total une banque de données de 158 séquences de 60 dimensions. Le nombre d'échantillons aurait pu être plus élevé, mais dû à des erreurs (oubli de démarrage d'enregistrement, problèmes techniques du jeu, difficulté d'ajustement du casque, erreur de calibrage du suivi du regard, grand nombre de données manquantes pour certains participants), nous n'avons pas pris en compte les données de plusieurs participants lors de la première analyse, et nous avons calculés les résultats indiqués ci-dessous à partir des données de neuf participants. Les participants étaient rejetés lorsque les sources de données (les mesures reliées à la pupille ou l'électroencéphalographie) n'étaient pas capturées de façon permanente, nous empêchant de les analyser.

4.1.1 Prédiction de la réussite

D'abord, nous avons entraîné un modèle de régression logistique sur une étiquette ne venant pas des questionnaires, mais de la rectitude de la solution offerte lors de la séquence. Ainsi, un modèle binaire suffit, mais considérant que la grande majorité des séquences présentent une étiquette négative, nous avons utilisé l'algorithme SMOTE (Synthetic Minority Over-sampling Technique) afin de permettre au classifieur de ne pas être biaisé envers la classe négative considérant qu'elle est largement majoritaire [38]. Nous avons choisi l'algorithme de régression logistique lors de cette première analyse car des tests préliminaires favorisaient ce type de modèles par rapport aux modèles de type SVM, "Random Forest", et k-NN. Nous avons utilisé une stratégie de "leave-one-participant-out" pour sélectionner un modèle, où le modèle est entraîné sur tous les participants sauf un, puis testé sur celui mis de côté. Ce type de validation croisée est souvent utilisé sur des études où on veut obtenir un modèle performant bien pour de différents participants, mettant une emphase sur sa généralisation [39]. La figure 14 présente les résultats sous forme d'aires sous la courbe ROC (Receiver Operating Characteristic) pour chacun des participants, avec une moyenne de 0.66, présentant une amélioration de 16 points au-dessus du niveau de chance de 50%. Le pourcentage de prédictions correctes de ce modèle est de 65%. Nous n'avons pas poursuivi cette avenue (la prédiction de la réussite) lors des analyses suivantes vu le faible taux de prédictions correctes, et nous nous sommes plutôt concentrés sur la prédiction du besoin d'aide désiré par les participants.

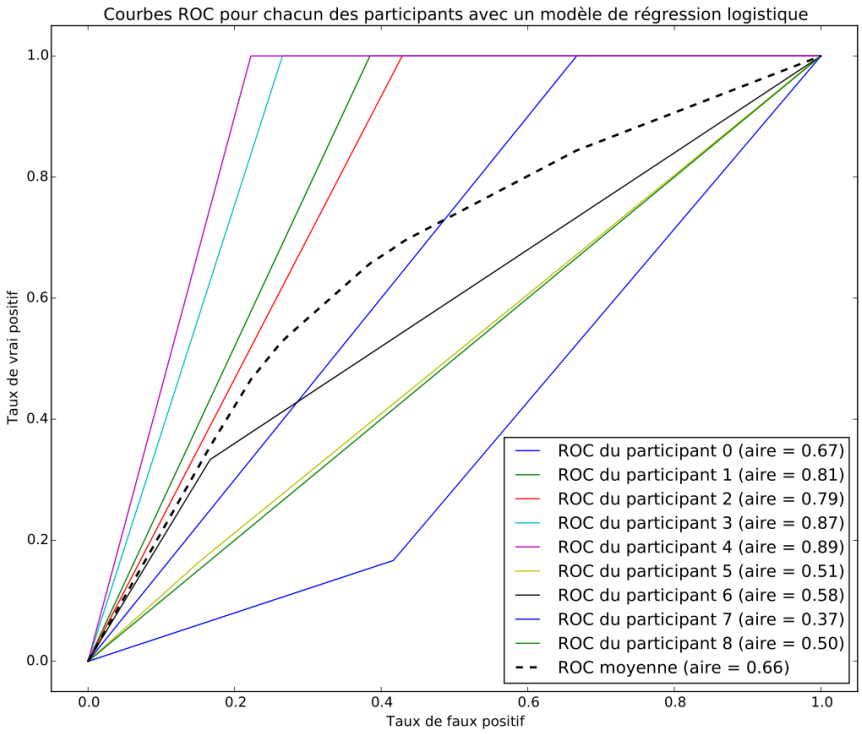

FIGURE 14 – Modèle généralisé de régression logistique, avec un schéma de "leave-one-participant-out" pour prédire si la solution sera correcte ou non.

4.1.2 Prédiction du besoin d'aide

Nous avons également sélectionné un modèle par leave-one-participant-out avec les mêmes vecteurs mais sur les étiquettes obtenues par le questionnaire, indiquant si les participants avaient désiré davantage d'aide pour chacune des tâches. Le modèle performant le mieux était un modèle de plus proches voisins uniforme, de distance Euclidienne, avec 7 voisins. La performance pour chacun des participants est illustrée par la figure 15. La moyenne

des prédictions correctes est de 74%, représentant une amélioration de 24 points au-dessus du niveau de chance de 50%.

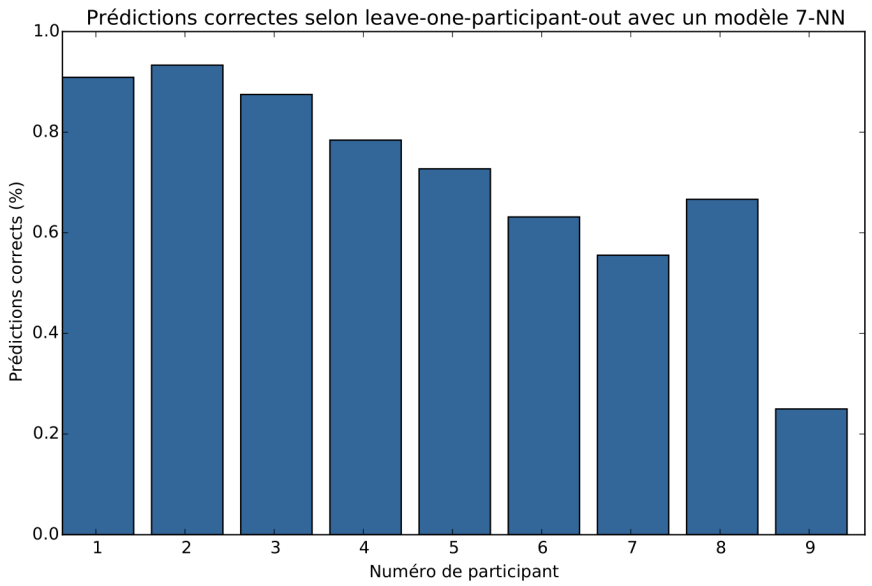

FIGURE 15 – Modèle généralisé de plus proches voisins, avec un schéma de leave-one-participant-out. Moyenne égale à 74% de prédictions correctes.

Une différence importante est visible dans la figure 15 au niveau du nombre de prédictions correctes entre les participants, démontrant l'importance des différences individuelles dans ce type d'application. Un facteur pouvant expliquer cette différence était l'utilisation d'un formulaire rempli par le participant afin d'associer une étiquette de prédiction aux mesures psychophysiologiques, puisqu'il est possible qu'un participant se soit trompé lors de la complétion du questionnaire (p.ex. un participant pensant ne pas avoir besoin d'aide alors que c'était le cas). Également, les signaux acquis lors

de l'expérience variaient en qualité (p.ex. le casque porté par un participant chauve permettant un excellent contact par comparaison à un participant aux cheveux longs). Ainsi, une analyse d'un modèle personnalisé, utilisant une validation croisée (en 4 plis) sur un participant à la fois fut effectuée. Le nombre de participants considéré pour cette approche est cependant bas (six), puisqu'il est nécessaire qu'un participant indique un niveau de difficulté différent pour deux tâches, ce qui ne fut pas le cas pour tous les participants (p.ex. un participant trouvant que toutes les tâches étaient trop difficiles). Également, nous avons effectué le découpage des vecteurs différemment afin d'obtenir plusieurs séquences de plus courtes durée. Un segment de jeu en entier pouvait durer plusieurs minutes, et le découper permettait au modèle d'apprentissage machine d'avoir une meilleure granularité temporelle. À l'intérieur d'une tâche, la séquence fut découpée par tranche de 60 secondes, avec un délai de 2 secondes, et ces tranches furent réduites en vecteur d'attributs, produisant donc plusieurs vecteurs pour une tâche pour chacun des participants. Nous avons également trouvé ces paramètres de découpe lors de la sélection du modèle.

Le meilleur modèle était un modèle de "Random Forest" [40], avec un nombre d'arbres de décision égal à 100, avec sa performance illustrée dans la figure 16. La moyenne de prédictions correctes pour tous les participants est de 85%, une augmentation de 35 points au-dessus du niveau de chance et de 11 points par rapport au modèle généralisé, visible dans la table I.

Une limite de l'approche personnalisée est le besoin d'entraîner un modèle

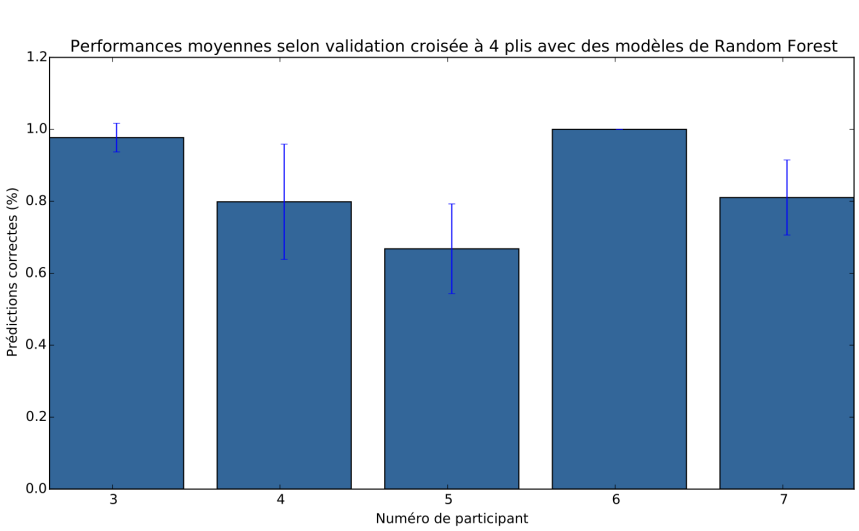

FIGURE 16 – Modèles individualisés de "Random Forest", pourcentage de prédictions correctes avec barres d'écart-types.

TABLE I – Performances de classification des différents modèles pour la première expérience

-	Modèle généralisé (réussite)	Modèle généralisé (besoin d'aide)	Modèle personnalisé (besoin d'aide)	Hasard
Performance en %	65	74	**85**	50

individuellement. Pour la seconde expérience, nous avons ajouté une étape au début du jeu consistant en deux pré-tâches, telles que décrites précédemment. Celles-ci avaient différentes difficultés afin de servir de calibrage pour un modèle personnalisé, permettant à l'environnement d'adaptation de pouvoir se servir de cette approche.

Finalement, la différence entre le post-test et le pré-test indique que le jeu

a instruit les participants par rapport aux diagrammes de Lewis. La moyenne de la différence est de 35%, la taille d'effet tel que mesuré par la méthode de Cohen (Cohen's d) est de 1.11, avec une p-value significative < 0.002 tel que mesuré par un test de Student pair. Il faut toutefois noter que le pré-test et post-test étaient courts et ces valeurs peuvent être exagérées, de plus qu'il est possible de deviner les solutions sans parfaitement les comprendre.

4.2 Analyse de la seconde expérience

Bien que nous avons initialement conduit la seconde expérience afin de tester l'adaptation, des problèmes techniques sont survenus qui ont empêché le système de fonctionner tel que prévu. Particulièrement, les valeurs du suivi du regard n'étaient pas enregistrées lors des sessions. Par conséquent, nous avons davantage porter d'attention sur l'analyse globale des deux expériences, détaillée dans la section suivante, afin de profiter d'un plus grand échantillon.

Spécifiquement à la seconde expérience, la moyenne de la différence entre le post-test et le pré-test est de 45%, avec une taille d'effet de 1.82 et une p-value < 0.000001. On peut noter que la session de jeu pendant la seconde expérience avait un effet plus large que celle de la première, bien qu'encore une fois, les valeurs étaient possiblement exagérées par la petite quantité de questions présentées par les tests.

4.3 Analyse globale

Nous présentons d'abord une analyse semblable à celle de la première expérience pour l'entièreté des données amassées au cours des deux expériences, suivie d'une discussion des problèmes rencontrés.

Avec les deux expériences combinées, nous avons produit 633 vecteurs d'attributs de 60 dimensions à partir des données de 33 participants, de la même manière que pour l'analyse de la première expérience (chacun des 633 vecteurs représente une période de jeu où le participant a soumis une solution pour la résolution d'une tâche). Considérant que nous passions de 158 vecteurs à 633, nous avions suffisamment de données afin utiliser comme étiquettes les trois catégories du questionnaire, plutôt que combiner deux catégories en une seule afin d'éviter un trop grand déséquilibre des catégories.

L'algorithme pour les classifieurs utilisés pour la deuxième expérience est une machine à vecteurs de support (SVM) avec un noyau de fonction à base radiale (RBF, "radial basis function") [10]. Afin de choisir cet algorithme et ce noyau, nous avons effectué des tests préliminaires afin de comparer différents types de modèles et ils ont suggérés que le SVM-RBF était plus performant que les autres. Ce modèle, tel qu'implémenté dans Scikit-learn [35], permet de changer le poids des catégories afin de pénaliser un modèle favorisant les classes majoritaires, obtenant une meilleure généralisation pour des banques de données qui ne sont pas équilibrées. La pénalisation se fait lors de l'entraînement : lors de la lecture des échantillons par le modèle, si les échantillons ne sont pas bien prédits, une mesure d'erreur est fournie par l'al-

gorithme afin d'ajuster les prédictions, et l'erreur fournie par un échantillon associé à classe minoritaire sera plus grande que celle d'un échantillon associé à une classe majoritaire. Les hyperparamètres furent trouvés par recherche exhaustive par grille (avec C et γ) [41] pour chacun des modèles décrits, avec un C variant de 10^{-6} à 10^{4}, et γ variant de 10^{-9} à 10^{3}.

Nous avons trouvé un modèle généralisé par validation croisée leave-one-participant-out, avec sa performance indiquée dans le tableau II sous la première colonne, en prenant compte que le niveau de performance pour une décision au hasard est de 33.3%. Également, la performance indiquée ici n'est pas exactement le pourcentage de prédictions correctes mais la moyenne des moyennes de prédictions correctes pour les trois catégories, avec un poids identique aux trois catégories. Cette mesure de performance est plus robuste pour des banques de données où les échantillons ne sont pas également distribués entre les différentes catégories. Ainsi, un modèle réussissant à classifier correctement 0% des vecteurs pour deux catégories et 100% de la troisième catégorie donnerait une moyenne de 33.3%, pénalisant un modèle favorisant la classe majoritaire.

Les colonnes suivantes indiquent des modèles entraînés sur des sous-ensembles de mesures psychophysiologiques, avec des vecteurs d'attributs plus petits. Nous avons entraîné ces modèles afin d'observer si certaines sources de données pouvaient être ignorées lors de l'adaptation. La meilleure performance est trouvée lorsque seules les données du casque sont prises en compte (54.1%), et la pire performance est trouvée en l'enlevant du vecteur

TABLE II – Prédictions correctes moyennes dépendamment des attributs et p-value entre parenthèses pour comparaison avec le cas où la prédiction s'effectue avec seulement les indices Affectiv.

-	Tous les attributs	Sans diamètres de pupille	Sans indices Affectiv	Sans expressions faciales	Seulement indices Affectiv
Correctes en %	53.4 (0.532)	53.1 (0.569)	45.6 (0.775)	53.8 (0.715)	**54.1**

(45.6%). Nous avons testé si la différence entre le meilleur modèle était significative lorsque comparée à tous les autres modèles (entraînés sur des sous-ensembles différents). Nous avons utilisé le "Wilcoxon signed-rank test", souvent utilisé pour la comparaison de performances en apprentissage machine [42]. D'après ce test, aucun sous-ensemble n'est significativement meilleur qu'un autre .

Nous avons sélectionné les modèles personnalisés de sorte que la procédure soit semblable au contexte d'une session de jeu. Généralement, on entraînerait un modèle sur un grand nombre d'échantillons pour le tester sur un plus petit nombre mis de côté. Cependant, le contexte d'une session de jeu ne nous permettait pas de recueillir une grande quantité d'observations dans le seul but de calibrer le modèle. Ainsi, nous avons effectué l'entraînement sur les premières séquences étant associées à chacune des trois catégories ("Trop facile", "Adéquate", "Trop difficile"), reproduisant ainsi un petit échantillon semblable à celui offert par un calibrage fait au début du jeu. Nous avons également découpé les séquences découpées de façon similaire à l'analyse de la première expérience mais la période de temps couverte par une séquence

fut augmentée à 90 secondes, et le délai augmenté à 20 secondes (produisant donc des séquences voisines couvrant les mêmes 70 secondes). La matrice de confusion présentée dans la table III montre les pourcentages de prédictions correctes moyennes (à travers tous les participants) pour chacune des catégories, ainsi que le total de séquences que chacune des catégories représente. La moyenne avec un poids égal pour toutes les catégories est de 67.5%, 34.5 points au-dessus du niveau de chance de 33.3%. La matrice de confusion illustre également la difficulté de mesurer un niveau de difficulté Adéquat (51.5%) lorsque comparée aux deux catégories extrêmes (au-dessus de 73%), une observation compatible avec [17].

La plus grande performance de l'approche personnalisée, comparé au modèle généralisé, s'explique par la variation se présentant chez les signaux psychophysiologiques des individus. Les différences individuelles peuvent être particulièrement marquées en ce qui concerne l'activité cérébrale [43], et utiliser un modèle qui généralise un groupe d'individus est potentiellement limité dans ses capacités. Entraîner un modèle pour chacun des participants nous a permis de prendre en compte ces différences et ainsi d'obtenir des prédictions davantage correctes puisqu'elles sont appliqués au seul participant sur lequel le modèle fut entraîné. Ce type de processus personnalisé est commun dans des applications de EEG [44].

TABLE III – Matrice de confusion pour le meilleur type de modèle, avec la moyenne dans le coin inférieur droit. Les valeurs réelles sont horizontales, alors que les prédites sont verticales.

-	Trop facile	Adéquate	Trop difficile	Total
Trop facile	**0.776**	0.034	0.190	58
Adéquate	0.092	**0.515**	0.392	291
Trop difficile	0.101	0.165	**0.733**	907
Total	164	302	790	0.675

4.4 Problèmes rencontrés et limites de l'environnement

Au cours de la seconde expérience, une erreur technique dans le code source du jeu est survenue par rapport à l'enregistrement du diamètre de la pupille, et puisque le modèle généralisé entraîné sur les données de la première expérience s'attendait à des données similaires pour les attributs reliés à cette source, le modèle a toujours déterminé que le participant n'avait pas besoin d'aide. D'un autre côté, le calibrage de l'approche personnalisée n'était pas suffisant pour produire des modèles fiables car ils décidaient presque toujours que le participant avait besoin d'aide. Cela a donc produit un jeu sérieux offrant constamment de l'aide aux participants, ce qui n'était pas le but de l'expérience.

L'erreur technique pour le diamètre de la pupille était un problème dans la spécification de la méthode d'écriture des valeurs, la solution consistant simplement à modifier le code source du jeu afin de transmettre correctement la valeur au fichier accumulant les mesures. Deux solutions sont envisagées pour le calibrage. Le jeu sérieux ne bloquait pas le participant lors de l'acquisition des données pour la tâche difficile, alors qu'idéalement, l'échec du

participant aurait dû être communiqué, lui donnant d'autres chances d'essayer de résoudre la tâche. Une autre manière, si on permet à l'utilisateur de donner des commentaires au système d'adaptation (par simple entrée textuelle), serait de présenter davantage de tâches variées à l'utilisateur puis le laisser nous indiquer la difficulté de celles-ci. La seconde méthode risque de mieux fonctionner mais demande davantage de temps et d'instructions à l'utilisateur, ce qui serait un désavantage moins important dans le cadre d'un jeu sérieux avec une longue durée d'utilisation (ce qui n'était pas le cas dans l'étude présentée ici, où le jeu peut être complété en 45 minutes). Bien qu'une autre expérience était nécessaire afin de démontrer la capacité du système d'adaptation en temps réel, les résultats présentées dans la section précédente montrent qu'il est raisonnable de s'attendre à un système capable de décider le niveau d'aide requis par un utilisateur si on a accès à son opinion sur des tâches similaires. Pour cette même raison, permettre à l'utilisateur de communiquer son opinion au début du jeu sérieux semble être la meilleure voie.

L'environnement d'adaptation est limité dans ses capacités par le matériel utilisé, particulièrement le casque EPOC de Emotiv. Alors que les deux autres sources de données sont tout à fait détachées de l'utilisateur, le casque doit être installé sur la tête du participant et il est fréquent que les électrodes aient de la difficulté à établir un bon contact compte tenu de la variance dans la forme de la tête des gens et de leur cheveux. L'installation se trouvait fastidieuse, et les électrodes pouvaient perdre contact pendant la session,

ajoutant une quantité considérable de bruit à l'enregistrement. 9.6% des données utilisées lors de l'analyse étaient manquantes, indiquant des moments lors de l'enregistrement où les capteurs ont entièrement perdu contact avec le participant. De plus, les participants ont mentionné que le casque devient inconfortable au cours de la session, ce qui devient une distraction.

Les mesures pour le diamètre de la pupille présentent également un problème considérant la sensibilité de la pupille aux différents niveaux de luminosité émis par l'écran lors d'une session typique. Même si l'éclairage de la salle est constant, les différentes scènes produisent plus ou moins de lumière, ajoutant du bruit à la mesure du diamètre.

5 Conclusion

Le but de ce travail était de développer un jeu sérieux auquel était intégré un environnement d'adaptation en temps réel utilisant un casque EEG, un instrument de suivi du regard, et une analyse automatique d'expression faciales, et d'y tester les deux hypothèses suivantes : il est possible de détecter le niveau de besoin d'aide d'un apprenant grâces aux mesures psychophysiologiques mentionnées précédemment, et un jeu sérieux peut s'adapter à l'utilisateur afin de faciliter son apprentissage en modifiant la difficulté des tâches rencontrées.

Nous avons conduit deux expériences avec le jeu sérieux afin d'obtenir des données nous permettant d'explorer ces deux hypothèses. La première expérience ne possédait pas un système capable de modifier la difficulté des tâches mais servait de méthode d'accumulation de données ainsi que de point de comparaison avec la seconde expérience, où un tel système opérait. Grâce à la première expérience, nous avons pu développer un modèle général de détection de besoin d'aide avec un classifieur de plus proches voisins, identifiant si un apprenant avait besoin d'aide ou non pour 74% des observations. Nous avons également développé une approche nous permettant d'entraîner un modèle spécifiquement pour un apprenant grâce à un algorithme de "Random Forest", celui-ci pouvant prédire si de l'aide était requise ou non pour 85% du temps. Ces deux modèles représentaient une hausse respective de 48% et 70% de performance par rapport à un modèle se basant sur le hasard

(50%).

Nous avons tenté d'identifier en temps réel le besoin d'aide d'un apprenant grâce à la seconde expérience, où le modèle général et le modèle personnalisé indiquaient au jeu de modifier la difficulté des tâches. Des erreurs techniques ont cependant empêché le bon fonctionnement des modèles. L'expérience nous a toutefois permis d'accumuler davantage de données afin de raffiner nos approches de détection de besoin d'aide. Nous avons entraîné deux modèles pouvant détecter si la tâche était trop facile, adéquatement difficile, ou trop difficile. Le premier fonctionne généralement, sans individualisation, et détectait correctement à 54% le niveau de besoin d'aide. Le second modèle est personnalisé, se calibrant sur l'apprenant, et détectait correctement à 68% le niveau de besoin d'aide. Ces modèles représentent respectivement une hausse de 70% et 106% par rapport à un modèle se basant sur la chance (33%).

Quant à la première hypothèse, nous pouvons confirmer qu'il est possible de détecter le niveau de besoin d'aide grâce à deux approches : un modèle généralisé et un modèle personnalisé, entraînés grâce à des algorithmes d'apprentissage machine. Leur performance était plus élevée que le hasard, consistant en des hausses de 70% et 106% en terme de nombres de prédictions correctes.

Notre seconde hypothèse, que le jeu sérieux pouvait faciliter l'apprentissage d'un utilisateur grâce à des modèles de prédiction de besoin d'aide, n'a pas pus être évaluée en détails. Le système d'adaptation n'a pas fonctionné

de façon adéquate lors de la seconde expérience dû à un problème technique avec le jeu, ce qui nous a empêché de tirer des conclusions à son sujet.

Une considération intéressante et qui permettrait de généraliser le travail décrit ici serait d'appliquer l'environnement d'adaptation à d'autre jeux ou logiciels. Il est peut-être possible d'entraîner un modèle à travers une session d'un jeu sérieux puis de l'appliquer sur un jeu différent sans avoir à le ré-entraîner, du moment que les types de tâches accomplies sont suffisamment semblables, un processus semblable à du transfert d'apprentissage dans le cas des modèles d'apprentissage machine.

Références

[1] T. M. Connolly, E. A. Boyle, E. MacArthur, T. Hainey, and J. M. Boyle, "A systematic literature review of empirical evidence on computer games and serious games," *Computers & Education*, vol. 59, no. 2, pp. 661–686, 2012.

[2] R. Lopes and R. Bidarra, "Adaptivity challenges in games and simulations : a survey," *Computational Intelligence and AI in Games, IEEE Transactions on*, vol. 3, no. 2, pp. 85–99, 2011.

[3] R. Ratan and U. Ritterfeld, "Classifying serious games," *Serious games : Mechanisms and effects*, pp. 10–24, 2009.

[4] T. Susi, M. Johannesson, and P. Backlund, "Serious games : An overview," 2007.

[5] I. Granic, A. Lobel, and R. C. Engels, "The benefits of playing video games.," *American Psychologist*, vol. 69, no. 1, p. 66, 2014.

[6] P. Wouters, C. Van Nimwegen, H. Van Oostendorp, and E. D. Van Der Spek, "A meta-analysis of the cognitive and motivational effects of serious games.," *Journal of Educational Psychology*, vol. 105, no. 2, p. 249, 2013.

[7] R. D. Roscoe, E. L. Snow, R. D. Brandon, and D. S. McNamara, "Educational game enjoyment, perceptions, and features in an intelligent writing tutor.," in *FLAIRS Conference*, 2013.

[8] S. B. Kotsiantis, I. D. Zaharakis, and P. E. Pintelas, "Machine learning : a review of classification and combining techniques," *Artificial Intelligence Review*, vol. 26, no. 3, pp. 159–190, 2006.

[9] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[10] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[11] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "Liblinear : A library for large linear classification," *The Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008.

[12] D. S. McNamara, G. T. Jackson, and A. Graesser, "Intelligent tutoring and games (ITaG)," *Gaming for classroom-based learning : Digital role-playing as a motivator of study*, pp. 44–65, 2010.

[13] H. S. Nwana, "Intelligent tutoring systems : an overview," *Artificial Intelligence Review*, vol. 4, no. 4, pp. 251–277, 1990.

[14] R. Ghali, S. Ouellet, and C. Frasson, "Using electrophysiological features in cognitive tasks : An empirical study," *International Journal of Information and Education Technology*, vol. 6, no. 8, p. 584, 2016.

[15] D. R. Michael and S. L. Chen, *Serious games : Games that educate, train, and inform.* Muska & Lipman/Premier-Trade, 2005.

[16] M. Ninaus, S. E. Kober, E. V. Friedrich, I. Dunwell, S. de Freitas, S. Arnab, M. Ott, M. Kravcik, T. Lim, S. Louchart, *et al.*, "Neurophysiological

methods for monitoring brain activity in serious games and virtual environments : A review," *International Journal of Technology Enhanced Learning*, vol. 6, no. 1, pp. 78–103, 2014.

[17] G. Chanel, C. Rebetez, M. Bétrancourt, and T. Pun, "Emotion assessment from physiological signals for adaptation of game difficulty," *Systems, Man and Cybernetics, Part A : Systems and Humans, IEEE Transactions on*, vol. 41, no. 6, pp. 1052–1063, 2011.

[18] A. I. A. Jabbar and P. Felicia, "Gameplay engagement and learning in game-based learning a systematic review," *Review of Educational Research*, vol. 85, no. 4, pp. 740–779, 2015.

[19] J. Rowe, B. Mott, S. McQuiggan, J. Robison, S. Lee, and J. Lester, "Crystal island : A narrative-centered learning environment for eighth grade microbiology," in *workshop on intelligent educational games at the 14th international conference on artificial intelligence in education, Brighton, UK*, pp. 11–20, 2009.

[20] C. Conati, "Probabilistic assessment of user's emotions in educational games," *Applied Artificial Intelligence*, vol. 16, no. 7-8, pp. 555–575, 2002.

[21] R. Ghali, S. Ouellet, and C. Frasson, "Lewispace : An educational puzzle game combined with a multimodal machine learning environment," in *KI 2015 : Advances in Artificial Intelligence*, pp. 271–278, Springer, 2015.

[22] A. D. McNaught and A. D. McNaught, *Compendium of chemical terminology*, vol. 1669. Blackwell Science Oxford, 1997.

[23] M. Persson and J. Bergensten, "Minecraft," *Computer software. Stockholm, Sweden : Mojang AB. Retrieved from http ://minecraft.net*, 2011.

[24] M. M. Cooper, N. Grove, S. M. Underwood, and M. W. Klymkowsky, "Lost in Lewis structures : An investigation of student difficulties in developing representational competence," *Journal of Chemical Education*, vol. 87, no. 8, pp. 869–874, 2010.

[25] Blender Online Community, *Blender - a 3D modelling and rendering package*. Blender Foundation, Blender Institute, Amsterdam,

[26] M. Den Uyl and H. Van Kuilenburg, "The facereader : Online facial expression recognition," in *Proceedings of Measuring Behavior*, vol. 30, pp. 589–590, 2005.

[27] P. Ekman and W. V. Friesen, "Facial action coding system," 1977.

[28] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 6, pp. 681–685, 2001.

[29] E. Emotiv, "Neuroheadset," 2012.

[30] E. Emotiv, "Software development kit," 2010.

[31] K. Stytsenko, E. Jablonskis, and C. Prahm, "Evaluation of consumer EEG device emotiv EPOC," in *MEi : CogSci Conference 2011, Ljubljana*, 2011.

[32] I. Ghergulescu and C. H. Muntean, "A novel sensor-based methodology for learner's motivation analysis in game-based learning," *Interacting with Computers*, vol. 26, no. 4, pp. 305–320, 2014.

[33] T. E. Tracking, "An introduction to eye tracking and tobii eye-trackers, white paper," 2010.

[34] E. H. Hess and J. M. Polt, "Pupil size in relation to mental activity during simple problem-solving," *Science*, vol. 143, no. 3611, pp. 1190–1192, 1964.

[35] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn : Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[36] L. R. Goldberg, "An alternative" description of personality" : the big-five factor structure.," *Journal of personality and social psychology*, vol. 59, no. 6, p. 1216, 1990.

[37] T. Buchanan, J. A. Johnson, and L. R. Goldberg, "Implementing a five-factor personality inventory for use on the internet," *European Journal of Psychological Assessment*, vol. 21, no. 2, pp. 115–127, 2005.

[38] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote : synthetic minority over-sampling technique," *Journal of artificial intelligence research*, pp. 321–357, 2002.

[39] G. Xu, J. Z. Huang, *et al.*, "Asymptotic optimality and efficient computation of the leave-subject-out cross-validation," *The Annals of Statistics*, vol. 40, no. 6, pp. 3003–3030, 2012.

[40] A. Liaw and M. Wiener, "Classification and regression by randomforest," *R news*, vol. 2, no. 3, pp. 18–22, 2002.

[41] C.-W. Hsu, C.-C. Chang, C.-J. Lin, *et al.*, "A practical guide to support vector classification," 2003.

[42] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *The Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.

[43] P. N. Mohr and I. E. Nagel, "Variability in brain activity as an individual difference measure in neuroscience ?," *The Journal of Neuroscience*, vol. 30, no. 23, pp. 7755–7757, 2010.

[44] M. Lang, "Investigating the emotiv epoc for cognitive control in limited training time," *Honours report, University of Canterbury*, p. 8, 2012.